УНИВЕРЗИТЕТ У БЕОГРАДУ

МАТЕМАТИЧКИ ФАКУЛТЕТ

Миљана В. Младеновић

ИНФОРМАТИЧКИ МОДЕЛИ У АНАЛИЗИ ОСЕЋАЊА ЗАСНОВАНИ НА ЈЕЗИЧКИМ РЕСУРСИМА

докторска дисертација

Београд, 2016.

UNIVERZITET U BEOGRADU

MATEMATIČKI FAKULTET

Miljana V. Mladenović

INFORMATIČKI MODELI U ANALIZI OSEĆANJA ZASNOVANI NA JEZIČKIM RESURSIMA

doktorska disertacija

Beograd, 2016.

UNIVERSITY OF BELGRADE

FACULTY OF MATHEMATICS

Miljana V. Mladenović

INFORMATION MODELS IN SENTIMENT ANALYSIS BASED ON LINGUISTIC RESOURCES

Doctoral Dissertation

Belgrade, 2016.

Ментор:

др Душко Витас, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Душко Витас, ванредни професор
Универзитет у Београду, Математички факултет

др Гордана Павловић-Лажетић, редовни професор
Универзитет у Београду, Математички факултет

др Ненад Митић, ванредни професор
Универзитет у Београду, Математички факултет

др Владан Девеџић, редовни професор
Универзитет у Београду, Факултет организационих наука

др Цветана Крстев, редовни професор
Универзитет у Београду, Филолошки факултет

Датум одбране:

_____

Наслов дисертације: Информатички модели у анализи осећања засновани на језичким ресурсима

Резиме: Почетак новог миленијума обележен је бурним развојем друштвених мрежа, интернет технологијама у облаку и применом вештачке интелигенције у веб алатима. Изузетно брз раст броја текстова на интернету (блогова, сајтова за електронску трговину, форума, дискусионих група, система за пренос кратких порука, друштвених мрежа и портала за објаву вести) увећао је потребу за развојем метода брзе, свеобухватне и прецизне анализе текста. Због тога је значајан развој језичких технологија чији су примарни задаци: класификација докумената (енг. Document classification), груписање докумената (енг. Document clustering), проналажење информација (енг. Information Retrieval), разрешавање значења вишезначних речи (енг. Word-sense disambiguation), екстракција из текста (енг. Text extraction), машинско превођење (енг. Machine translation), рачунарско препознавање говора (енг. Computer speech recognition), генерисање природног језика (енг. Natural language generation), анализа осећања (енг. sentiment analysis), итд. У рачунарској лингвистици данас је у употреби више различитих назива за област чији је предмет интересовања обрада осећања у тексту: класификација према осећању (енг. sentiment classification), истраживање мишљење (енг. opinion mining), анализа осећања (енг. sentiment analysis), екстракција осећања (енг. sentiment extraction). По својој природи и методама које користи, анализа осећања у тексту спада у област рачунарске лингвистике која се бави класификацијом текста. У процесу обраде осећања се, у општем случају, говори о три врсте класификације текстова:

- идентификацији субјективности (енг. opinion classification или subjectivity identification) којом се текстови деле на оне који носе емоционални садржај и оне који имају искључиво чињенични садржај;
- класификацији осећања (енг. sentiment classification) или идентификацији поларитета осећања (енг. polarity identification) којим се текстови који носе емоционални садржај деле на оне са позитивним и оне са негативним емоционалним садржајем;

- одређивању снаге или интензитета осећања (енг. strength of orientation).

У погледу нивоа на коме се анализа осећања врши, разликујемо три методологије: анализу на нивоу документа, на нивоу реченице и на нивоу атрибута. Класификација докумената се обично спроводи методама машинског учења (енг. machine learning) или помоћу система заснованих на правилима (енг. rule-based systems). Анализа осећања, као специфичан вид класификације докумената, такође користи ове методе.

Ова докторска дисертација, чији је основни задатак анализа осећања у тексту, представља истраживање које се односи на класификацију на основу осећања (енг. sentiment classification) на нивоу докумената написаних на српском језику, вероватносном методом машинског учења полиномијалне логистичке регресије, односно, максималне ентропије. Циљ овог истраживања је генерисање првог свеобухватног, флексибилног, модуларног система за анализу текстова на српском језику на основу осећања уз помоћ дигиталних ресурса какви су: семантичке мреже, специјализовани лексикони и доменске онтологије. У том смислу истраживање је подељено у две фазе. У првој фази се развијају методе и алати којима се открива поларитет осећања дословног значења текста. У овом делу рада се предлаже, примењује и оцењује нова метода редукције векторског простора предиктора који се користе у процесу класификације текста на основу осећања. Предложена метода редукције примењује се у класификационом моделу максималне ентропије, а ослања се на употребу лексичко-семантичке мреже *WordNet* и специјализованог лексикона сентименталних речи и израза. Предложена метода састоји се из два сукцесивна поступка. Први се односи се на проширење векторског простора предиктора флективним облицима предиктора. У истраживању је показано да у анализи осећања употреба стемера[1], као стандардног метода смањења векторског простора у класификаторима текста, може довести до непотпуних или нетачних ознака предиктора поларитета осећања, док се увођењем флективних облика

[1] Софтверски алат којим се врши уклањање суфикса речи при чему се не губи семантичка информација.

предиктора то се може избећи. У раду се, затим, показује да иницијално повећање векторског простора услед увођења флективних облика, може бити успешно редуковано применом другог предложеног поступка – семантичког пресликавања свих предиктора истог поларитета осећања у мањи број семантичких класа. Тиме се векторски простор предиктора редукује у односу на иницијални.

У другој фази дисертације описује се израда и имплементација формалних онтологија реторичких фигура српског језика - доменске онтологије и онтологије задатака. Приказана је имплементација онтологије задатака у процесу генерисања предиктора фигуративног говора. Примењени су поступци закључивања дескрпитивном логиком на основу правила дефинисаних у самој онтологији. Задатак истаживања друге фазе је препознавање фигуративног говора како би се унапредио постојећи скуп предиктора генерисан у првој фази истраживања. Резултати истраживања ове фазе показују да се неке класе стилских фигура могу аутоматски препознавати што може унапредити процес класификације на основу осећања.

У току рада на овој дисертацији пројектован је, имплементиран и статистички оцењен софтверски алат SAFOS (Sentiment Analysis Frameword for Serbian) који представља интегрисани систем за класификацију текстова на основу осећања на српском језику. Резултати истраживања у оквиру ове дисертације приказани су у радовима (Mladenović & Mitrović, 2013; Mladenović & Mitrović, 2014; Mladenović, Mitrović & Krstev, 2014; Mladenović, Mitrović, Krstev & Vitas, 2015; Mitrović, Mladenović, & Krstev, 2015; Mladenović, Mitrović & Krstev, 2016).

Дисертација садржи седам поглавља следеће структуре. У поглављу 1 уведени су и дефинисани појмови и методе који се користе у првој фази истраживања: класификација текста, класификација текстова на основу осећања, машинско учење, надгледано машинско учење, вероватносно надгледано машинско учење, модели репрезентације докумената. На крају уводног дела дефинисани су задаци и циљеви истраживања. У поглављу 2 приказан је математички модел метода класификације текстова и

класификације текстова на основу осећања. Дат је математички модел вероватносне класификације текста и примена вероватносне класификације у регресионим моделима. На крају поглавља приказан је математички модел метода максималне ентропије као једног од регресионих модела који се успешно примењује у задацима обраде природног језика. У поглављу 3 приказани су лексички ресурси српског језика као и методе и алати њихове обраде. Поглавље 4 бави се свеобухватним истраживањем тренутно актуелних типова и метода класификације текстова на основу осећања. Поглавље 5 разматра допринос ове дисертације методама редукције векторског простора предиктора коришћених у класификационом моделу максималне ентропије. Најпре се врши анализа метода редукције предиктора. Предлаже се метода редукције којом се унапређује класификација текстова на основу осећања. Дефинише се математички модел предложене методе. Уводе се и описују скупови за учење и скупови за тестирање. Дефинишу се лексичко-семантички ресурси који се користе у примени предложене методе. У поглављу 5 дат је опис изградње и евалуације система за класификацију текстова на основу осећања – SAFOS у коме се примењује и оцењује предложени метод редукције векторског простора предиктора. Дефинишу се параметри и функције SAFOS-а, као и параметри за оцену система: прецизност, одзив, F1-мера и тачност. Даје се опис одабране методе за оцену статистичке значајности предложеног модела и начин примене над скуповима за тестирање у систему SAFOS. На крају поглавља дат је приказ изведених експеримената, резултата и оцена система. Поглавље 6 односи се на поступке препознавања фигуративног говора ради побољшања класификације текстова на основу осећања. Уведен је појам доменске онтологије. Дат је појам реторичке фигуре. Размотрен је значај фигуративног говора у класификацији текстова на основу осећања. Дат је опис изградње и структуре прве формалне доменске онтологије реторичких фигура за српски језик *RetFig.owl* и одговарајуће онтологије задатака дефинисане на основу правила за препознавање одређених класа реторичких фигура применом дескриптивне логике. На крају овог одељка

дат је приказ изведених експеримената, резултата и оцена софтверског додатка систему SAFOS који препознаје фигуративни говор.

Закључно поглавље ове дисертације бави се приказом резултата, проблема и недостатака у систему SAFOS. Најзад, указује се и на технолошки, друштвени, педагошки и научни значај анализе осећања и препознавања фигуративног говора и дају даље смернице у процесу развоја система SAFOS.

Кључне речи: обрада природног језика, анализа осећања, класификација текстова према осећањима, метода максималне ентропије, екстракција предиктора, доменска онтологија, онтологија задатака, реторичке фигуре, WordNet

Научна област: Рачунарство

Ужа научна област: Рачунарска лингвистика

УДК број: [004.822+004.838.2]:004.912(043.3)

Title of the dissertation: Information Models in Sentiment Analysis Based on Linguistic Resources

Abstract: The beginning of the new millennium was marked by huge development of social networks, internet technologies in the cloud and applications of artificial intelligence tools on the web. Extremely rapid growth in the number of articles on the Internet (blogs, e-commerce websites, forums, discussion groups, and systems for transmission of short messages, social networks and portals for publishing news) has increased the need for developing methods of rapid, comprehensive and accurate analysis of the text. Therefore, remarkable development of language technologies has enabled their applying in processes of document classification, document clustering, information retrieval, word sense disambiguation, text extraction, machine translation, computer speech recognition, natural language generation, sentiment analysis, etc. In computational linguistics, several different names for the area concerning processing of emotions in text are in use: sentiment classification, opinion mining, sentiment analysis, sentiment extraction. According to the nature and the methods used, sentiment analysis in text belongs to the field of computational linguistics that deals with the classification of text. In the process of analysing of emotions we generally speak of three kinds of text classification:

- identification of subjectivity (opinion classification or subjectivity identification) used to divide texts into those that carry emotional content and those that only have factual content
- sentiment classification (polarity identification) of texts that carry emotional content into those with positive and those with negative emotional content
- determining the strength or intensity of emotional polarity (strength of orientation).

In terms of the level at which the analysis of feelings is carried out, there are three methodologies: an analysis at the document level, at the sentence level and at the level of attributes. Standardized methods of text classification usually use machine learning methods or rule-based techniques. Sentiment analysis, as a specific type of classification of documents, also uses these methods.

This doctoral thesis, whose main task is the analysis of emotions in text, presents research related to the sentiment classification of texts in Serbian

language, using a probabilistic method of machine learning of multinomial logistic regression i.e. maximum entropy method. The aim of this research is to create the first comprehensive, flexible, modular system for sentiment analysis of Serbian language texts, with the help of digital resources such as: semantic networks, specialized lexicons and domain ontologies. This research is divided into two phases. The first phase is related to the development of methods and tools for detecting sentiment polarity of literal meaning of the text. In this part of the work, a new method of reducing the feature vector space for sentiment classification is proposed, implemented and evaluated. The proposed method for reduction is applied in the classification model of maximum entropy, and relies on the use of lexical-semantic network WordNet and a specialized sentiment lexicon. The proposed method consists of two successive processes. The first process is related to the expansion of feature vector space by the inflectional forms of features. The study has shown that usage of stemming in sentiment analysis as a standard method of reducing feature vector space in text classification, can lead to incomplete or incorrect sentiment-polarity feature labelling, and with the introduction of inflectional feature forms, this problem can be avoided. The paper shows that a feature vector space, increased due to the introduction of inflectional forms, can be successfully reduced using the other proposed procedure – semantic mapping of all predictors with the same sentiment-polarity into a small number of semantic classes. In this way, the feature vector space is reduced compared to the initial one, and it also retains the semantic precision.

The second phase of the dissertation describes the design and implementation of formal ontologies of Serbian language rhetorical figures – the domain ontology and the task ontology. Usage of the task ontology in generating features representing figurative speech is presented. The research aim of the second phase is to recognize figurative speech to be used in improving of the existing set of predictors generated in the first phase of the research. The research results in this phase show that some classes of figures of speech can be recognized automatically.

In the course of working on this dissertation, a software tool SAFOS (Sentiment Analysis Framework for Serbian), as an integrated system for

sentiment classification of text in Serbian language, has been developed, implemented and statistically evaluated. Results of the research within the scope of this thesis are shown in papers (Mladenović & Mitrović, 2013; Mladenović & Mitrović, 2014; Mladenović, Mitrović & Krstev, 2014; Mladenović, Mitrović, Krstev & Vitas, 2015; Mladenović, Mitrović & Krstev, 2016).

The dissertation consists of seven chapters with the following structure. Chapter 1 introduces and defines methods, resources and concepts used in the first phase of research: text classification, sentiment classification, machine learning, supervised machine learning, probabilistic supervised machine learning, and language models. At the end of the introductory section, the tasks and objectives of the research have been defined. Chapter 2 presents a mathematical model of text classification methods and classification of sentiment methods. A mathematical model of a probabilistic classification and an application of the probabilistic classification in regression models are presented. At the end of the chapter it is shown that the method using the mathematical model of maximum entropy, as one of the regression models, has been successfully applied to natural language processing tasks. Chapter 3 presents the lexical resources of the Serbian language and the methods and tools of their processing. Chapter 4 deals with the comprehensive research on the currently available types and methods of sentiment classification. It shows the current work and research in sentiment classification of texts. It also presents a comparative overview of research in sentiment classification of texts using the method of maximum entropy. Chapter 5 discusses the contribution of this thesis to methods of feature space reduction for maximum entropy classification. First, a feature space reduction method is analysed. A new feature space reduction method which improves sentiment classification is proposed. A mathematical model containing proposed method is defined. Learning and testing sets and lexical-semantic resources that are used in the proposed method are introduced. Chapter 5 also describes building and evaluation of a system for sentiment classification – SAFOS, which applies and evaluates the proposed method of a feature vector space reduction. The parameters and the functions of SAFOS are defined. Also, measures for evaluation of the system were discussed – precision, recall, F1-measure and accuracy. A

description of the method for assessing the statistical significance of a system is given. Also, implementation of the statistical test in the system SAFOS is discussed. The chapter provides an overview of the presented experiments, results and evaluation of the system. Chapter 6 deals with methods of recognizing figurative speech which can improve sentiment classification. The notion of domain ontology is introduced, the role of rhetorical figures and domain ontology of rhetorical figures. The importance of figurative speech in the sentiment classification has been explored. The description of the construction and structure of the first domain ontology of rhetorical figures in Serbian language, *RetFig.owl*, is given. Also, the description of the construction and structure of the corresponding task ontology that contains rules for identification of some classes of rhetorical figures is given. At the end of this chapter, an overview of the performed experiments, results and evaluation of the SAFOS system plugin that improved the recognition of figurative speech is given.

The final chapter of this study deals with the achievemnts, problems and disadvantages of the SAFOS system. The conclusion of this thesis points to the great technological, social, educational and scientific importance of the sentiment analysis and recognition of the figurative speech and gives some routes in further development of the SAFOS system.

Key words: natural language processing, opinion mining, sentiment analysis, maximum entropy method, feature extraction, domain ontology, task ontology, rhetorical figures, WordNet

Scientific field: Computer science

Scientific subfield: Computational linguistics

UDK number: [004.822+004.838.2]:004.912(043.3)

Садржај

Листа слика

Листа табела

Листа алгоритама

Скраћенице

ACL	...	Association for Computational Linguistics
AMI	...	Augmented Multi-party Interaction
AOL	...	Acronym, abbreviation, shorthand or slang term
API	...	Application Program Interface
ASCII	...	American Standard Code for Information Interchange
BCS	...	Base Concepts
BILI	...	Balkan Interlingual Index
BWN	...	BalkaNet
CES	...	Corpus Encoding Standard
CSV	...	Comma-separated Values
CV	...	Cross-validation
DDC	...	Dewey Decimal Classification
DELA	...	Dictionnaires électroniques du LADL
ДЈР	...	Дигитални језички ресурси
DL	...	Description Logic
DOLCE	...	Descriptive Ontology for Linguistic and Cognitive Engineering
ELRA	...	The European Language Resources Association
E-MELD	...	Electronic Metastructure for Endangered Languages Data
ESL	...	English as a Second Language
EWN	...	EuroWordNet
FE	...	Feature Extraction
FS	...	Feature Selection
FTC	...	Fairy Tale Corpus
GI	...	General Inquirer
GIS	...	Generalized Iterative Scaling
GML	...	Geographical Markup Language
GOLD	...	General Ontology for Linguistic Description
HEO	...	Human Emotion Ontology
HMC	...	Hate Mail Corpus
HTML	...	HyperText Markup Language
IC	...	Information Content
IIS	...	Improved Iterative Scaling
ILI	...	Interlingual Index
IMDB	...	Internet Movie DataBase
IR	...	Information Retrieval
IRLS	...	Iterative Reweighted Least Squares
JANTOR	...	Java Annotation Tool Of Rhetoric
KIF	...	Knowledge Interchange Format
L-BFGS	...	Limited-memory Broyden-Fletcher-Goldfarb-Shanno Method

LDA	...	Linear Discriminant Analysis
LIWC	...	Linguistic Inquiry and WordCount
LLC	...	Love Letters Corpus
LM	...	Language Model
LR	...	Linear Regression
LSA	...	Latent Semantic Analysis
LSO	...	Lexical Semantic Orientated Methods
MAP	...	Maximum a Posteriory Probability
MaxEnt	...	Maximum Entropy Model
MDS	...	Multidimensional Scaling
ME	...	Maximum Entropy Classification
MEDS	...	Electronic Morphological Dictionaries for Serbian
ML	...	Machine Learning
MLE	...	Maximum Likelihood Estimation
MPQA	...	Multi-Perspective Question Answering
MRL	...	Morphologically Rich Language
MWU	...	Multi-word Units
NB	...	Naive Bayes
NER	...	Named Entity Recognition
NHS	...	National Health Service
NRC	...	National Research Council Canada
OCC	...	Ortony, Clore and Collins Model
OGC	...	Open Geospatial Consortium
OWL	...	Web Ontology Language
PCA	...	Principal Component Analysis
PMI	...	Pointwise Mutual Information
PMI-IR	...	Pointwise Mutual Information for Information Retrieval
POS	...	Part of Speech
PWN	...	Princeton WordNet
RDF	...	Resource Description Framework
RDFS	...	Resource Description Framework Schema
ROC	...	Receiver Operating Characteristic
SA	...	Sentiment Analysis
SAFOS	...	Sentiment Analysis Frameword for Serbian
SNC	...	Suicide Notes Corpus
SO	...	Semantic Orientation
SO-CAL	...	Semantic Oriented Calculator
SPARQL	...	SPARQL Protocol and RDF Query Language
SPSS	...	Statistical Package for the Social Sciences
SRP	...	Serbian Interlingual Index
SrpMED	...	Elektronski morfološki rečnici srpskog jezika
SUMO	...	Suggested Upper Merged Ontology

SUO	. . .	Standard Upper Ontology
SVM	. . .	Support Vector Machines
SWL	. . .	Stop Words List
SWN	. . .	Serbian WordNet
SWNonto	. . .	Ontologija Srpski vordnet
SWRL	. . .	Semantic Web Rule Language
TEI	. . .	The Text Encoding Initiative
TF	. . .	Term Frequency
TF-IDF	. . .	Term Frequency – Inverse Document Frequency
TOEFL	. . .	Test of English as a Foreign Language
URI	. . .	Uniform Resource Identifier
W3C	. . .	World Wide Web Consortium
WFOL	. . .	WonderWeb Foundational Ontologies Library
WS4LR	. . .	Workstation for Lexical Resources
XML	. . .	Extensible Markup Language
XPATH	. . .	XML Path Language
XSD	. . .	XML Schema Definition
XSLT	. . .	eXtensible Stylesheet Language Transformations
ZB	. . .	zettabyte (10^{21}bytes)

1. УВОД

Снажан развој интернета и његов утицај на сва поља људског рада обележили су прву декаду новог миленијума. Веб, најважнији део светске мреже, преплављен је текстуалним садржајима различите намене - почев од дневних вести, образовних, научних, пословних, па до забавних информација. Све то праћено је исто тако брзим развојем различитих уређаја за генерисање, пренос и претрагу текстова - лаптопова, мобилних уређаја, таблета, читача дигиталних књига и комуникационе опреме. Потреба за брзом, свеобухватном и прецизном претрагом текста постаје све већа, што условљава динамичан развој језичких технологија чији су примарни задаци: класификација докумената (енг. Document classification), груписање докумената (енг. Document clustering), проналажење информација (енг. Information retrieval), разрешавање значења речи (енг. Word-sense disambiguation), екстракција из текста (енг. Text extraction), машинско превођење (енг. Machine translation), рачунарско препознавање говора (енг. Computer speech recognition), генерисање природног језика (енг. Natural language generation). С друге стране, језичке технологије обухватају три основне компоненте: језичке ресурсе, језичке алате и производе језичких технологија. Језички ресурси обухватају корпусе, дигиталне речнике, лексичке базе, семантичке мреже, лингвистичке онтологије и др. Језички алати су методе и поступци над језичким ресурсима, а производи су заокружене апликације које примењују достигнућа језичких алата у обради језичких ресурса.

Изузетно брз раст броја текстова на интернету: блогова, сајтова за електронску трговину, форума, дискусионих група, система за пренос кратких порука, друштвених мрежа и портала за објаву вести, нагласио је важност и субјективног мишљења аутора текста. Постало је врло важно добити, осим чињеничне информације, и субјективну оцену неког догађаја, услуге или производа. Данас постоји општи интерес јавности и за таквим

видом информација. Од појединачних интереса купаца који желе да знају мишљења других људи пре куповине производа или услуге, преко производних и трговинских компанија заинтересованих за позитивне и негативне критике о својим производима и услугама, до политичких организација и влада држава које желе да имају информације о мишљењу бирача. У рачунарској лингвистици данас је у употреби више различитих назива за област чији је предмет интересовања обрада осећања у тексту: класификација на основу осећања (енг. sentiment classification), истраживање мишљења (енг. opinion mining), анализа осећања (енг. sentiment analysis), екстракција осећања (енг. sentiment extraction), екстракција процена (енг. appraisal extraction).

Појам „истраживање мишљења“ (opinion mining) први пут је употребљен у раду (Dave et al., 2003) да би означио скуп процеса – идентификацију објекта који је предмет истраживања, генерисање атрибута којима се објекат-предмет истраживања описује, проналажење, оцену и систематизацију мишљења и ставова о вредностима тих атрибута и сумаризацију опште оцене о посматраном објекту, при чему се истраживање врши над текстовима, пре свега добијених са веба. Појам „анализа осећања“ (sentiment analysis) појавио се 2001. у радовима (Tong, 2001; Das & Chen, 2001) и односио се на поступак аутоматске анализе текстова проналажењем и оценом тврђења којима се исказују осећања у текстовима објављеним на вебу. Према Пенгу и Ли[2] (Pang & Lee, 2008, pp. 6) анализа осећања (sentiment analysis) и истраживање мишљења (енг. opinion mining) представљају синониме и могу се користити равноправно да означе скуп метода и техника којима се у тексту прате и истражују информације субјективног карактера. С друге стране, информације субјективног карактера се могу односити на различите врсте емоционалних реакција. У наставку ћемо укратко изложити шта се подразумева под појмом осећање и који су синоними овог појма.

Осећај је основни чулни податак (нпр. љуто, кисело, светло, мрачно). Осећај, као чулна сензација производи емоционалну реакцију. Емоционалне

[2] Bo Pang and Lillian Lee

реакције могу бити различите. Обично разликујемо више типова на основу њихове сложености, интензитета и дужине трајања. Према Брковићу (Brković, 2011, pp. 279-280) то су:

- афективна реакција – интензивно али краткотрајно изражавање доживљаја пријатности или непријатности, праћено изразитим телесним феноменима (нпр. бес, паника);
- осећање или емоција – психички процес којим вреднујемо сазнато, изражавамо субјективни однос према догађајима, особама и појавама (нпр. радост, жалост, страх, љутња). Осећање (емоција) представља резултат синтезе више осећаја, односно, по својој структури је комплексније од осећаја као чулне сензације;
- сентимент – сложена емоција, трајни афективни и конативни однос према нечему (другој особи, објекту, апстрактној идеји) који не захтева рационално образложење. Сентимент се одређује како објектом тако и самим односом између носиоца сентимента и тог објекта (нпр. љубав према отаџбини - патриотизам);
- расположење – стање продуженог трајања осећања које се одликује мањим и равномернијим интензитетом од осећања. Зависи од темперамента, климе, периода године, ендокриног система, физиолошког стања и не мора бити усмерено ка конкретној особи, објекту или идеји.

С обзиром да је у рачунарској лингвистици обрада осећања у тексту везана за посматрање и праћење конкретног објекта, појаве или особе, може се рећи да се највећим делом ради о изражавању осећања или емоција. У одељку 4.4.4 изложићемо и актуелна истраживања која се односе на исказивање расположења.

У процесу обраде осећања у текстовима говори се о три врсте идентификације осећања:

- идентификацији субјективности (енг. opinion classification или subjectivity identification) (Esuli & Sebastiani, 2006) којом се текстови деле на оне који носе емоционални садржај и оне који имају искључиво чињенични садржај;

- класификацији на основу осећања (енг. sentiment classification) или идентификацији поларитета осећања (енг. polarity identification) (Turney & Littman, 2003) којим се текстови који носе емоционални садржај деле на оне са позитивним и оне са негативним емоционалним садржајем;
- одређивању снаге или интензитета осећања (енг. strength of orientation) што се може урадити на различите начине: кроз идентификовање осећања израженог у тексту претходним утврђивањем квалитативне скале од слабог ка јаком позитивном или негативном осећању (Wilson, Wiebe & Hwa, 2006), претходним утврђивањем класа осећања (Strapparava & Mihalcea, 2008) или утврђивањем нумеричке (квантитативне) скале оцене јачине осећања израженог у тексту и увођењем метода за оцену јачине утврђеног осећања (Hatzivassiloglou & Wiebe, 2000; Pang & Lee, 2005; Taboada, Brooke, Tofiloski, Voll & Stede, 2011).

По својој природи и методама које користи, анализа осећања у тексту спада у област рачунарске лингвистике која се бави класификацијом текста. При томе, идентификација на основу субјективности (енг. subjectivity identification) и класификација на основу осећања (енг. sentiment classification) су класификациони процеси код којих постоје две класе (објективно-субјективно и позитивно-негативно) због чега су сврстани у подврсту бинарних класификатора текста.

У погледу нивоа на коме се анализа осећања врши, разликујемо три методологије. У првом случају, сваки документ се посматра као јединствени објекат анализе осећања (SA), а сва мишљења потичу од једног посматрача који називамо „извор осећања“. Под таквим претпоставкама сви документи се могу класификовати као субјективни или објективни, а затим се субјективни сврставају у позитивну или негативну класу (Missen, Boughanem & Cabanac, 2009) зависно од утврђеног поларитета израженог осећања. У том случају ради се о анализи осећања на нивоу документа. Друга позната методологија у анализи осећања, као објекат посматрања узима реченицу. У том случају се документ дели на реченице, уз претпоставке да је сада свака од њих понаособ носилац осећања о једном објекту и сва осећања у реченици потичу од једног носиоца осећања. У датом случају имамо анализу осећања

на нивоу реченице и можемо утврдити да ли је реченица субјективна или објективна, а ако је субјективна, могуће је одредити (класификовати) да ли је поларитет њоме израженог осећања позитиван или негативан (Liu, 2010). Коначно, анализа осећања може бити и на нивоу атрибута. У случајевима када је објекат посматрања сложен или постоји више извора осећања, он мора бити посматран као скуп атрибута од којих је сваки понаособ предмет анализе осећања (на пример фото-апарат, као објекат посматрања, може бити оцењен сумаризацијом појединачних оцена свих његових значајних саставних делова) (Liu, 2011). Класификација докумената се обично изводи методама машинског учења (енг. machine learning) (Sebastiani, 2002) или помоћу система заснованих на правилима (енг. rule-based systems) (Liu, Hsu & Ma, 1998). Анализа осећања, као специфичан вид класификације докумената, користи исте групе метода машинског учења (Blinov, Klekovkina, Kotelnikov & Pestov, 2013). Оне могу бити:

- методе ненадгледаног учења (енг. unsupervised learning) (Turney, 2002) помоћу којих се врши обележавање субјективности и поларитета осећања речи или фраза у текстовима на основу правила лексичко-семантичких мрежа као што су *WordNet*, *ConceptNet* или знања о осећањима похрањена у доменским онтологијама или онтологијама високог нивоа.
- методе надгледаног учења (енг. supervised learning) (Khairnar & Kinikar, 2013) које могу бити вероватносне класификационе методе машинског учења као што су методе засноване на Бајесовом[3] правилу (енг. Naive Bayes) (Gamallo & Garcia, 2014) и регресионе методе каква је метода максималне ентропије (енг. Maximum Entropy) (Mehra, Khandelwa & Patel, 2002), или пак могу бити невероватносне класификационе методе каква је метода потпорних вектора (енг. Support vector machines) (Mullen & Collier, 2004).

[3] Thomas Bayes

1.1 Машинско учење

Машинско учење (енг. Machine Learining – ML) је интердисциплинарна област која се ослања на информатику, вештачку интелигенцију и статистику. Постоји више различитих дефиниција, од којих су најзначајније:

- дефиниција Артура Семјуела[4] (Samuel, 1959)

Дефиниција 1.1 *Машинско учење је истраживачка област која рачунарима пружа способност рада „учењем“ из података, а не искључиво извршавањем програмског кода.*

- формална дефиниција Тома Мичела[5] (Mitchel, 1997)

Дефиниција 1.2 *За један рачунарски програм се може рећи да „учи“ из искуства E у односу на неку класу задатака T и меру перформансе П, ако он може да дате перформансе решавања задатака класе T, мерених перформансом П, побољшава на основу искуства E.*

У новије време, дефиниције појма машинског учења разматрају се са два становишта: инжењерско-рачунарског и математичког, односно статистичког, па сходно томе, према (Janičić & Nikolić, 2010, pp. 161) оне могу бити дате респективно као:

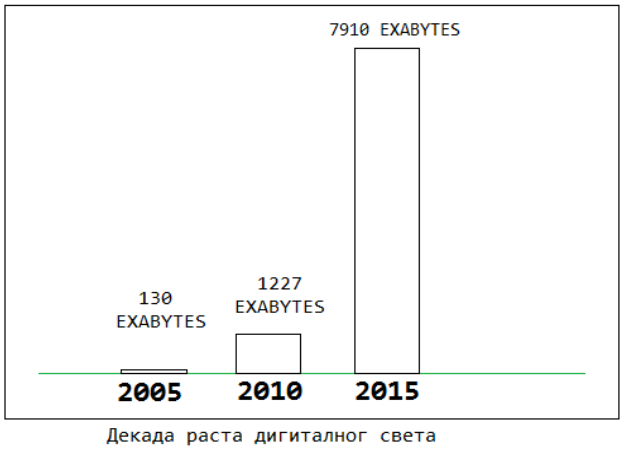
Дефиниција 1.3 *Машинско учење је дисциплина која се бави изградњом прилагодљивих рачунарских система који су способни да побољшавају своје перформансе користећи информације искуства.*

Дефиниција 1.4 *Машинско учење је дисциплина која се бави проучавањем генерализације и конструкцијом и анализом алгоритама који генерализују.*

Значај примене ML расте са порастом количине и броја врста дигиталних података (текстова, слика, база података, мултимедијалних записа, итд.). Генерисање знања из „мора“ података један је од главних циљева машинског

[4] Arthur Lee Samuel
[5] Tom Michael Mitchell

учења. Истраживање Дицитал Јуниверса (Digital Universe)[6] из 2011. године показује да количина дигиталних података која се ствара и репродукује на годишњем нивоу износи, у 2011. години, 1.8 ZB (зетабајтова) или 10^{21} B (бајтова), а да ће за пет година порасти шест пута (слика 1.1).



Слика 1.1 Раст количине дигиталних података у свету
(Извор: www.emc.com)[7]

Неопходност откривања знања које лежи у великим количинама разноврсних дигиталних података један је од главних разлога развитка ове дисциплине. Тај генерални захтев, постављен као циљ примене машинског учења, доживљава своју конкретизацију у многим применама као што су:

- препознавање образаца (Pattern Recognition)
- оптичко препознавање знакова (Optical Character Recognition)
- препознавање говора (Speech Recognition)
- роботика и интелигентни системи са способношћу учења
- обрада природног језика (Natural Language Processing)
- претраживање информација (Search Mashines)
- медицинска дијагностика (Computer-aided Dagnosis)
- биометријска идентификација и ауторизација (Biometrics Authentication), итд.

[6] www.emc.com
[7] http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf

Међутим, постоје и други разлози због којих технике машинског учења имају предност над традиционалним програмским решењима.

- Неки се проблеми због своје сложености не могу једноставно и једнозначно дефинисати (препознавање лица, говора, итд.).
- Неки проблеми нису довољно познати, истражени или нису познати параметри њихових модела у тренутку када се систем за њихово решавање пројектује.
- Некада су проблеми оптерећени огромном количином улазних података између којих није могуће експлицитно и једнозначно утврдити релације.
- Проблеми који се мењају динамички, захтевају програмска решења способна да се прилагођавају тим променама у времену.

Најзад, на основу разматрања дефиниције појма и задатака машинског учења, можемо указати и на двосмисленост значења превода „Машинско учење“ из изворног назива „Machine Learning“ и указати на могућност одабира назива (Учење машина, Обучавање машина и сл.) који би једнозначно указивао на чињеницу да су машине објекти процеса учења и да се процес учења, пре свега, односи на машине, а да тек потом, продукт тог учења може донети одређена знања и самом човеку.

1.2 Процес закључивања

Интелигентни системи поседују два основна вида закључивања:

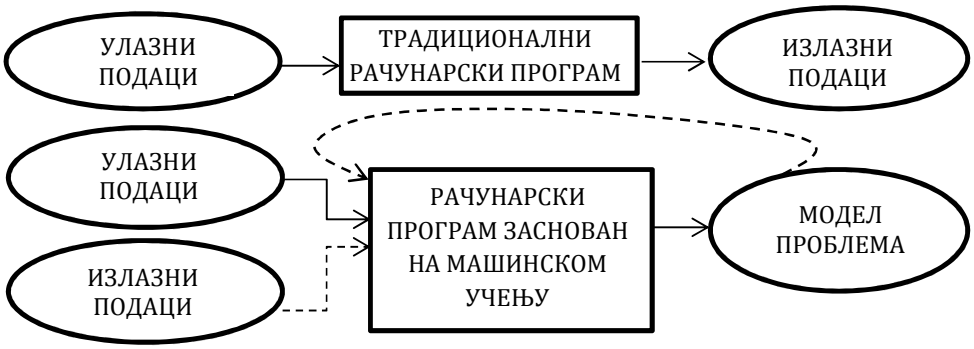
- дедуктивно закључивање
- индуктивно закључивање.

Дедуктивно закључивање (специјализација) засновано је на правилима математичке логике и подразумева процес у коме се знање примењено на решавање неког скупа проблема преноси и на све његове подскупове. Индуктивно закључивање, или генерализација, заснива се на претходном поступку апстракције неких особина неких надскупова проблема, како би било могуће да се знања примењена на нижем нивоу (на неком од подскупова) могу пропагирати, односно пренети на надскуп. Апстракција и

генерализација представљају кључне компоненте сваког система машинског учења. Недостатак метода индуктивног закључивања је могућност генерисања грешака у закључивању, а предност је да су врло често, нарочито у случају обраде велике количине података, ефикасније од метода дедуктивног закључивања.

1.3 Методологије машинског учења

Сваки систем који учи захтева постојање скупа података тј. скупа за учење (енг. training set) из којих се генерише скуп атрибута и њихових вредности релевантних за проблем који се решава. Приликом формулације проблема дефинише се **циљна функција** (c) која је најчешће непозната и коју треба апроксимирати на најбољи начин самим процесом учења. Функција којом се апроксимира циљна функција назива се **хипотеза** (h) или модел проблема. Апроксимирајућа функција (хипотеза) бира се из скупа допустивих хипотеза који се назива **простор хипотеза** (*H*). Са рачунарског становишта, процес учења можемо схватити као претрагу простора хипотеза чији је резултат она хипотеза која даје најбољу апроксимацију циљне функције. Математички посматрано, процес учења своди се на проблем математичке оптимизације којим се проналази најбоља хипотеза у датом простору, уз задати скуп ограничења. На слици 1.2 представљен је однос компоненти традиционалног програмског система и система машинског учења.



Слика 1.2 Функционални однос компоненти класичног програмског система и система заснованог на машинском учењу

Зависно од тога да ли је у процес учења укључен скуп излазних података, односно да ли је познат скуп вредности циљне функције у тренутку имплементације система, као и да ли постоји повратна спрега система, модела и спољашње средине, могу се дефинисати три методологије машинског учења:

- надгледано учење (енг. supervised learning)
- ненадгледано учење (енг. unsupervised learning)
- учење са подстицајем (енг. reinforced learning).

Надгледано учење (Mitchel, 1997; Alpaydin, 2010; Russell & Norvig, 2011; Murphy, 2012; Murphy, 2012) обухвата скуп проблема и техника за њихово решавање који користи скуп улазних података и скуп излазних података (вредности циљне функције добијених применом над скупом улазних података). Рачунарски посматрано, може се рећи да методе надгледаног учења захтевају скуп улазних података такав да је сваком вектору улазних података придружена вредност, ознака или обележје (енг. label) које се може интерпретирати као вредност циљне функције. Задатак система је да „научи“ како да произвољном вектору података додели прихватљиво тачну излазну вредност или ознаку.

Ненадгледано учење (Mitchel, 1997; Alpaydin, 2010; Murphy, 2012) обухвата скуп проблема и техника за њихово решавање који користи само скуп улазних података. Циљна функција није позната, а систем има задатак да открије скривене законитости у улазним подацима. Најчешће примењиване технике ненадгледаног учења су груписање (енг. clustering), откривање изузетака (енг. outlier detection), скривени Марковљеви модели, методе раздвајања сигнала без помоћних информација (енг. blind signal separation), итд.

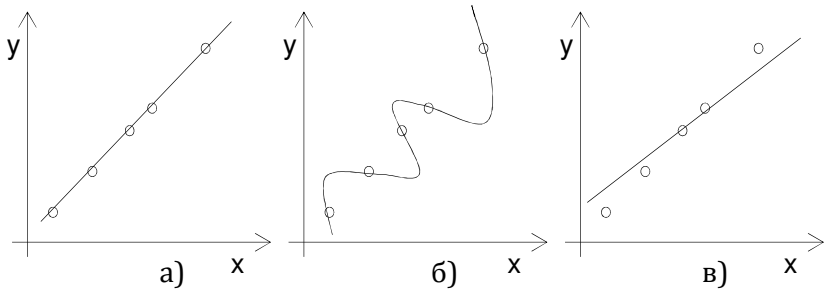
Учење са подстицајима (Alpaydin, 2010; Russell & Norvig, 2011; Murphy, 2012) укључује враћање повратних информација систему, које се интерпретирају као „награде“ или „казне“ у зависности од интеракције актуелног модела и задатог окружења. Циљ учења оваквих система је максимизација „награда“, односно минимизација „казни“ у времену у непознатом окружењу.

1.3.1 Надгледано машинско учење

Нека је дат скуп улазних података, односно скуп за обуку, од n примера парова улаз-излаз, облика:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

где је свако y_i резултат примене непознате функције $y = f(x)$ над улазним податком x_i. Задатак надгледаног учења је налажење оне функције h која представља најбољу апроксимацију непознате циљне функције f. Скуп хипотеза може садржати различите типове функција. На слици 1.3 дати су примери неких хипотеза за скуп од 5 парова (x_i, y_i) скупа за учење $(n = 5)$. Хипотеза је конзистентна ако садржи све дате парове података из скупа за учење, а неконзистентна у супротном. Може се претпоставити да је свака конзистентна хипотеза оно тражено решење проблема избора које најбоље апроксимира непознату циљну функцију f. Међутим, насупрот овој детерминистичкој претпоставци стоји принцип одабира хипотезе који се може свести на одабир најједноставније хипотезе, на основу принципа који је познат као Окамова оштрица[8].



Слика 1.3 Примери могућих хипотеза у простору хипотеза за дати проблем а) конзистентна хипотеза линеарне функције б) конзистентна хипотеза нелинеарне функције в) неконзистентна хипотеза линеарне функције

[8] Ockham's razor (лат. lex parsimoniae) принцип решавања проблема који се заснива на учењу филозофа Вилијама од Окама (William of Ockham) у XIV веку и који из скупа хипотеза за решење неког проблема узима ону која је најједноставнија.

Узимајући у обзир ова два супротстављена принципа, долазимо до хеуристике која представља оптимизацију између постизања максималне конзистенције у случају познатих обележја (сложености модела) и најједноставнијих претпоставки у свим осталим случајевима (једноставности модела). Овај услов долази из потребе да хипотезу треба бирати тако да буде довољно тачна и за све друге парове изван понуђеног скупа за учење. Може се десити да конзистентна хипотеза буде нефлексибилна у односу на те непознате вредности. Овакав проблем називамо проблемом превише прилагођеног модела (енг. overfitting) - хипотеза је конзистентна или, пак, даје најбоље могуће резултате у скупу за учење, али резултати које постиже ван тог скупа нису прихватљиви. Зато, у општем случају, правимо компромис између сложености хипотезе која је добро прилагођена подацима обучавања и једноставније хипотезе која боље генерализује тј. уопштава (исправно предвиђа изван скупа за обуку). Некада нисмо сигурни да ли посматрани простор хипотеза садржи праву функцију, односно да ли је проблем учења остварив. Зато, врло често говоримо о избору оне h^* хипотезе из простора хипотеза H која је највероватнија у односу на дате податке скупа за учење ($data$):

$$h^* = \underset{h \in H}{argmah}\,(P(h|data))$$

Можемо рећи да хипотезе могу бити детерминистичке и тада оне, у односу на скуп за учење, могу бити конзистентне или неконзистентне, или могу бити вероватносне, и тада, у односу на скуп за учење, могу бити мање или више вероватне. Сходно томе и алгоритми надгледаног машинског учења могу бити невероватносни:

- алгоритми засновани на стаблима одлучивања (енг. Decision-tree algorithms) (Mitchel, 1997; Alpaydin, 2010; Murphy, 2012);
- алгоритми засновани на примерима, нпр. алгоритам к-најближих суседа (енг. k-Nearest Neighbors algorithm) (Manning, Raghavan & Schütze, 2008; Alpaydin, 2010; Murphy, 2012);
- алгоритми потпорних вектора (енг. Support Vector Machines algorithm) (Cristianini & Shawe-Taylor, 2000; Alpaydin, 2010; Murphy, 2012);

- алгоритми засновани на вештачким неуронским мрежама (енг. Artificial Neural Networks), нпр. алгоритам перцептрон (Mitchel, 1997; Alpaydin, 2010; Murphy, 2012);

или вероватносни:

- Марковљеви модели максималне ентропије (енг. Maximum Entropy Markov Model - MEMM) (McCallum, Freitag & Pereira, 2000; Jurafsky & Martin, 2009; Alpaydin, 2010; Murphy, 2012);
- алгоритми засновани на Бајесовом правилу (енг. Naive Bayes algorithms) (Jurafsky & Martin, 2009; Murphy, 2012;Gamallo & Garcia, 2014);
- алгоритми засновани на регресионој анализи (енг. Regression Analysis algorithms) (Jurafsky & Martin, 2009; Hastie, Tibshirani & Friedman, 2011; Murphy, 2012);
- алгоритми условних случајних поља (енг. Conditional random fields -CRFs) (Lafferty, McCallum & Pereira, 2001).

Метода коју примењујемо у овој тези припада групи вероватносних алгоритама заснованих на регресионој анализи, па ће у наредним одељцима бити више речи о надгледаном машинском учењу, алгоритмима вероватносног надгледаног машинског учења, регресионој анализи као једној групи алгоритама из те врсте и њеној примени у процесу класификације текстова на основу осећања.

1.3.2 Надгледано машинско учење засновано на регресионој анализи

У машинском учењу постоје две методе којима се скуп за учење (вредности атрибута којима се описује посматрани процес) пресликава у неку излазну вредност. У случају када је та излазна вредност континуална (најчешће реалан број) реч је о регресији, а у случају дискретне вредности (обично се назива квалитативним излазом) говоримо о класификацији. Класификација и регресија (Murphy, 2012) могу узети у обзир и више од једног (енг. multiple regression) улазног атрибута, а такође могу бити и са

више излазних вредности (енг. multinomial regression), или са више класа у случају класификације.

Регресија[9] (регресиона анализа) је скуп статистичких метода помоћу којих се утврђује постојање функционалних зависности (закономерних веза) међу појавама или догађајима. Регресионом анализом можемо утврдити степен посматране функционалне зависности као и њен смер (да ли промене једне независне појаве или процеса утичу на варијације других у истом или супротном смеру). Стога, регресиона анализа као статистички алат, није применљива у појединачним мерењима. Међутим, у већем броју понављања неке појаве или догађаја може се утврдити просечна зависност међу њима. Најчешће се посматране појаве представљају помоћу скупова карактеристичних атрибута (променљивих) и затим се испитује просечна зависност излазне (одзивне) променљиве у односу на једну или више независних променљивих (предиктора). Математички формулисана, просечна зависност представља регресиону функцију. Зависно од природе функције, регресија може бити линеарна или нелинеарна. Међутим, може се десити да регресиона функција нема аналитички облик, односно да предиктори нису искључиво непрекидне променљиве, већ могу бити и дискретне. Стога, у општем случају говоримо о регресионом моделу. Ако и одзивна променљива има дискретну (најчешће дихотомну) природу, онда говоримо о класификационом моделу.

1.4 Модели репрезентације докумената

У основи обраде сваког дигиталног документа лежи израда његовог модела репрезентације. Када је реч о дигиталном текстуалном документу онда се може говорити о три основна модела репрезентације у области проналажења информација (IR) и то су:

- логички модел – настао раних 50-тих, проширен радовима (van Rijsbergen, 1979; Salton, Fox & Wu, 1983)

[9] Појам регресије као и закон о регресији (лат. regresio - узвраћање) увео је крајем XIX века сер Галтон (F. Galton), енглески природњак, истражујући наследне особине живих бића.

- модел векторског простора (Salton, Wong & Yang, 1975; Turney & Pantel, 2010)
- вероватносни модел (Maron & Kuhns, 1960; Goodman, 2001; Zhai, 2009).

Логички модел је једноставан систем репрезентације текста заснован на његовом представљању помоћу коначног скупа међусобно независних речи или група речи – термина (енг. terms), садржаних у посматраном тексту. У овом моделу није релевантна позиција речи у тексту. Једини релевантан параметар у односу на посматрану реч је индикатор њеног појављивања у датом тексту. Добра страна овако усвојене представе је поједностављење алгоритама у области проналажења информација. Лакоћа постављања упита логичким изразима у којима се користе појмови и логички оператори AND, OR, NOT овај модел чини још увек доминантним у комерцијалним библиографским системима и неким машинама за претрагу. С друге стране, овакви модели тешко се боре са различитим морфолошким варијантама, са проблемима вишезначности и са високом непрецизношћу из разлога што се документи не рангирају, а стратегија претраживања заснива се на бинарном критеријуму одлучивања (документ је релевантан, документ није релевантан).

Модел векторског простора почео је да се развија на Универзитету Корнел (енг. Cornell University) седамдесетих година прошлог века и он уводи вектор као начин репрезентације сваког појединачног документа.
Посматрајмо, на пример, скуп *D* од *m* докумената:

$$D = (d_1, d_2, \dots, d_m)$$

где је сваки документ d_i идентификован са n установљених атрибута – (најчешће међусобно независних речи или група речи)

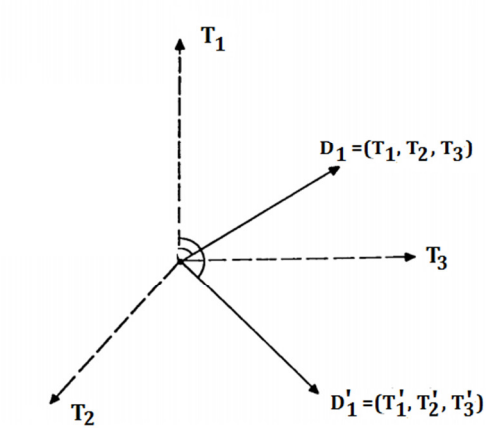
$$d_i = (t_{i1}, t_{i2}, \dots, t_{in})$$

где је t_{ij} тежина *j*-тог атрибута *i*-тог документа (најчешће мерена бројем појављивања тог атрибута у датом документу).

N-димензионални векторски простор, чије димензије представљају установљени атрибути, је модел у коме је сваки документ представљен вектором чији је почетак у координатном почетку, а дужина

$$|d_i| = \sqrt{\sum_{j=1}^{n} {t_{ij}}^2}$$

На слици 1.4 приказан је пример векторског простора који садржи 2 документа, који су представљени са 3 атрибута.



Слика 1.4 Пример модела векторског простора докумената

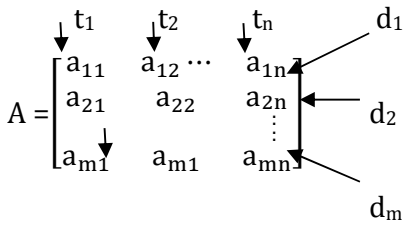
Модел векторског простора користи се у процесима проналажења и филтрирања информација и код претраге веба. Даје добре резултате (Turney & Pantel, 2010) у алгоритмима мерења сличности семантичке оријентације[10] на нивоу речи, фраза и докумената и алгоритмима машина за претрагу за утврђивање степена семантичке сличности[11] између корисничког упита и докумената коришћењем тзв. технологије инвертовања индекса и рангирања чија је суштина да се пронађу они документи у којима се највећи број пута употребио појам који је предмет претраге. Слабости примене ове методе су значајно процесорско време и неопходност рекалкулације свих вектора у случају промене димензије векторског

[10] Семантичка оријентација (енг. акроним SO) је мера субјективности неког текста (Taboada, Brooke, Tofiloski, Voll & Stede, 2011).

[11] Семантичка сличност је мера дефинисана између два концепта или двеју текстуалних целина којом се изражава степен сличности њиховог значења или семантичког садржаја (Bollegala, Matsuo & Ishizuka, 2009).

простора. Такође, овај модел претпоставља међусобну независност атрибута и не узима у обзир појаву полисемије (атрибут може имати различита значења у различитим контекстима) и синонимије (различити атрибути могу имати исто значење).

Уколико се документи представљају скупом атрибута тј. термина, а структура коју граде није вектор већ матрица, онда говоримо о матрици термина и докумената (енг. term-document matrix) где је сваки документ - један ред матрице, свака колона - један атрибут, а вредност сваког елемената матрице једнака је броју појављивања датог појма у датом документу.



Главна слабост претходно описаних, тзв. традиционалних, модела репрезентације докумената је искључивост. Репрезентативни атрибути које описују неки документ представљају ограничен скуп, а резултати претраге, у том случају, дају дисјунктне скупове у којима или постоји или не постоји дати документ, односно он је или пронађен или није пронађен. Побољшање традиционалне технологије дошло је у виду увођења оцене степена повезаности документа и задатог услова. За сваки документ из посматране колекције се израчунава вероватноћа да ће дати документ бити у резултујућем скупу. Пошто се ради о везаним догађајима, вероватноћа се добија као резултат примене Бајесових (енг. Bayes) формула.

Вероватносни језички модел или краће - језички модел (енг. Language Model – *LM*) заснива се на концептима речи и реченица, где је свака реченица уређени низ речи, а свака реч припада унапред дефинисаном скупу који се назива речник. Тада се језички модел дефинише речником и дистрибуцијом вероватноћа појављивања речи дефинисаних у речнику.

Када се моделује неки текст језичким моделом, тада се, у општем случају, свакој речи ω_i у документу, у низу од m речи, додељује вероватноћа $P(\omega_i|\omega_{i-1},\dots,\omega_1)$ помоћу расподеле вероватноћа свих претходних речи у секвенци (Goodman, 2001), (Zhai, 2009).

Вероватноћа $P(\omega_1, \omega_2, \omega_3,..., \omega_m)$ документа који је дат низом речи $\omega_1,\omega_2,\omega_3,...,\omega_m$ апроксимира се производом условних вероватноћа[12]

$$P(\omega_1, \omega_2, \omega_3, \dots, \omega_m) = \prod_{i=1}^{m} P(\omega_i|\omega_1, \dots, \omega_{i-1}) .$$

Вероватносни језички модели користе се у системима вероватносне обраде текста. Како језички модел моделује сваки документ понаособ, примена оваквог модела над већим корпусом је готово немогућа, па се примењују различите методе апроксимације (нпр. *n*-грамски модели) (Manning, Raghavan & Schütze, 2008). Најједноставнији *n*-грамски језички модел је модел униграма који подразумева да појављивање једне речи не зависи од појављивања других, односно да је дужина уређеног низа речи који се посматра *n=1*. У том случају се дистрибуција вероватноћа даје на нивоу појединачних речи. Ако претпоставимо да је V речник униграма, за модел униграма θ важи:

$$\sum_{\omega \in V} p(\omega|\,\theta) = 1.$$

1.4.1 N-грамски модел језика

Појам *n*-грама (Manning & Schütze, 1999) увео је Клод Шенон[13] у свом раду о теорији информација 1948. године истражујући вероватноћу појављивања следећег слова у датој секвенци карактера. Он је показао да се у датој секвенци од n карактера, у n-грамском моделу, вероватноћа сваког карактера може изразити помоћу вероватноћа $(n-1)$ претходних карактера у документу. Такође, вероватноћа $P(\omega_1, \omega_2, \omega_3, \dots, \omega_m)$ секвенцијалног низа речи $\omega_1, \omega_2, \omega_3,..., \omega_m$ апроксимира се производом условних вероватноћа

[12] Због чега се *језички модел*, према (Zhai, 2009), назива *генеративни модел текста*.
[13] Claude Shannon

$$P(\omega_1, \omega_2, \omega_3, \ldots, \omega_m) = \prod_{i=1}^{m} P(\omega_i | \omega_{i-1}, \ldots, \omega_{i-n+1})$$

То значи да је вероватноћа i-те речи ω_i зависна од вероватноћа претходних $(n - 1)$ речи. На пример, за $n = 3$ и реченицу: „Ја волим да једем салату“, где је <s> ознака почетка реченице, вероватноћа реченице је:

P(Ја, волим, да, једем, салату) ≈ P(Ја | <s>, <s>) * P(волим | <s>, Ја) *

P(да | Ја, волим) * P(једем | волим, да) * P(салату | да, једем).

У процесима обраде текста, n-грам је ниска од n елемената, који могу бити слова, слогови, речи или веће целине. У техникама машинског учења којима се генеришу различити модели класификације текста, често се користе n-грамски профили, тј. уређени парови облика (n-грам, фреквенција_ n-грама) (Graovac, 2012, 2014).

Задаци класификације текста такође користе језичке моделе. Најчешће коришћени модели репрезентације овог типа су: n-грамски модели на нивоу карактера и n-грамски модели на нивоу речи. И један и други тип, зависно од броја елементарних јединица које користе, најчешће су модели нивоа: 1(unigrams), 2(bigrams), 3(trigrams) или 4(four-grams). Модели векторске репрезентације и n-грамски модели носе заједнички назив – модели вреће токена (енг. bag-of-tokens models) (Giannakopoulos, Mavridi, Paliouras, Papadakis & Tserpes, 2012). У задацима класификације текстова они се користе за груписање међусобно сличних токена у посматраном документу, на основу унапред одабране мере сличности, тако да на најбољи начин репрезентују (генеришу) или одвајају (дискриминишу) посматране документе (в. одељак 2.3).

У овом раду биће приказана метода машинског учења, из класе вероватносних регресионих метода, којим се генерише модел класификације текста чије предикторе представљају n-грами речи где је $n \leq 3$, а зависна променљива је дихотомна величина која може узети вредности из скупа

{„позитиван поларитет осећања“, „негативан поларитет осећања“}.

2. Класификација

Класификација је метода изградње математичког модела којим се, помоћу низа предиктора једне врсте појава или ентитета и скупа унапред дефинисаних класа, утврђује којој од датих класа нека инстанца те појаве или ентитета припада. У фази изградње модела класификације, а такође и касније, у самој класификацији, ентитет за који се гради класификатор се мора уопштити помоћу коначног скупа својстава (енг. features)[14] релевантних за процес класификације (предиктори класификације). Стога, можемо рећи да класификација, у ширем смислу, подразумева три подпроцеса и то су:

- процес уопштавања објекта класификације, односно селекција релевантних предиктора
- процес генерисања класификационог модела на основу проучавања понашања познатих инстанци објекта класификације
- примена класификационог модела над непознатим инстанцама објекта класификације.

Процес селекције релевантних предиктора (Tang, Alelyani & Liu, 2014; Hastie, Tibshirani & Friedman, 2011) састоји се у оцени и одбацивању редундантних, нерелевантних и међусобно зависних својстава како би се добио репрезентативан скуп својстава који на оптималан начин репрезентује посматрани објекат у контексту класификационих услова. О методама селекције предиктора и оценама селекције оптималног скупа биће речи у поглављу 5.

Процес генерисања класификационог модела (Nigam, Lafferty & McCallum, 1999) је итеративан поступак у коме се долази до квантитативних оцена

[14] У областима машинског учења и препознавања образаца, појам *feature* односи се на појединачно мерљиво својство неке појаве, догађаја или објекта који су предмет проучавања. У методама машинског учења реч је о независној нумеричкој или дихотомној променљивој која учествује у генерисању модела. У методама које се користе у лексичко-семантичким методама класификације, појам *feature* може бити садржај текста или неког другог формата, као и неко њихово квалитативно својство. У овој тези овај појам биће коришћен у оба значења, што ће бити наглашено.

(тежина) како предиктори појединачно утичу на исход класификације. О методама генерисања биће речи у наставку овог поглавља. И најзад, добијени класификациони модел може се применити над произвољном инстанцом посматраног ентитета, односно скупом својстава којим је таква инстанца представљена, како би се оценила вредност излазних параметара класификације, тј. класа којој дата инстанца припада.

2.1 Класификација текста

Класификација текста представља подврсту општије методе, познате под именом *класификација,* у којој је предмет обраде текстуални документ или неки његов део. Постоји више различитих дефиниција класификације текста. Према Менингу[15] (Manning, Raghavan & Schütze, 2008, pp. 253):

Дефиниција 2.1 *Класификација текста је одређивање класе којој неки документ припада, ако је унапред познат скуп класа*.

Формална дефиниција коју такође даје Менинг (Manning, Raghavan & Schütze, 2008, pp. 256) претпоставља дефиницију класификације методом машинског учења:

Дефиниција 2.2 *Класификација текста је функција пресликавања текстуалног документа у одговарајућу класу*

$$\gamma: X \rightarrow C$$

за унапред познат тренинг скуп $< d, c > \in X \times C$.

Формална дефиниција Себастијанија[16] (Sebastiani, 2002, pp. 3) користи појам категоризације и даје дефиницију којом се појмови класификације и категоризације поистовећују:

Дефиниција 2.3 *Категоризација текста је задатак доделе логичке вредности сваком пару* $< d_j, c_i > \in D \times C$ *, где је D скуп докумената, а C је скуп предефинисаних категорија. Вредност T (тачно) додељена пару* $< d_j, c_i >$

[15] Christopher D. Manning
[16] Fabrizio Sebastiani

означава одлуку да се документ d_j сврста у категорију c_i, а вредност F (нетачно) означава одлуку да се документ d_j не сврста у категорију c_i.

Себастијани даље даје још формалнију дефиницију (Sebastiani, 2002, pp. 3):

Дефиниција 2.4 *Категоризација текста је апроксимација непознате циљне функције* $\breve{\boldsymbol{\Phi}}: \boldsymbol{D} \times \boldsymbol{C} \rightarrow \{\boldsymbol{T}, \boldsymbol{F}\}$ *хипотезом (класификатором)* $\boldsymbol{\Phi}: \boldsymbol{D} \times \boldsymbol{C} \rightarrow \{\boldsymbol{T}, \boldsymbol{F}\}$ *таква да је* $\boldsymbol{\Phi}$ најбоља апроксимација непознате циљне функције $\breve{\boldsymbol{\Phi}}$.

Себастијани, као и Менинг, не прави формалну разлику између појмова класификације и категоризације. Напротив, Менинг (Manning, Raghavan & Schütze, 2008, pp. 253) сматра да се у класификацији текстова подједнако могу користити изрази: „text classification", „text categorization", „topic classification", „topic spotting", а да се појам „класа" равноправно замењује појмом „тема" („topic"). С друге стране, постоје истраживања (Jacob, 1991, 2004; Glushko, 2013) која указују на разлике у значењу, а нарочито у методама за решавање проблема који се дефинишу као класификација и категоризација. Према Јакобу[17] (Jacob, 2004):

Дефиниција 3.1 *Категоризација је процес поделе у групе ентитета (категорије) такве да су чланови исте групе слични један другом на известан начин, односно деле неке сличности које су видљиве унутар датог контекста.*

Дефиниција 3.2 *Класификација је процес уређене и систематичне доделе сваког ентитета једној и само једној класи унутар система међусобно искључивих и непреклапајућих класа који се спроводи у складу са утврђеним скупом принципа који регулишу структуру класа и класне односе уз обавезу доследне примене ових принципа.*

Јакоб даље идентификује 6 системских својстава на основу којих се може утврдити да ли је један систем класификациони или категоризациони. То су: 1. процес (process), 2. границе (boundaries), 3. припадност (membership), 4. критеријуми доделе (criteria for assignment), 5. типизација (typicality), 6. структура (structure).

[17] Elin K. Jacob

Уколико је један систем категоризациони, процес разврставања врши се на основу утврђивања сличности са осталим познатим члановима или узорцима категорије, у односу на неки контекст. Границе међу категоријама су меке и преклапајуће, а припадност већем броју категорија није искључена. Критеријуми доделе морају бити мерљиви и формирају се на основу неког уопштеног знања, а сви чланови групе не репрезентују подједнако категорију. Обично не постоји хијерархија класа и ако се гради, ради се о груписању у неке видове хијерархијских структура.

Уколико је један систем класификациони, процес разврставања врши се на основу анализе потребних и довољних услова припадања класи. Класе су међусобно искључиве, не преклапају се и непроменљиве су. Припадност класи је такође искључива (припада или не припада). Критеријуми доделе су унапред познати, а сви чланови подједнако репрезентују класу. Уколико постоји хијерархија класа, она је фиксна и предефинисана.

Уколико усвојимо Јакобову анализу, можемо рећи да су методе ненадгледаног машинског учења типични представници система категоризације[18]. Систем груписања није унапред познат, или, уколико је познат, није стриктно дефинисан. Он се генерише на основу примене општих знања о датом проблему, а документи се аутоматски групишу (енг. clustering) оценом растојања које представља меру сличности између двају ентитета[19]. Методе надгледаног машинског учења типични су представници система класификације. Класе и правила о припадности класама су једнозначно и унапред дефинисани усвајањем јединственог модела класификације, па су границе фиксне и не преклапају се. Јединствен модел класификације може бити дат у облику скупа правила као код методе *стабла одлуке* (decision tree), језгрене функције као код методе *потпорних вектора* (support vector machine) или вероватносне функције као код вероватносних класификатора и др.

[18] У обради природног језика (енг. NLP) уместо појма категоризација користе се појмови груписање или кластеровање (енг. "clustering").
[19] Пример примене методе, у задатку препознавања језика којим је текст написан, може се наћи код Јаничића и Николића, (Janičić & Nikolić, 2010, одељак 11.1.2)

2.2 Класификација докумената према осећању

Класификација докумената према осећању (анализа осећања израженог у тексту) је специфичан вид класификације текстуалних докумената где излазна променљива узима вредност из скупа од две дискретне вредности {“позитивно” , “негативно”}. Како у класификацији докумената на основу осећања постоје 2 класе, реч је о бинарном класификатору, а све методе које се користе у задацима машинског учења ради класификације текстова могу бити коришћене и у класификацији докумената на основу осећања. Једну од општих дефиниција анализе осећања (SA) даје Лиу[20] (Liu, 2010, pp. 629):

Дефиниција 4.1 *Анализа осећања и истраживање мишљења представљају поље истраживања рачунарских наука које се бави проучавањем осећања и мишљења изражених у неком тексту.*

Општа формална дефиниција изводи се из дефиниције општег класификационог проблема:

Дефиниција 4.2 *Ако са* $X = \{x_1, x_2, \dots, x_m\}$ *обележимо скуп предиктора документа d, а са* $C = \{c_1, c_2\}$ *скуп класа којима се врши класификација према осећањима, онда је класификатор докумената на основу осећања дефинисан пресликавањем* f*, таквим да је:*

$f\ (c, x)$: $C \times X \ \rightarrow \{\ \mathrm{T}, \mathrm{F}\}$,

бинарна функција која узима вредност „T“ ако се посматраним документом изражавају позитивна осећања, а вредност „F“ у супротном (уколико се посматраним документом изражавају негативна осећања), тј.

$$f\ (c, x) = \begin{cases} T, & \text{документ } x \text{ носи позитивни поларитет осећања} \\ F, & \text{документ } x \text{ носи негативни поларитет осећања} \end{cases}$$

Формалну дефиницију анализе осећања која обухвата и истраживање мишљења (opinion mining), а која се не изводи из дефиниције општег

[20] Bing Liu

класификационог проблема, такође даје Лиу (Liu, 2010, pp. 631-633). Он најпре уводи дефиниције *ентитета* (*entity*), *компоненте* (*component*), *атрибута* (*attribute*), *својства* (*feature*),[21] *предмета осећања, оцене или мишљења* (*opinion target*) *и носиоца осећања или мишљења* (*opinion holder*):

Дефиниција 4.3.1 *Ентитет је производ, особа, организација, тема или догађај, представљен хијерархијом својих компоненти и подкомпоненти.*

Дефиниција 4.3.2 *Свака компонента једног ентитета је скуп уређених парова <атрибут, вредност_атрибута>.*

Дефиниција 4.3.3 *Својство је компонента или атрибут.*

Дефиниција 4.3.4 *Носилац осећања, оцене или мишљења је особа која изражава своје осећање, оцену или мишљење.*

Дефиниција 4.3.5 *Предмет осећања, оцене или мишљења је ентитет са свим његовим својствима.*

Дефиниција 4.3.6 *Поларитет осећања носиоца осећања или мишљења у односу на својство посматраног ентитета.*

На основу претходних дефиниција, Лиу и сарадници дају дефиницију анализе осећања као уређене петорке**:**

Дефиниција 4.3 *Осећање (емоција, сентимент), оцена или мишљење је уређена петорка* $< \mathrm{e}_j, \mathrm{a}_{jk}, so_{ijkl}, h_i, t_l >$ *таква да је:*

e_j – ентитет који је предмет исказаног осећања, оцене или мишљења,

a_{jk} – к-то својство ентитета e_j

so_{ijkl} – је поларитет осећања или мишљења којим носилац осећања оцењује својство a_{jk} , ентитета e_j , у тренутку t_l

[21] Формална дефиниција Лија дата је са становишта лексичко-семантичких метода класификације, па се појам *feature* схвата као битна квалитативна особина или својство посматраног објекта.

h_i – носилац осећања

t_l – време када се исказује осећање или даје оцена или мишљење.

Анализа осећања, као класификациони проблем, може се реализовати како методама ненадгледаног, тако и методама надгледаног учења. У овом раду је изграђен и оцењен систем за класификацију текстова на основу осећања SAFOS, методом надгледаног машинског учења која имплементира вероватносни регресиони класификатор, о чему ће детаљно бити речи у поглављу 5.

2.3 Вероватносни класификатори

Нека је дат коначан скуп Ω могућих исхода неког експеримента, који, иначе, називамо *простор елементарних догађаја* и нека је случајан догађај E његов подскуп. Нека је дата функција

$p: \Omega \rightarrow [0,1]$

таква да је $\sum_{\omega \in \Omega} p(\omega) = 1$.

Тада је вероватноћа догађаја $E \subset \Omega$ дефинисана са

$P(E) = \sum_{\omega \in E} p(\omega)$,

а уређена тројка $(\Omega, \mathrm{E}, \mathrm{P})$ представља *вероватносни модел* или *простор вероватноћа*.

У проблемима бинарне класификације, променљива која представља класу може узети једну од две могуће вредности, а у случајевима вероватносне бинарне класификације можемо говорити само о вероватноћи да променљива узме једну од две могуће вредности из скупа {0,1}, што означавамо са:

P(C=1), P(C=0),

и представља вероватноћу која означава знање о резултату класификације пре било каквог оцењивања (још се зове и *почетна вероватноћа,* „a priori“ или на енглеском *prior probability*). Када немамо почетних знања, онда су обе класе једнако вероватне, тј важи:

P(C=1) = P(C=0)=0,5.

Вероватносни класификатор је простор вероватноћа (назива се и класификациони модел) који на основу расподеле вероватноћа улазних променљивих или предиктора (атрибута посматраног ентитета) даје расподелу вероватноћа припадности посматраног ентитета скупу датих класа. Другим речима, вероватносни класификатор је дистрибуција условних вероватноћа $P(C|X)$, односно расподела вероватноћа да посматрани ентитет X буде класификован у сваку од класа из коначног скупа класа C. У случају бинарног класификационог модела, имамо расподелу облика:

$$P(C = 1|X),\ P(C = 0|X)$$

а на основу Бајесовог правила и условних вероватноћа (*likelihood*)

$$P(X|C = 1),\ P(X|C = 0)$$

можемо изразити „*a постериори*“ вероватноће да се посматрани ентитет класификује у једну ($C = 1$) или другу класу ($C = 0$) помоћу:

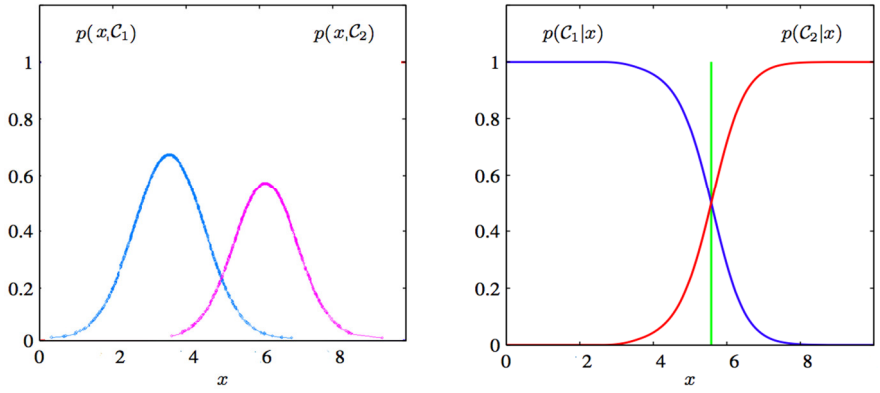
$$\hat{y} = P(C = 1|X) = \frac{p(X|C=1)P(C=1)}{p(X)},\ 1 - \hat{y} = P(C = 0|X) = \frac{p(X|C=0)P(C=0)}{p(X)}$$

У том случају, класификација произвољне инстанце x_i ентитета X дефинише се као она класа $\hat{y}^*$ која максимизује „*a постериори*“ (назива се још МАП услов – *maximum a posteriori probability*) вероватноћу $\hat{y}$ да класификатором оцењена класа инстанце x_i буде баш она којој x_i заиста припада (y):

$$\hat{y}^*_{MAP} = argmax_{\hat{y}} P(C = y|x_i\ )$$

Вероватносни класификатори се деле на генеративне и дискриминативне (Ng & Jordan, 2002). Генеративни класификатори (Su & Srihari, 2011) генеришу модел на основу заједничке вероватноће $p(x, c)$ улазне променљиве x и класе c. Процесом учења моделују се дистрибуције класа. Најпознатији алгоритми ове групе су: „наивни“ Бајесов класификатор, скривени Марковљев модел, n-грамски модел, итд.

Дискриминативни класификатори генеришу модел апостериорних вероватноћа $p(c|x)$ тежећи да установе што веће разлике између класа. Процесом учења моделују се границе између класа. Типични алгоритми ове групе су: метода потпорних вектора, регресиони модели, метода условних случајних поља, метода најближих суседа, итд. На слици 2.1 приказане су дистрибуције вероватноћа генеративног и дискриминативног модела, у случају бинарног класификатора са једним предиктором.



а) б)

Слика 2.1 Дистрибуције вероватноћа: а) генеративног б) дискриминативног класификационог модела над истим скупом независних променљивих

Поређење ових метода можемо размотрити на задатку препознавања језика којим је текст написан (Janičić & Nikolić, 2010, одељак 11.1.2). Код генеративог приступа овом проблему потребно је научити сваки од језика и

на основу тога препознати језик непознатог текста, а код дискриминативног приступа потребно је научити разлике између језика, али не и саме језике[22].

Дискриминативни класификатори најчешће постижу виши степен тачности класификације, мање су осетљиви на грешке у подацима за обуку и процес њихове евалуације је бржи. Такође, добро решавају задатке класификације секвенцијалних структура. С друге стране, генеративни класификатори се брже обучавају и лакше је разумевање утицаја предиктора. Такође, дају боље резултате на малим скуповима за обуку (Murphy, 2012).

2.4 Модел линеарне регресије у класификацији докумената

Регресиона анализа представља класу метода машинског учења којима се, између осталог, успешно решавају проблеми класификације текста. Најједноставнији и најчешће примењивани регресиони модел је линеарна регресија. Размотрићемо је на проблему моделирања процеса предвиђања продајне цене стана на основу скупа за учење добијеног на основу онлајн огласа за продају некретнина. Дакле, желимо да моделујемо систем за предвиђање продајне цене неког стана и при том посматрамо једну променљиву (предиктор) коју ћемо назвати „број епитета којим се стан описује".

Узећемо у разматрање, за потребе скупа за учење, низ огласа о продатим становима какав је следећи пример:

„Продаје се фантастичан стан, на атрактивној локацији, са изузетним, функционалним распоредом".

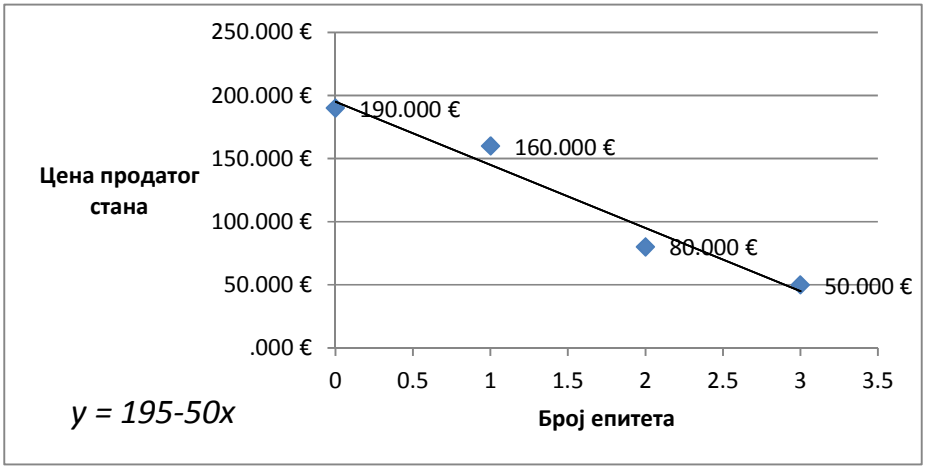
Анализом продатих станова, за сваки од њих добија се корелација између броја епитета и вредности по којој је стан продат. У складу са тим, а примера ради, овде наводимо табелу 2.1 која ће представљати скуп за учење добијен из те корелације:

[22] Ову аналогију дао је Срихари (Sargur Srihari) http://www.cedar.buffalo.edu/~srihari/CSE574/

Табела 2.1 Скуп за учење представљен подацима о броју епитета у огласу на основу кога је стан продат и вредности тог стана

Број епитета	Цена продатог стана
3	50.000€
2	80.000€
1	160.000€
0	190.000€

Резултате можемо приказати и графички, у правоуглом координатном систему, где свака тачка представља један пример (продат стан) из скупа за учење (слика 2.2).



Слика 2.2 Графичко приказ скупа за учење од једног предиктора и линеарне апроксимације тог скупа

Линеарни модел (линеарна апроксимација скупа тачака добијених из скупа за учење) дат једначином праве линије, општег облика $y = a + bx$, у посматраном примеру је:

$$y = 195 - 50x.$$

У статистици, а такође и рачунарској лингвистици, променљива x је посматрано својство и у овом примеру је предиктор процеса продаје стана. Коефицијенти a и b су тежине (енг. weights). У датом примеру, линеарни модел је вектор тежина

$$\vec{w} = (195, -50)$$

Овако генерисан модел можемо користити за процену цене по којој се може продати понуђени стан, уколико узмемо у обзир дату променљиву, односно предиктор „број епитета“. На пример, можемо проценити вредност стана наведеног у горњем примеру и описаног са четири епитета.

У процесу моделирања система уобичајено је коришћење већег броја предиктора. Нпр. можемо претпоставити да на цену утичу и укупан број станова у понуди у датом делу града и висина каматне стопе стамбеног кредита. Тада би линеарни модел била линеарна функција више променљивих:

$$cena = w0 + w1 * broj\_epiteta + w2 * stopa + w3 * broj\_stanova\_u\_ponudi$$

У општем случају претпоставићемо да је дат скуп за учење X са m исхода[23] неког догађаја помоћу скупа парова

$$(X, y_1), (X, y_2), \dots, (X, y_m)$$

где је y_i исход i-тог извођења догађаја X. Ако X представимо помоћу n репрезентативних својстава (features),[24] односно помоћу вектора предиктора $\vec{\mathrm{X}} = (x_1, x_2, \dots x_n)^T$, једна репрезентација скупа за учење је матрица $(n+1) \cdot m$ облика:

$$\begin{matrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} & y^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} & y^{(2)} \\ . & . & . & . & \\ . & . & . & . & \\ . & . & . & . & \\ x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} & y^{(m)} \end{matrix}$$

Геометријски посматрано, ради се о скупу од m тачака у $n-$димензионалном векторском простору. Претпоставимо, такође, да је дат вектор тежина

[23] Основна претпоставка у статистици и машинском учењу је да су све инстанце скупа за учење из исте заједничке дистрибуције $p(x, y)$ и да су узорковане независно (енг. independent and identically distributed - i.i.d.)

[24] Овде се појам *feature* схвата као независна нумеричка или дихотомна променљива која учествује у генерисању модела.

$\overrightarrow{W} = (w_1, w_2, \dots w_n)^T$ којима се мери утицај[25] сваког од n предиктора инстанце x скупа за учење на коначан исход y. Тада можемо рећи да важи:

$$\mathrm{y}^{(j)} = \sum_{\mathrm{i}=1}^{\mathrm{n}} x_i^{(j)} \; w_i^{(j)}$$

односно, да је свака инстанца скупа за учење скаларни производ вектора предиктора и вектора тежина у тој инстанци

$$y^{(j)} = \overrightarrow{x^{(j)}} \cdot \overrightarrow{w^{(j)}}$$

С друге стране, ако посматрамо једначину праве аналитички, имамо, у случају линеарне зависности y од n променљивих:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

Уз претпоставку да је $x_0 = 1$, имаћемо

$$\mathrm{y} = \sum_{\mathrm{i}=0}^{\mathrm{n}} w_i \, x_i$$

а аналитички,

$$y = \vec{w}^{\,T} \cdot \vec{x}$$

На основу тога можемо дати дефиницију линеарне регресије из перспективе машинског учења.

Дефиниција 5.1 Линеарна регресија је онај модел скупа за учење који претпоставља да је исход скупа за учење линеарна комбинација вредности његових предиктора и одговарајућих тежина предиктора.

2.4.1 Процес учења у моделима линеарне регресије

Основне претпоставке (Janičić & Nikolić, 2010) регресионе анализе су да број чланова скупа за учење није мањи од броја предиктора и да случајна променљива која представља шум има нормалну расподелу, односно да величина грешке не зависи од величине излазне променљиве. Нека имамо

[25] Тежина $w_i^{(j)}$ показује за колико се јединица мења вредност $y^{(j)}$ у случају промене предиктора $x_i^{(j)}$ за 1.

скуп за учење од $j \in \{1,2,\dots,m\}$ посматрања за која знамо све улазне предикторе $x_i^{(j)}$, где је $i \in \{1,2,\dots,n\}$ број предиктора у сваком посматрању[26] и вредности излазних променљивих (посматраних вредности) $y_{obs}^{(j)}$, тако да важи $m \geq n$. Линеарни регресиони модел је облика

$$y_{pred}^{(j)} = \sum_{i=1}^{n} x_i^{(j)} * w_i^{(j)} \ , \quad j \in \{1,2,\dots,m\}$$

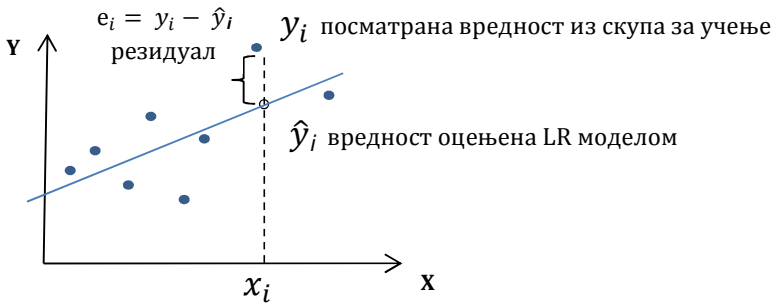
Основни проблем учења линеарног регресионог модела је како пронаћи вектор $\vec{w}$ тежина, такав да важи

$$y_{pred}^{(j)} = y_{obs}^{(j)} \ , \quad j \in \{1,2,\dots,m\}. \tag{1}$$

У случају важења једнакости (1) реч је о конзистентном моделу. Међутим, најчешће није могуће (када је $m \neq n$),[27] или није најоптималније наћи конзистентан модел (о чему је било речи у одељку 1.3.1). Стога се најчешће решење овог система тражи тако да се вредности $w_i^{(j)}$ оцене тако да вредности непознате функције $y_{pred}^{(j)}$ буду што је ближе могуће вредности циљне функције (вредности коју имамо у скупу за учење $y_{obs}^{(j)}$). Тада кажемо да је линеарна регресија над скупом X за учење она функција y_{pred} (често се означава и са $\hat{y}$) која представља најбољу линеарну апроксимацију функције y. Разлика између стварне и оцењене (претпостављене) вредности назива се резидуалом (слика 2.3) и означава се са e :

$$e^{(j)} = y^{(j)} - \hat{y}^{(j)} = y^{(j)} - \sum_{i=1}^{n} x_i^{(j)} \ w_i^{(j)}, \qquad j \in \{1,2,\dots,m\}$$

[26] У раду је усвојен уобичајени начин индексирања - горњи индекс означава индекс члана скупа за учење, а доњи означава индекс предиктора.
[27] Не добија се јединствено решење система линеарних једначина.

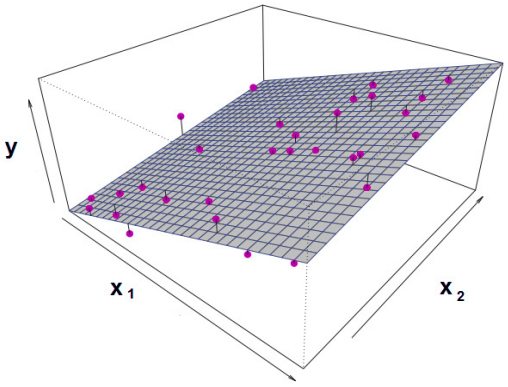


Слика 2.3 Стварна (посматрана) вредност циљне функције, моделом оцењена (претпостављена) вредност апроксимативне линеарне функције и њихов резидуал у једнодимензионалном линеарном регресионом моделу

Дакле, учење у LR моделу представља проналажење оног вектора тежина $\vec{w} = (w_1, w_2, \dots w_n)^{\mathrm{T}}$ којим се минимизира разлика између оцењених (претпостављених) вредности y_{pred} и стварних (посматраних) y_{obs} вредности из скупа за учење. Проблем налажења овог вектора један је од проблема конвексне оптимизације и назива се метода најмањих квадрата, јер се тражени вектор $\vec{w}$ може изразити као минимум суме квадрата резидуала:

$$\vec{w} = min \sum_{j=0}^{m} (y_{pred}^{(j)} - y_{obs}^{(j)})^2$$

На слици 2.4 дат је графички приказ линеарног регресионог модела са два предиктора где су посматране вредности y_{obs} скупа за учење дате тачкама лила боје, а одговарајуће вредности процењене моделом y_{pred} леже на регресионој равни.



Слика 2.4. Линеарни регресиони модел са два предиктора x_1 и x_2 који минимизира суму квадрата разлика скупа за учење

Решавање овог оптимизационог проблема показаћемо на примеру једнодимензионалног линеарног модела у скупу за учење од m тачака. Линеарни регресиони модел је права:

$$\hat{y}^{(j)} = a + bx^{(j)}\ ,\ \ j \in \{1,2,\dots,m\}$$

а сума квадрата резидуала:

$$Q_e{}^{(j)} = \sum_{j=0}^{m} (y^{(j)} - a - bx^{(j)})^2$$

За услов минимума $\frac{\partial Q_e}{\partial a} = \frac{\partial Q_e}{\partial b} = 0$, имамо:

$$\frac{\partial Q_e}{\partial a} = -2\sum_{j=0}^{m}(y^{(j)} - a - bx^{(j)}) = 0 \quad (2)$$

$$\frac{\partial Q_e}{\partial b} = -2x^{(j)}\sum_{j=0}^{m}(y^{(j)} - a - bx^{(j)}) = 0 \quad (3)$$

Из (2) и (3):

$$\sum_{j=0}^{m}\left(y^{(j)} - a - bx^{(j)}\right) = 0 \quad (4)$$

$$\sum_{j=0}^{m} x^{(j)}(y^{(j)} - a - bx^{(j)}) = 0 \quad (5)$$

Из (4) и (5):

$$\sum y^{(j)} - am - b\sum x^{(j)} = 0 \quad (6)$$

$$\sum x^{(j)}y^{(j)} - a\sum x^{(j)} - b\sum (x^{(j)})^2 = 0 \quad (7)$$

Из (6) и (7):

$$\sum y^{(j)} = am + b\sum x^{(j)} \quad (8)$$

$$\sum x^{(j)}y^{(j)} = a\sum x^{(j)} + b\sum (x^{(j)})^2 \quad (9)$$

Ако уведемо

$$\bar{X} = \frac{\sum x^{(j)}}{m} \quad (10)$$

$$\bar{Y} = \frac{\sum y^{(j)}}{m} \quad (11)$$

имаћемо:

$$\sum x^{(j)}y^{(j)} = m\bar{X}\sum y^{(j)} = m\bar{Y}\sum x^{(j)} \quad (12)$$

$$\bar{X}\sum y^{(j)} = \bar{Y}\sum x^{(j)} = m\bar{X}\bar{Y} \quad (13)$$

Из (8), (10) и (11):

$$a = \bar{Y} - b\bar{X} \quad (14)$$

Ако (14) заменимо у (9), имаћемо:

$$\sum x^{(j)}y^{(j)} = (\bar{Y} - b\bar{X})\sum x^{(j)} + b\sum (x^{(j)})^2 \quad (15)$$

$$\sum x^{(j)}y^{(j)} = \bar{Y}\sum x^{(j)} - b\bar{X}\sum x^{(j)} + b\sum (x^{(j)})^2 \quad (16)$$

Решавањем (16) по b, имаћемо:

$$b = \frac{\sum x^{(j)}y^{(j)} - \bar{Y}\sum x^{(j)}}{\sum (x^{(j)})^2 - \bar{X}\sum x^{(j)}} \quad (17)$$

Узимајући у обзир (10), бројилац се може трансформисати:

$$\sum x^{(j)}y^{(j)} - \bar{Y}\sum x^{(j)} = \sum x^{(j)}y^{(j)} - \bar{Y}\sum x^{(j)} - m\bar{X}\bar{Y} + m\bar{X}\bar{Y}$$

$$\sum x^{(j)}y^{(j)} - \bar{Y}\sum x^{(j)} = \sum x^{(j)}y^{(j)} - \bar{Y}\sum x^{(j)} - \bar{X}\sum y^{(j)} + m\bar{X}\bar{Y}$$

$$\sum x^{(j)}y^{(j)} - \bar{Y}\sum x^{(j)} = \sum (x^{(j)} - \bar{X})\,(y^{(j)} - \bar{Y}) \quad (18)$$

а именилац, такође узимајући у обзир (10):

$$\sum (x^{(j)})^2 - \bar{X}\sum x^{(j)} = \sum (x^{(j)})^2 - \bar{X}\sum x^{(j)} + \bar{X}\sum x^{(j)} - \bar{X}\sum x^{(j)}$$

$$\sum (x^{(j)})^2 - \bar{X}\sum x^{(j)} = \sum (x^{(j)})^2 - 2\bar{X}\sum x^{(j)} + m\bar{X}^2$$

$$\sum (x^{(j)})^2 - \bar{X}\sum x^{(j)} = \sum (x^{(j)} - \bar{X})^2 \quad (19)$$

Коначно решење за b из (18) и (19) је:

$$b = \frac{\sum (\boldsymbol{x}^{(j)} - \bar{\boldsymbol{X}})\,(\boldsymbol{y}^{(j)} - \bar{\boldsymbol{Y}})}{\sum (\boldsymbol{x}^{(j)} - \bar{\boldsymbol{X}})^2} \quad (20)$$

Размотримо сада исти LR модел у матричној нотацији. Скуп за учење трансформишимо у матрицу X, где сваки ред матрице представља вектор вредности предиктора једног посматрања $x^{(j)}$.

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(m)} \end{bmatrix} \quad (21)$$

Вектор $\vec{y}$ представља све посматране вредности $y_{obs}^{(j)}$

$$y = [y^{(1)} \quad y^{(2)} \quad \cdots \quad y^{(m)}]^{\mathrm{T}} \quad (22)$$

Вектор тежина $\vec{b}$

$$b = [a \quad b_1]^{\mathrm{T}} \quad (23)$$

LR модел је онда облика:

$$\vec{y} = X\vec{b} \quad (24)$$

Ако једначине (8) и (9) такође представимо матрично, уз замену $a = b_0$ ради униформности означавања, имаћемо:

$$\begin{bmatrix} \sum y^{(j)} \\ \sum x^{(j)} y^{(j)} \end{bmatrix} = \begin{bmatrix} m & \sum x^{(j)} \\ \sum x^{(j)} & \sum (x^{(j)})^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad (25)$$

Лева страна (25) се може помоћу (21) и (22) представити

$$\begin{bmatrix} \sum y^{(j)} \\ \sum x^{(j)} y^{(j)} \end{bmatrix} = \boldsymbol{X^T Y} \quad , \qquad \text{а десна страна} \quad \begin{bmatrix} m & \sum x^{(j)} \\ \sum x^{(j)} & \sum (x^{(j)})^2 \end{bmatrix} = \boldsymbol{X^T X}$$

где је $\boldsymbol{X^T} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \end{bmatrix}$

Сада (25) може да се трансформише у:

$$\boldsymbol{X^T y = X^T X\, b} \quad (26)$$

а вектор тежина израчунавамо множењем (26) слева инверзном матрицом $(X^T X)^{-1}$, чиме решавамо проблем учења LR модела у коме важи:

$$\boldsymbol{\hat{b} = (X^T X)^{-1} X^T y} \quad (27)$$

Ова метода генерисања LR модела пати од неких недостатака. Ако постоје међусобно линеарно зависни предиктори, могуће је да не постоји матрица $(X^T X)^{-1}$ или је, пак, лоше условљена. Ипак тај проблем може да се решава различитим поступцима регуларизације којима се смањује утицај појединих предиктора. Други проблем који се тиче израчунљивости постаје значајан када број предиктора расте. Једначина (27) користи се у многим статистичким алатима (SPSS, R, Excel) за релативно мали број предиктора ($n \leq 10$). Када број предиктора расте, једначина (27) захтева знатно процесорско време услед пораста димензије матрица, па се користе различити поступци редукције векторског простора предиктора и итеративни поступци учења модела, о чему ће бити више речи у поглављу 5. У одељку 2.7 увешћемо параметре и методе којима се оцењују регресиони модели.

2.5 Логистички регресиони модел класификације докумената

У одељку 1.3.2 говорили смо о ограничењима класификационих модела који су у основи искључиви и у поступку класификације генеришу строго дисјунктне скупове (документ припада класи – документ не припада класи).

Међутим, често уместо одговора да ли неки документ припада или не припада датој класи, желимо да знамо вероватноћу припадања документа датој класи. У том случају, линеарни модел је функција која представља вероватноћу да посматрани документ из скупа за учење припада класи у коју је сврстан скупом за учење:

$$P(y = true|x) = \sum_{i=0}^{n} w_i \, f_i = \vec{w} \cdot \vec{f} \tag{28}$$

Међутим, проблем са оваквим моделом је што је вредност десне стране реалан број $y \in \{-\infty, +\infty\}$, а леве $p \in [0,1]$. Стога можемо, уместо вероватноће, посматрати шансу (енг. odds) тј. однос вероватноће припадања класи и вероватноће неприпадања класи:

$$\frac{P(y=true|x)}{1 - P(y=true|x)} = \sum_{i=0}^{n} w_i \, f_i \tag{29}$$

У том случају имамо леву страну из интервала $[0, +\infty\}$. Међутим, ако леву страну логаритмујемо, тј. уведемо природан логаритам

$$\ln\left(\frac{P(y=true|x)}{1 - P(y=true|x)}\right) = \vec{w} \cdot \vec{f} \tag{30}$$

онда су обе стране у интервалу $\{-\infty, +\infty\}$. Логаритам шансе назива се *логит функцијом*:

$$logit\big(P(x)\big) = \ln\left(\frac{P(x)}{1-P(x)}\right) \tag{31}$$

Модел регресије који користи линеарну функцију да оцени логит функцију вероватноће да документ припада класи зове се логистичка или лог-линеарна регресија. Иако носи назив „регресија“, логистичка регресија је, по својој природи, вероватносна дискриминативна класификациона метода. Низом трансформација (30) добићемо вероватноћу да посматрани документ x припада класи y, на основу два вектора: вектора тежина $\vec{w}$ и вектора предиктора $\vec{f}$ посматраног документа x.[28]

[28] Ради једноставнијег писања, у наставку, скаларни производ $\vec{w} \cdot \vec{f}$ биће написан помоћу $w \cdot f$

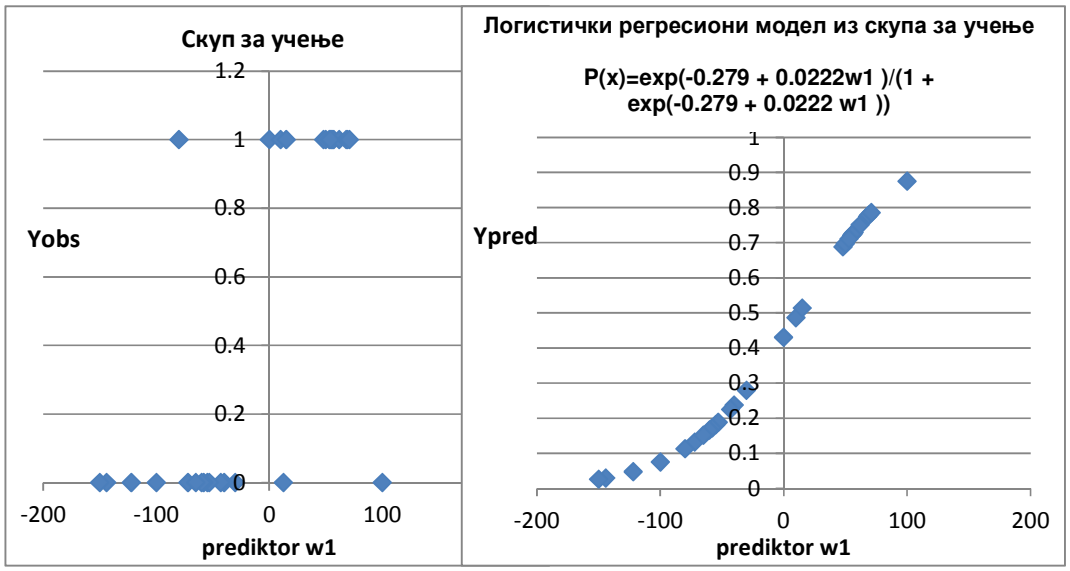
$$\ln\left(\frac{P(y)}{1\ -\ P(y)}\right) = \vec{w}\cdot\vec{f}$$

$$\frac{P(y)}{1\ -\ P(y)} = e^{w\cdot f}$$

$$P(y = true|x) = \frac{e^{w\cdot f}}{1+ e^{w\cdot f}} \tag{32}$$

$$P(y = false|x) = 1 - P(y = true|x)\ = \frac{1}{1+ e^{w\cdot f}} \tag{33}$$

Функција вероватноће (32) да документ припада посматраној класи је у нелинеарној вези са предикторима и назива се *сигмоидалном функцијом* или *основном логистичком функцијом*. У општем случају, она пресликава скуп вредности предиктора из интервала $\{-\infty, +\infty\}$ у затворени интервал вероватноћа [0,1]. На слици 2.5 приказан је однос вредности једног предиктора x_1 случајне променљиве x и вероватноће $P(x)$.



Слика 2.5 Логистички регресиони модел а) Скуп за учење са једним дихотомним предиктором б) Логистички регресиони модел скупа за учење - модел који даје максималну условну вероватноћу да $p(y_{pred}|x) = y_{obs}$

Једначинама (32) и (33) показали смо да логистичка функција изражава вероватноћу класификације. Међутим, у одељку 2.3 истакли смо да

дискриминативни класификатори генеришу модел апостериорних вероватноћа $p(c|x)$ и моделују границе између класа. Покажимо да се логистичком функцијом моделује управо апостериорна вероватноћа $P(C = c_1|X)$.

Нека посматрамо бинарни класификатор са класама c_1 и c_2 једнаких варијанси $\sigma_1 = \sigma_2 = \sigma$. Тада важи $P(C = c_1|X) = 1 - P(C = c_2|X)$.

Ако је случајна променљива X дефинисана помоћу n међусобно независних предиктора, тада важи

$$P(X|c_1) = P(x_1, \dots, x_n|c_1) = \prod_{i=1}^{n} P(x_i|c_1)$$

па је апостериорна вероватноћа:

$$P(c_1|X) = \frac{P(c_1)\prod P(x_i|c_1)}{P(c_1)\prod P(x_i|c_1) + P(c_2)\prod P(x_i|c_2)} = \frac{1}{1+\frac{P(c_2)\prod P(x_i|c_2)}{P(c_1)\prod P(x_i|c_1)}}$$

Ако уведемо замену $e^{-\alpha} = \frac{P(c_2)\prod P(x_i|c_2)}{P(c_1)\prod P(x_i|c_1)}$

апостериорна вероватноћа зависи од параметра α и има облик сигмоидалне функције .

$$P(c_1|X) = \frac{1}{1+e^{-\alpha}} \tag{34}$$

Али, да би једначина (34) могла бити усвојена као *логистичка* функција потребно је да параметар α представља линеарну комбинацију предиктора и тежина, односно мора важити (35)

$$\alpha = ln\frac{P(c_1)\prod P(x_i|c_1)}{P(c_2)\prod P(x_i|c_2)} = \sum \beta_i x_i \tag{35}$$

Трансформишимо део једначине (35)

$$\alpha = ln\frac{P(c_1)\prod P(x_i|c_1)}{P(c_2)\prod P(x_i|c_2)} = \sum ln\frac{P(x_i|c_1)}{P(x_i|c_2)} + ln\frac{P(c_1)}{P(c_2)}$$

Условне вероватноће $P(X|c_1)$ и $(X|c_2)$ изразимо помоћу Гаусове расподеле.

$$\sum ln\frac{P(x_i|c_1)}{P(x_i|c_2)} = \sum ln\frac{\frac{1}{\sqrt{2\pi\sigma_i}}\exp(-\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi\sigma_i}}\exp(-\frac{(x_i-\mu_{i2})^2}{2\sigma_i^2})}$$

$$\sum ln\frac{P(x_i|c_1)}{P(x_i|c_2)} = \sum\frac{(x_i-\mu_{\iota 2})^2-(x_i-\mu_{\iota 1})^2}{2\sigma_\iota^2} = \sum\frac{2x_i(\mu_{\iota 1}-\mu_{\iota 2})+{\mu_{\iota 2}}^2-{\mu_{\iota 1}}^2}{2\sigma_\iota^2}$$

$$\sum ln\frac{P(x_i|c_1)}{P(x_i|c_2)} = \sum(\frac{(\mu_{\iota 1}-\mu_{\iota 2})}{\sigma_\iota^2}x_i + \frac{{\mu_{\iota 2}}^2-{\mu_{\iota 1}}^2}{2\sigma_\iota^2})$$

Ако уведемо замене:

$$\beta_\iota = \frac{(\mu_{\iota 1}-\mu_{\iota 2})}{\sigma_\iota^2},\ i \in \{1,\dots,n\} \qquad \text{и} \qquad \beta_0 = \sum\frac{{\mu_{\iota 2}}^2-{\mu_{\iota 1}}^2}{2\sigma_\iota^2} + ln\frac{P(c_1)}{P(c_2)}$$

параметар α можемо изразити као:

$\alpha = \beta_0 + \sum\beta_\iota x_i$ чиме добијамо линеарни модел дат једначином (35)

где су $\beta_\iota, i \in \{0,\dots,n\}$ регресиони коефицијенти.

2.5.1 Процес учења у логистичкој регресији

У линеарној регресији процес учења се састоји од проналажења оног вектора тежина $\vec{w}$ који минимизује суму квадратне грешке свих предиктора из скупа за учење. У логистичкој регресији се до вектора тежина $\vec{\beta}$ долази оценом максималне веродостојности (енг. maximum-likelihood estimation MLE) (Myung, 2003; Mladenović, 2008), што значи да вектор $\vec{\beta}$ чине оне тежине које дају највећу вероватноћу да класификација сваког документа $X^{(j)}$ из посматраног скупа за учење буде $y^{(j)} = {y_{obs}}^{(j)}$.

Функција веродостојности $l(\beta)$ дефинише се као производ условних вероватноћа тако да сваки пар за учење $(X^{(j)}, y^{(j)})$, где је $X^{(\mathrm{j})} = \left(x_1^{(\mathrm{j})}, x_2^{(\mathrm{j})}, \dots, x_n^{(\mathrm{j})}\right)$, има вероватноћу $\mathrm{P}\left(y^{(\mathrm{j})} = 1 \middle| X^{(\mathrm{j})}\right)$ уколико је $y_{\mathrm{obs}}^{(\mathrm{j})} = 1$, односно вероватноћу $\mathrm{P}\left(y^{(\mathrm{j})} = 0 \middle| x^{(\mathrm{j})}\right) = 1 - P\left(y^{(j)} = 1 \middle| X^{(j)}\right)$ уколико је $y_{\mathrm{obs}}^{(\mathrm{j})} = 0$, што на целом скупу за учење представља:

$$l(\beta) = \prod_{j=1}^{m}\begin{cases} P\left(y^{(j)} = 1 \middle| X^{(j)}\right)\ za\ y^{(j)} = 1 \\ P\left(y^{(j)} = 0 \middle| X^{(j)}\right)\ za\ y^{(j)} = 0 \end{cases}$$

Оцене параметара $\hat{\beta}$ налазимо из услова максималне вредности функције веродостојности. Међутим, у ML уобичајено је да се уместо максимума

функције веродостојности тражи минимум њеног негативног логаритма $L(\beta)$.

$$L(\beta) = -\log\big(l(\beta)\big) = -\sum_{j=1}^{m} \log \begin{cases} P\big(y^{(j)} = 1 \big| X^{(j)}\big) \ za \ y^{(j)} = 1 \\ P\big(y^{(j)} = 0 \big| X^{(j)}\big) \ za \ y^{(j)} = 0 \end{cases}$$

односно:

$$L(\beta) = -\sum_{j=1}^{m} \Big(y^{(j)} \log P\big(y^{(j)} = 1 \big| X^{(j)}\big) + \big(1 - y^{(j)}\big) \log P\big(y^{(j)} = 0 \big| X^{(j)}\big) \Big)$$

Скуп оцена регресионих параметара $\hat{\beta}$ добићемо из

$$\hat{\beta} = \underset{\beta}{argmax}\, L(\beta)$$

односно, ако применимо (34) и (35)

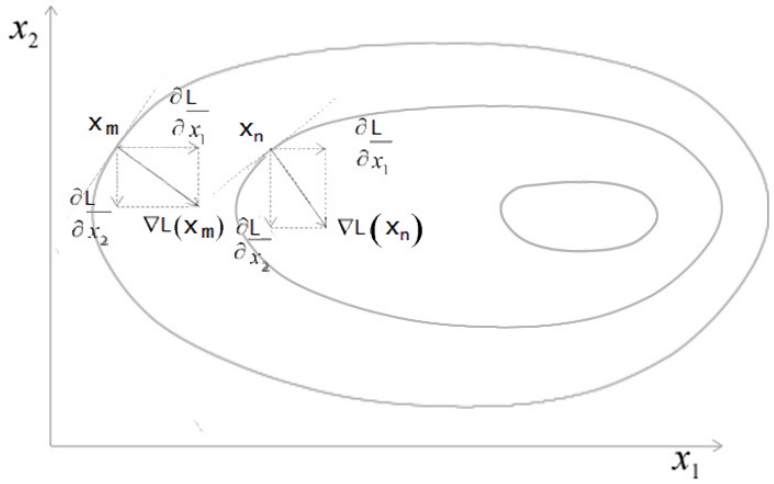
$$\hat{\beta} = \underset{\beta}{argmax} - \sum_{i=1}^{m} \left(y^{(j)} \log \frac{1}{1 + e^{-\beta \cdot x^{(j)}}} + \big(1 - y^{(j)}\big) \log \frac{e^{-\beta \cdot x^{(j)}}}{1 + e^{-\beta \cdot x^{(j)}}} \right) \qquad (36)$$

где је $\beta \cdot x^{(j)}$ скаларни производ вектора регресионих коефицијената $\vec{\beta}$ и вектора предиктора j-тог елемент скупа за учење $X^{(j)}$.

За разлику од линеарног регресионог модела, где се модел учења може дати у аналитичком облику (једначина 27), у логистичкој регресији решење скупа нелинеарних једначина (36) представља један од проблема безусловне оптимизације (Myung, 2003). Најчешче коришћене методе учења у логистичкој регресији су градијентне методе (Myung, 2003; Stanimirović, 2014). Оне спадају у методе безусловне оптимизације итеративног типа, где се у сваком кораку итерације генерише тачка $x^{(k+1)} \in R^n$ таква да важи $x^{(k+1)} = x^{(k)} + l^{(k)} \nabla L^{(k)}$, $k = 0,1,2,\ldots$ где је $\nabla L^{(k)} \in R^n$ градијент, односно правац кретања из тачке $x^{(k)}$ ка траженом екстремуму функције веродостојности, а $l^{(k)} > 0$, $l^{(k)} \in R^n$ је дужина корака дуж градијентног вектора $\nabla L^{(k)}$, вектора који у свакој тачки $x^{(k)}$ има смер најбржег раста посматране функције у тој тачки.

$$\nabla L\big(x^{(k)}\big) = \text{grad} L\big(x^{(k)}\big) = \left\{ \frac{\partial L}{\partial x_1^{(k)}}, \frac{\partial L}{\partial x_2^{(k)}}, \ldots, \frac{\partial L}{\partial x_n^{(k)}} \right\}$$

За случај функције од два предиктора $X=[x_1, x_2]^T$, на слици 2.6 дата је пројекција градијента који представља нормалу на тангентну раван у тачкама x_m и x_n функције L.



Слика 2.6. Градијентни вектори у тачкама x_n и x_m

Постоји више различитих критеријума заустављања итеративног процеса. Идеалан критеријум је $L(\beta^{(k+1)}) = L(\beta^{(k)})$, међутим у реалним условима се обично узима као услов она тачка функције веродостојности LR модела $L(\beta)$ за коју важи $L(\beta^{(k+1)}) - L(\beta^{(k)}) \leq \varepsilon$, $\varepsilon > 0$. То такође може бити и релативна грешка функције веродостојности $\frac{|L(\beta^{(k+1)}) - L(\beta^{(k)})|}{|L(\beta^{(k)})|} \leq \varepsilon$ или неки други значајан услов.[29] У градијентним методама је важно обезбедити почетну тачку од које креће итеративни поступак и утврдити да ли функција која представља LR модел задовољава потребне и довољне услове за поступак безусловне итеративне оптимизације.

Побољшање градијентних метода може се постићи Њутн-Рафсоновим[30] методом. Реч је о квадратној апроксимацији циљне функције $f(x)$ у области дате тачке $x^{(k)}$ која користи први и други извод циљне функције, па се често

[29] У статистичким софтверским пакетима обично се као критеријум завршетка итеративног процеса узима унапред постављен укупан број итерација.
[30] Isaac Newton, Joseph Raphson

назива и градијентна метода другог реда. Максимизацијом квадратне апроксимације добијамо апроксимацију максимума циљне функције.

Квадратну апроксимацију представља развој функције у Тејлоров[31] ред функције f у тачки $x^{(k)}$ члановима првог и другог степена

$$\tilde{f}(x) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)\left(x - x^{(k)}\right) + \frac{1}{2}\left(x - x^{(k)}\right)^T \nabla^2 f\left(x^{(k)}\right)\left(x - x^{(k)}\right) \qquad (37)$$

Једначина (37) представља Тејлоров развој функције $f: R \to R$ (са једним предиктором). Како функција $L(\beta)$ у општем случају има n предиктора, то је потребно представити Тејлоров развој функције $f: R^n \to R$.

Нека је $f: R^n \to R$ двапут диференцијабилна. Тада у околини тачке $x^{(k)}$ важи:

$$f(x) \approx f\left(x^{(k)}\right) + \mathrm{g}^T\left(x - x^{(k)}\right) + \frac{1}{2}\left(x - x^{(k)}\right)^T \mathrm{H}\left(x - x^{(k)}\right)\left(x - x^{(k)}\right) \qquad (38)$$

где је $g = \nabla f\left(x^{(k)}\right)$ градијент функције, а $\mathrm{H} = \nabla^2 f\left(x^{(k)}\right)$ Хесијан функције у тачки $x^{(k)}$.

Хесијан или Хесеова[32] матрица је квадратна матрица парцијалних извода другог реда функције $f: R^n \to R$ која је двапут диференцијабилна на посматраном интервалу.

$$\nabla^2 f(x) = H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Сређивањем (38) по x имаћемо:

$$f(x) \approx \mathrm{g}^T x + \frac{1}{2} x^T \mathrm{H} x - H x^{(k)} x + c \qquad (39)$$

где је $c = f\left(x^{(k)}\right) - \mathrm{g} x^{(k)} + \frac{1}{2} {x^{(k)}}^2 \mathrm{H}$,

Ако са $q(x)$ означимо израз на десној страни (39) имаћемо:

[31] Brook Taylor
[32] Ludwig Otto Hesse

$$q(x) = \mathrm{g}^T x + \frac{1}{2}\, x^T \mathrm{H} x + \left(\mathrm{g}^T - H x^{(k)}\right) x + c$$

а применом услова за екстремум функције, $\frac{\partial q(x)}{\partial x} = 0$, добијамо:

$$\mathrm{H} x + \left(\mathrm{g}^T - H x^{(k)}\right) = 0$$

$$x = H^{-1}\left(H x^{(k)} - \mathrm{g}^T\right)$$

$$\boldsymbol{x = x^{(k)} - H^{-1} g^T} \qquad (40)$$

где је x тачка локалног екстремума функције f.

Конвергенција Њутнове методе могућа је ако је функција f двапут диференцијабилна и ако постоји инверзна Хесеова матрица. Тада, уколико је Хесеова матрица позитивно дефинитна[33], односно важи $H^{-1}\mathrm{g}^T > 0$, Њутнова метода конвергира ка минимуму функције f. Уколико је Хесеова матрица негативно дефинитна, односно важи $H^{-1}\mathrm{g}^T < 0$, Њутнова метода конвергира ка максимуму функције f. Међутим, уколико је почетна тачка итерације далеко од екстремума функције или не постоји инверзна Хесеова матрица, или она није дефинитна, Њутнова метода не конвергира. Тада се могу користити различити алгоритми побољшања ове методе попут Левенберг-Маркардове[34] модификације (Myung, 2003), квази-њутновских метода и метода коњугованих градијената (Stanimirović, 2014). Ако пођемо од (36), важиће

$$L(\beta) = -\sum_{\mathrm{j}=1}^{m} y_j \log \frac{1}{1+ e^{-\beta^T \cdot x_j}} + \left(1 - y_j\right) \log \frac{e^{-\beta^T \cdot x_j}}{1+ e^{-\beta^T \cdot x_j}} \qquad (41)$$

Уведимо смену $\alpha_j = \frac{1}{1+ e^{-\beta^T \cdot x_j}}$ у (41) где $j = 1,2,\dots,m$ (42)

$$L(\beta) = -\sum_{j=1}^{m} y_j \ \log\ \alpha_j + \left(1 - y_j\right) \log(1 - \alpha_j) \qquad (43)$$

Тада важи:

$$\frac{\partial}{\partial \beta_k} log \alpha_j = \frac{\partial}{\partial \beta_k} log \frac{1}{1+ e^{-\beta^T \cdot x_j}} = \frac{\partial}{\partial \beta_k}\left(-\log\left(1 + e^{-\beta^T \cdot x_j}\right)\right) = \frac{x_{jk} e^{-\beta^T \cdot x_j}}{1+e^{-\beta^T \cdot x_j}} = x_{jk} \frac{e^{-\beta^T \cdot x_j}}{1+e^{-\beta^T \cdot x_j}} = x_{jk}(1 - \alpha_j) \qquad (44)$$

[33] Симетрична, квадратна матрица H је позитивно дефинитна ако важи $\mathrm{z}^T H\, \mathrm{z} > 0$, за било који нетривијални вектор z реалних n бројева ($\mathrm{z}_i \neq 0$)

[34] Kenneth Levenberg, Donald Marquardt

где $x_j = (x_{j1}, x_{j2}, \dots x_{jn})$.

Други део израза из (42) може се трансформисати:

$$\log(1-\alpha_j) = \log\frac{e^{-\beta^T \cdot x_j}}{1+e^{-\beta^T \cdot x_j}} = \log\left(e^{-\beta^T \cdot x_j}\right) - \log\left(1+e^{-\beta^T \cdot x_j}\right) = -\beta^T x_j - \log\left(1+e^{-\beta^T \cdot x_j}\right)$$

$$\frac{\partial}{\partial \beta_k}\log(1-\alpha_j) = \frac{\partial}{\partial \beta_k}\left(-\beta^T x_j - \log\left(1+e^{-\beta^T \cdot x_j}\right)\right)$$

$$\frac{\partial}{\partial \beta_k}\log(1-\alpha_j) = -x_{jk} + x_{jk}(1-\alpha_j) = -x_{jk}\,\alpha_j \qquad (45)$$

Изразе из (44) и (45) уврстимо у (43) и нађимо $\frac{\partial L}{\partial \beta_k}$

$$\frac{\partial \mathrm{L}}{\partial \beta_\mathrm{k}} = -\sum_{j=1}^{m} y_j x_{jk}(1-\alpha_j) - (1-y_j)x_{jk}\alpha_j$$

$$\frac{\partial \mathrm{L}}{\partial \beta_\mathrm{k}} = \sum_{j=1}^{m}(\alpha_j - y_j)x_{jk} \qquad (46)$$

Можемо уочити да су x_{jk} заправо чланови скупа за учење, па можемо дефинисати матрицу X.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

Сума у (46) представља скаларни производ j-те колоне матрице X и вектора $(\alpha_j - y_j)^{\mathrm{T}}$.

$$\frac{\partial \mathrm{L}}{\partial \beta_\mathrm{k}} = (\boldsymbol{\alpha} - \boldsymbol{y})^{\mathrm{T}}\mathbf{X}$$

Израз $\frac{\partial \mathrm{L}}{\partial \beta_\mathrm{k}}$ називамо градијентом функције $L(\beta)$ и важи:

$$\boldsymbol{\nabla_\beta \mathbf{L}} = \boldsymbol{\frac{\partial \mathbf{L}}{\partial \beta}} = \boldsymbol{X^T(\alpha - y)} \qquad (47)$$

Размотримо сада Хесијан функције $L(\beta)$ и покажимо да је Хесеова матрица позитивно дефинитна, што обезбеђује конвексност функције (Myung, 2003), односно конвергентност Њутнове методе ка минимуму функције $L(\beta)$, а самим тим и максимуму функције веродостојности $l(\beta)$ и оценама параметара $\hat{\beta}$ као решењу проблема учења у LR моделу. Диференцирањем (46) имаћемо:

$$H(x) = \frac{\partial^2 L}{\partial\beta_l \partial\beta_k} = \frac{\partial L}{\partial\beta_l}\frac{\partial L}{\partial\beta_k} = \frac{\partial L}{\partial\beta_l}\left(\sum_{j=1}^{m}\left(\alpha_j - y_j\right)x_{jk}\right)$$

$$H(x) = \sum_{j=1}^{m} x_{jk}\frac{\partial\alpha_j}{\partial\beta_l} \tag{48}$$

Како генерално важи да је $\partial(\log(\alpha)) = \frac{\partial\alpha}{\alpha}$, имаћемо

$$\partial\alpha = \alpha\partial(\log(\alpha)) \tag{49}$$

С обзиром да према (44) важи: $\partial(\log(\alpha)) = x_j(1-\alpha_j)$, то се (49) може изразити као:

$$\partial\alpha_j = \alpha_j\ x_j(1-\alpha_j) \tag{50}$$

Ако (50) употребимо у (48) биће:

$$H(x) = \sum_{j=1}^{m} x_{jk}\frac{\partial}{\partial\beta_l}\left(\alpha_j\ x_j(1-\alpha_j)\right) = \sum_{j=1}^{m} x_{jk}x_{jl}\alpha_j(1-\alpha_j)$$

Ако уведемо ознаке:

$z_k = (x_{1k}, x_{2k}, \dots, x_{mk})^T$

$z_l = (x_{l1}, x_{l2}, \dots, x_{ln})$

$B = \begin{bmatrix} \alpha_1(1-\alpha_1) & 0 & \cdots & 0 \\ 0 & \alpha_2(1-\alpha_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_m(1-\alpha_m) \end{bmatrix}$ дијагонална матрица

тада Хесијан функције је:

$$\boldsymbol{\nabla_\beta}^2\mathbf{L} = H(x) = z_k^T B\ z_l \tag{51}$$

Како низ вектора z_k представља колоне матрице X, а низ вектора z_l редове исте матрице, (51) можемо записати као:

$$\boldsymbol{\nabla_\beta}^2\mathbf{L} = H(x) = \mathrm{A}^{\mathrm{T}} B\ \mathrm{A} \tag{52}$$

Како увек важи $\alpha_i(1-\alpha_i) > 0$, јер је према (42) α_i позитиван број из интервала (0,1), то је матрица B дијагонална, чији су чланови на главној дијагонали позитивни. У том случају се она може изразити као $\mathrm{B} = \mathrm{B}^{1/2}\mathrm{B}^{1/2}$, па је квадратна форма у (52) матрица чији су сви чланови позитивни јер је:

$$\mathrm{A}^{\mathrm{T}} B\ \mathrm{A} = \mathrm{A}^{\mathrm{T}}\ B^{1/2}B^{1/2}\mathrm{A} = (B^{1/2}A)^{\mathrm{T}}(B^{1/2}A)$$

У том случају Хесеова матрица функције $L(\beta)$ је позитивно семидефинитна, па према теореми о конвексности функција (Myung, 2003) може се применити у итеративним поступцима минимизације како бисмо добили оптималан скуп тежина предиктора скупа за учење.

Примена Њутнове методе у логистичкој регресији носи ознаку IRLS – метод најмањих квадрата прерачунатих итеративно (енг. Iterative Reweighted Least Squares).

Ако су градијент и Хесијан функције L: $R^n \to R$ дефинисани са (47) и (52)

$$\nabla_\beta L = \frac{\partial L}{\partial \beta} = A^T(\alpha - y), \quad \nabla_\beta{}^2 L = H(x) = \mathrm{A}^\mathrm{T} B\ \mathrm{A}$$

где је $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$, $y = (y_1, y_2, \dots, y_m)$, $A(m \times n)$ матрица скупа за учење.

Према Њутновој методи, $x^{(k+1)}$ је тачка локалкног минимума функције дата са (40)

$$x^{(k+1)} = x^{(k)} - H^{-1} g^T = x^{(k)} - \left(\mathrm{A}^\mathrm{T} B\ \mathrm{A}\right)^{-1} A^T(\alpha - y) \tag{53}$$

Ако претходну једнакост проширимо тако да важи:

$$x^{(k)} = \left(\mathrm{A}^\mathrm{T} B\ \mathrm{A}\right)^{-1} A^T B A\ x^{(k)} \text{ и } (\mathrm{A}^\mathrm{T} B\ \mathrm{A})^{-1} A^T(\alpha - y) = (\mathrm{A}^\mathrm{T} B\ \mathrm{A})^{-1} A^T B B^{-1}(\alpha - y)$$

Тада (53) може да се изрази као:

$$x^{(k+1)} = \left(\mathrm{A}^\mathrm{T} B\ \mathrm{A}\right)^{-1} A^T B\left(\mathrm{A} x^{(k)} - B^{-1}(\alpha - y)\right) \tag{54}$$

Ако резултујући вектор дела израза (54) запишемо као

$$z_k = \mathrm{A} x^{(k)} - B^{-1}(\alpha - y)$$

једначина (54) се трансформише у једначину IRLS модела:

$$\boldsymbol{x^{(k+1)} = \left(\mathrm{A}^\mathrm{T} B\ \mathrm{A}\right)^{-1} A^T B z_k} \tag{55}$$

Можемо уочити да је (55) аналогна једначини (27) која представља линеарни регресиони модел. У (55) се укључује дијагонална матрица B чији чланови

зависе од тежина предиктора па се сматра додатним тежинским коректором у IRLS моделу.

Најзад, када се итеративним поступком дође до оптималног скупа тежина предиктора, генерисани логистички модел је спреман да буде тестиран на непознатим инстанцама истог типа.

Класификациони процес у логистичком моделу за произвољни документ d који не припада скупу за учење оцењује класу за коју је највероватније да јој посматрани документ припада. Уколико претпоставимо да документ d припада класи y, тада важи:

$$P(y = true|d) > P(y = false|d)$$

$$\frac{P(y = true|d)}{P(y = false|d)} > 1$$

$$\frac{P(y = true|d)}{1 - P(y = true|d)} > 1$$

$$e^{\beta \cdot f} > 1$$

$$\beta \cdot f > 0$$

$$\sum_{i=0}^{n} \beta_i f_i > 0$$

Последња једначина представља једначину хиперравни у $n-$димензионалном простору која одваја документе који припадају класи од оних који јој не припадају.

2.6 Метода максималне ентропије

Логистичка регресија се може користити и у случају када у процесу класификације имамо скуп дискретних излазних вредности, односно више класа. У рачунарској лингвистици метода полиномијалне логистичке регресије се назива метода максималне ентропије или *MaxEnt* метода. Разматрајући проблем моделирања система машинског превођења и ослонивши се на претходна истраживања (Darroch & Ratcliff, 1972; Jaynes,

1991), Бергер са сарадницима[35] (Berger, Della Pietra & Della Pietra, 1996; Della Pietra, Della Pietra & Lafferty, 1997) усваја принцип максималне ентропије према коме, између свих дистрибуција вероватноћа које задовољавају постављена ограничења, најоптималније је изабрати ону са максималном ентропијом, односно са најуниформнијом расподелом[36]. Они даље доказују да је модел полиномијалне логистичке регресије који постиже максималну веродостојност еквивалентан моделу са максималном ентропијом који поштује ограничења дефинисана скупом за учење и да је решење оптимизационог проблема којим се налази вектор тежина $\vec{\beta}$ којим модел постиже максималну веродостојност истовремено и решење проблема којим се налази модел са максималном ентропијом под ограничењима задатих скупом за учење. Основна идеја ове методе је да се генерише модел (односно таква дистрибуција условних вероватноћа $p^*(y|x)$) који има максималну ентропију, али да је, при том, конзистентан са информацијама скупа за учење.

Метода максималне ентропије (енг. Maximum entropy model – *MaxEnt*) се успешно користи у задацима обраде природног језика: машинском превођењу (Berger, Della Pietra & Della Pietra, 1996), означавању врстама речи[37] (Ratnaparkhi, 1996), моделирању језика (Rosenfeld, 1996), екстракцији информација (McCallum, Freitag & Pereira, 2000), класификацији докумената (Manning & Schütze, 1999; Nigam, Lafferty & McCallum, 1999), парсирању (Ratnaparkhi, Roukos & Ward, 1994; Charniak, 2000) и др. У наставку овог поглавља показаћемо који су услови дуалности функција максималне веродостојности и максималне ентропије са ограничењима и како то можемо употребити у моделовању класификационих система.

[35] Adam Berger
[36] филозофско учење Окама – в. одељак 1.3.1
[37] POS tagging

2.6.1 Моделирање система помоћу методе максималне ентропије

У задацима обраде природног језика се врло често, уместо нумеричких предиктора, користе бинарни (дихотомни) предиктори и они се називају индикаторске функције или контекстни предиктори. У статистици (James, Witten, Hastie & Tibshirani, 2014, pp. 84) се још називају индикатори или "dummy variables", док се у класификацији текстова (Manning & Schütze, 1999, pp. 590) користи појам *својство*. Скуп за учење у *MaxEnt* моделу представља се помоћу скупа индикаторских функција. Индикаторска функција се означава са $f_j(x, y)$ и интерпретира као предиктор j за класу y у односу на посматрани контекст x. На пример, уколико текст посматрамо као низ речи $words_i$ (i=0,1,2,...) где $words_0 =< s >$ представља ознаку почетка реченице, индикаторска функција се може дефинисати тако да је посматрани контекст:

а) појава да је i-та реч $words_i$ у документу $words$ властита именица (LI)

$$f_j(x,y) = \begin{cases} 1, & if\ (words_{i-1}\ != < s >)\ and\ (is\_upper\_case(words_i[0]))\ and\ (y = LI) \\ 0, & otherwise \end{cases}$$

б) појава да је i-та реч у документу $words$ присвојни придев (PP)

$$f_j(x,y) = \begin{cases} 1, & if\ (words_{i-1}\ != < s >)\ and\ \left(is\_lower\_case(words_i[0])\right) \\ & and\ (suffix(word_i) = \text{"ски"})\ and\ (y = PP) \\ 0, & otherwise \end{cases}$$

Можемо уочити да структура предиктора може бити различита и комплекснија (може представљати и логички израз) него у регресионој анализи и она зависи од доменских знања која се примењују у грађењу датог модела. У горњем б) примеру, употребљено је морфолошко правило о грађењу придева у српском језику које препознаје једну класу присвојних придева (београдски, школски, планински, спортски,...) и притом се не узимају у обзир властита имена са истоветним завршетком (Бродски, Чомски,...). Једна индикатор функција има вредност 1, уколико у скупу за учење дође до задовољења услова предиктора, у супротном вредност је 0.

За разлику од регресионе анализе, где предикторима моделујемо документе, у моделу максималне ентропије моделујемо класе. Ако имамо скуп Y од k различитих класа $y_1, y_2, \dots, y_k$, тада *MaxEnt* модел даје расподелу

условних вероватноћа класификације у скупу Y, тако да је класа представљена колекцијом парова (x_{ij}, y_j) где је x_{ij} релативна фреквенција појављивања предиктора f_i у j-тој класи и важи:

$$\tilde{p}(x,y)_{f_{ij}} = \frac{n_{ij}}{N}$$

Вероватноћа $\tilde{p}(x,y)$ зове се емпиријска вероватноћа и представља укупан број појављивања предиктора f_i, који задовољавају постављени услов у скупу за учење, у односу на N - димензију скупа за учење. Уколико је $n_{ij} > 0$, одговарајућа индикаторска функција имаће вредност 1 на датом скупу.

Ако са $X = \{x_1, x_2, \dots, x_q\}$ обележимо скуп релативних фреквенција појављивања свих предиктора f по класама, а са $Y = \{y_1, y_2, \dots, y_k\}$ скуп свих класа, тада је математичко очекивање, односно дистрибуција емпиријске вероватноће, предиктора који је дат индикаторском функцијом $f(x,y)$, где $x \in X, y \in Y$, а скуп $Z = X \times Y$[38]

$$\tilde{E}_p(f_x) = \tilde{p}(x_1,y_1)\mathrm{f}(x_1,y_1) + \tilde{p}(x_1,y_2)\mathrm{f}(x_1,y_2) + \cdots + \tilde{p}(x_q,y_k)\mathrm{f}(x_q,y_k),$$

односно

$$\tilde{E}_p(f_x) = \sum_{x,y\in Z} \tilde{p}(x,y) f(x,y)$$

Аналогно томе, математичко очекивање модела који генеришемо је:

$$E_p(f_x) = \sum_{x,y\in Z} p(x,y) f(x,y)$$

Како је вероватноћа $p(x,y)$ заправо заједничка вероватноћа да предиктор узме вредност x, а класа вредност y, то се претходна једнакост може приказати уз помоћ условне вероватноће да класа узме вредност y, уколико предиктор узме вредност x помоћу

$$E_p(f_x) = \sum_{x,y\in Z} p(x)p(y|x)f(x,y) \qquad (56)$$

У изразу (56), вероватноћа $p(x)$ може се апроксимирати емпиријском $\tilde{p}(x)$ на основу истраживања (Lau, Rosenfeld & Roukos, 1993), па важи

[38] Уочити да скуп Z представља скуп свих могућих комбинација x и y, а не само оних које се добијају из скупа за учење.

$$E_p(f_x) \approx \sum_{x,y \in Z} \tilde{p}(x) p(y|x) f(x,y) \quad (57)$$

где $E_p(f_x)$ изражава математичко очекивање условне вероватноће модела.

Да би модел био конзистентан са скупом за учење мора важити да дистрибуција сваког предиктора дата моделом буде управо она коју имамо у скупу за учење, односно да је:

$$\tilde{E}_p(f_i) = E_p(f_i), \quad i \in \{1,2,..,m\} \quad (58)$$

Једнакост (58) назива се ограничење модела и може се, помоћу (56) и (57), изразити и као:

$$\sum_{x,y \in Z} \tilde{p}(x,y) f(x,y) = \sum_{x,y \in Z} \tilde{p}(x) p(y|x) f(x,y) \quad (59)$$

У реалним условима, дистрибуција емпиријских вероватноћа предиктора у скупу за учење није једнака дистрибуцији коју генерише модел. С друге стране, циљ је да модел генерише управо ону дистрибуцију која најбоље апроксимира емпиријску и при том задовољава скуп ограничења (59).

Када градимо класификациони модел, желимо да нађемо дистрибуцију условне вероватноће $p(y|x)$ како би из тако оцењене дистрибуције утврдили највероватнију класу узорка x

$$p^*(y|x) = \underset{y \in C}{argmax}\, P(y|x), \quad (60)$$

где је $C = \left\{ p \in P \middle| E_{\tilde{p}}(f_i) = E_p(f_i),\ i \in \{1,2,\dots,n\} \right\}$.

У раду (Berger, Della Pietra & Della Pietra, 1996) аутори у поступку моделовања оваквог система примењују функцију ентропије. Ентропија је мера неуређености неког система, позната из термодинамике и теорије гасова. У теорији информација посебно је значајна функција бинарне ентропије $H(p)$ као мера неизвесности случајне промењљиве која може узети једну од две вредности 0 и 1. Дефинише се као функција вероватноће неке случајне промењљиве x по свим исходима:

$$H(p) = -\sum_x p(x) \log\big(\mathrm{p}(x)\big)$$

Моделирање посматраног класификационог система функцијом ентропије обезбеђује стратегију да се догађаји за које не постоји информација по којој се може оценити разлика у вероватноћи њиховог појављивања третирају једнако вероватним, што се постиже максимизацијом функције ентропије, а конзистентност са скупом за учење обезбеђује се задовољавањем датог скупа ограничења. Стога, проблем добијања класификационог модела максималне ентропије представља проблем оптимизације функције са ограничењима.

У том смислу, овде ћемо, без доказивања,[39] навести теореме о потребним и довољним условима постојања локалних максимума[40] ове класе функција када су ограничења дата помоћу скупа линеарних једнакости.[41]

Теорема 1 (О потребним условима првог реда о постојању локалног максимума функције при условима типа једнакости) Ако је тачка X^* локални максимум функције $f: R^n \to R$ при услову $g(X) = 0$, где је $g: R^n \to R^m, m \leq n$ и за коју важи да је регуларна тачка површи $S = \{X \in R^n | g(X) = 0\}$, тада постоји $\Lambda^* = (\lambda_1, \dots, \lambda_m)^T \in R^m$ за које важи $\nabla f(X^*) + (\Lambda^*)^T \nabla g(X^*) = 0$.

Теорема 2 (О потребним условима другог реда о постојању локалног максимума функције при условима типа једнакости) Ако је регуларна тачка X^* локални максимум функције $f: R^n \to R$ при услову $g(X) = 0$, где су f и g двапут диференцијабилне и где је $g: R^n \to R^m, m \leq n$, тада постоји $\Lambda^* \in R^m$ за које важе:

- $\nabla f(X^*) + (\Lambda^*)^T \nabla g(X^*) = 0$
- за свако $Y \neq 0$ из тангентног простора $T(X^*)$, Хесијан Лагранжове функције у тачки X^* је негативно семидефинитна матрица, тј. $Y^T H_L(X^*, \Lambda^*)\, Y \leq 0$.

Теорема 3 (О довољним условима о постојању локалног максимума функције при условима типа једнакости) Ако су функције $f: R^n \to R$ и

[39] Докази се могу наћи у књизи (Stanimirović, 2014).
[40] Аналогно се формирају и доказују теореме о постојању минимума функције.
[41] Ограничења могу бити дата помоћу линеарних једнакости или неједнакостима.

$g: R^m \to R$ реалне, двапут диференцијабилне и ако постоје тачке $X^* \in R^n$ и $\Lambda^* \in R^m$ за које важе

$\nabla f(X^*) + (\Lambda^*)^T \nabla g(X^*) = 0$ (Лагранжов услов)

и ако за свако $Y \neq 0$ из тангентног простора $T(X^*)$ Хесијан Лагранжове функције је негативно дефинитна матрица, тј. $Y^T H_L(X^*, \Lambda^*)\, Y < 0$, тада је тачка X^* строги локални максимум функције f при услову $g(X^*) = 0$.

Ако тачка $X^* \in R^n$ задовољава Лагранжов услов и Хесијан Лагранжове функције $H_L(X^*, \Lambda^*)$ је негативно дефинитан, тада је X^* тачка строгог локалног максимума функције f.

Дакле, проблем налажења екстремних вредности двапут диференцијабилне функције f са са задатим ограничењима типа једнакости своди се на проблем налажења екстремних вредности одговарајуће Лагранжове[42] функције без ограничења. На основу претходне теореме, проблем налажења

$\max f(x),\ x \in R^n,$

при задатим условима

$g(x) = C,\ x \in R^m\ ,\ m \leq n$

своди на проналажење оне тачке X^* у којој је градијент функције f паралелан градијенту функције g, односно проналажење оне тачке X^* у којој је вектор градијента функције једнак линеарној комбинацији услова, што се може дати помоћу (61)

$$\nabla f(X^*) \ = \ -(\Lambda^*)^T \nabla g(X^*) \qquad (61)$$

одакле је:

$$\nabla f(X^*) + (\Lambda^*)^T \nabla g(X^*) = 0\,. \qquad (62)$$

Вектор $\Lambda = (\lambda_1, \dots, \lambda_m)^T \in R^m$ је вектор Лагранжових множитеља, а Лагранжова функција

[42] Giuseppe Luigi Lagrangia (Joseph-Louis Lagrange)

$F_L: R^n \times R^n \to R,$ дефинисана је као:

$F_L(X, \Lambda) = f(X) + (\Lambda)^T g(X)$.

У тачки X^* Лагранжова функција је облика:

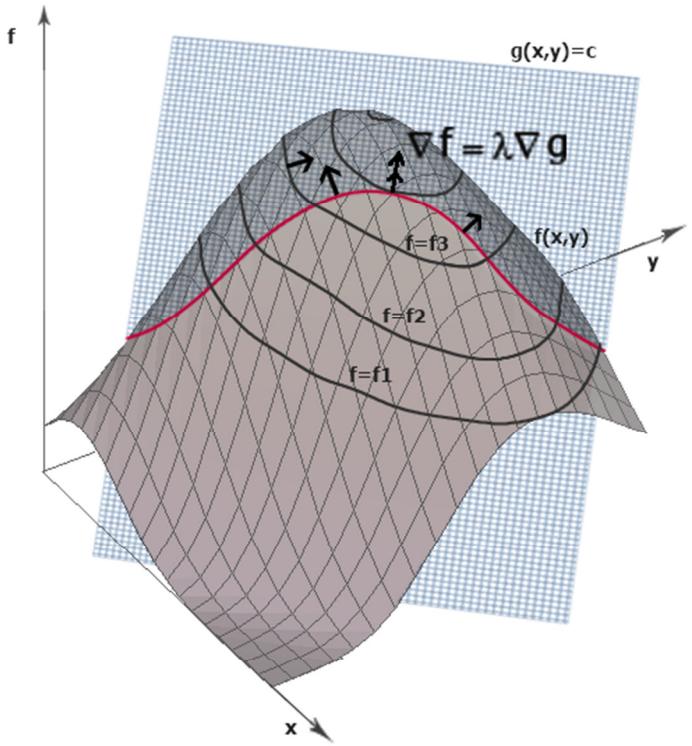
$F_L(X^*, \Lambda^*) = f(X^*) + (\Lambda^*)^T g(X) = f(X) + \sum_{i=1}^{m} \lambda_\iota g_i(X)$

а њен градијент је тада $\nabla F_L(X^*, \Lambda^*) = [\nabla_X F_L(X^*, \Lambda^*) \quad \nabla_\Lambda F_L(X^*, \Lambda^*)]^T,$

па се Лагранжов услов (62) може формулисати и као (63).

$$\left.\begin{array}{l} \nabla_X F_L(X^*, \Lambda^*) = \nabla_X f(X^*) + (\Lambda^*)^T \nabla_X g(X^*) = 0 \\ \nabla_\Lambda F_L(X^*, \Lambda^*) = g(X^*) = 0 \end{array}\right\} \qquad (63)$$

На слици 2.7 дат је приказ функције $f: R^2 \to R$, чији максимум тражимо при услову типа једнакости датог функцијом $g(x, y) = C,\ g: R^2 \to R$



Слика 2.7 Лагранжова функција – тачка максимума у којој су градијенти функције и функције ограничења паралелни.

Применимо Лагранжову методу над функцијом ентропије дефинисаном са (64) уз ограничења (59) и (65).

$$H\big(p(y|x)\big) = -\sum_{(x,y)\in Z} \tilde{p}(x)p(y|x)\log(p(y|x)) \qquad (64)$$

Такође, уведимо још ограничење нормираности:[43]

$$\sum_{y\in Y} p(y|x) = 1 \quad \text{за свако } x \qquad (65)$$

Подразумевано[44] важи и $p(y|x) \geq 0$, за све x, y.

Нека су λ_i Лагранжови множитељи. Тада је Лагранжова функција:

$$L(p,\Lambda) = \overbrace{H(y|x)}^{\text{f(x)}} + \overbrace{\sum_{i=1}^{m} \lambda_i \left(E(f_i) - \tilde{E}(f_i)\right)}^{\text{g}_i\text{(x)}} + \overbrace{\lambda_{m+1}\left(\sum_{y\in Y} p(y|x) - 1\right)}^{\text{g}_{m+1}} \qquad (66)$$

$$L(p,\Lambda) = -\sum_{(x,y)\in Z} \tilde{p}(x)p(y|x)\log p(y|x) + \sum_{i=1}^{m} \lambda_i \left(\sum_{(x,y)\in Z} \tilde{p}(x)p(y|x)f_i(x,y) - \sum_{(x,y)\in Z} \tilde{p}(y|x)f_i(x,y) \right) + \lambda_{m+1}\left(\sum_{y\in Y} p(y|x) - 1\right)$$

Максимум Лагранжове функције налазимо за $p^*(y|x)$ из услова: $\frac{\partial L(p,\Lambda)}{\partial p(y|x)} = 0$

$$\frac{\partial L(p,\Lambda)}{\partial p(y|x)} = -\tilde{p}(x)(\log p(y|x) + 1) + \sum_{i=1}^{m} \lambda_i \tilde{p}(x)\, f_i(x,y) + \lambda_{m+1}$$

$$-\tilde{p}(x)(\log(p(y|x) + 1) + \sum_{i=1}^{m} \lambda_i \tilde{p}(x)\, f_i(x,y) + \lambda_{m+1} = 0$$

$$\log p(y|x) + 1 = \sum_{i=1}^{m} \lambda_i\, f_i(x,y) + \frac{\lambda_{m+1}}{\tilde{p}(x)}$$

$$\log p(y|x) = \sum_{i=1}^{m} \lambda_i\, f_i(x,y) + \frac{\lambda_{m+1}}{\tilde{p}(x)} - 1$$

$$p(y|x) = \exp\left(\sum_{i=1}^{m} \lambda_i\, f_i(x,y) + \frac{\lambda_{m+1}}{\tilde{p}(x)} - 1\right)$$

$$p(y|x) = exp\big(\sum_{i=1}^{m} \lambda_i\, f_i(x,y)\big)\exp\left(\frac{\lambda_{m+1}}{\tilde{p}(x)} - 1\right) \qquad (67)$$

Заменом (67) у (65) добиће се:

[43] Основно својство вероватноће као функције догађаја, в. (Mladenović, 2008).
[44] Основно својство ненегативности вероватноће као функције догађаја, в. (Mladenović, 2008)

$$\exp\left(\frac{\lambda_{m+1}}{\tilde{p}(x)} - 1\right) \sum_{y \in Y} exp\left(\sum_{i=1}^{m} \lambda_i \, f_i(x,y)\right) = 1$$

одакле је

$$exp\left(\frac{\lambda_{m+1}}{\tilde{p}(x)} - 1\right) = \frac{1}{\sum_{y \in Y} exp\left(\sum_{i=1}^{m} \lambda_i f_i(x,y)\right)} \tag{68}$$

а заменом (68) у (67) добиће се:

$$p(y|x) = exp\left(\sum_{i=1}^{m} \lambda_i \, f_i(x,y)\right) \frac{1}{\sum_{y \in Y} exp\left(\sum_{i=1}^{m} \lambda_i f_i(x,y)\right)}$$

Уколико уведемо нормализациони фактор Z, такав да је

$$Z(x, \Lambda) = \sum_{y \in Y} exp\left(\sum_{i=1}^{m} \lambda_i \, f_i(x,y)\right)$$

биће

$$\boldsymbol{p^*(y|x) = exp\left(\sum_{i=1}^{m} \lambda_i \, f_i(x,y)\right) \frac{1}{Z}} \tag{69}$$

Једначина (69) представља модел максималне ентропије, односно ону дистрибуцију условних вероватноћа $p^*(y|x)$ којима се постиже максимум функције ентропије под ограничењима датих скупом за учење.

Ако дистрибуцију вероватноћа p^* у једначини (69), којом се постиже максимум Лагранжове функције $L(p, \Lambda)$ (66) када је вектор вредности Лагранжових коефицијената Λ константан, означимо са p_*^λ, односно када важи

$$p_*^\lambda \equiv \underset{p \in C}{argmax} \, L(p, \Lambda)$$

онда се функција $\Psi(\Lambda) \equiv L\left(p_*^\lambda, \Lambda\right)$ која представља Лагранжову функцију по Λ у тачки p_*^λ назива дуална функција функције дефинисане у (60).

$$\Psi(\Lambda) \equiv L\left(p_*^\lambda, \Lambda\right) = -\sum_{x,y} \tilde{p}(x) \, p_*^\lambda \log(p_*^\lambda) + \sum_{i=1}^{n} \lambda_i \left(\sum_{x,y} \tilde{p}(x) \, p_*^\lambda \, f_i - \sum_{x,y} \tilde{p}(y|x) \, f_i\right)$$

$$= -\sum_{x,y} \tilde{p}(x) \, p_*^\lambda \log(p_*^\lambda) + \sum_{x,y} \sum_{i=1}^{n} \lambda_i \, \tilde{p}(x) \, p_*^\lambda \, f_i - \sum_{x,y} \sum_{i=1}^{n} \lambda_i \, \tilde{p}(y|x) \, f_i$$

$$= -\sum_{x,y} \tilde{p}(x) \, p_*^\lambda \left(\log(p_*^\lambda) - \sum_{i=1}^{n} \lambda_i \quad f_i\right) - \sum_{x,y} \sum_{i=1}^{n} \lambda_i \, \tilde{p}(y|x) \, f_i$$

$$= -\sum_{x,y} \tilde{p}(x) \, p_*^\lambda \left(\log(p_*^\lambda) - \sum_{i=1}^{n} \lambda_i \quad f_i\right) - \sum_{x,y} \sum_{i=1}^{n} \lambda_i \, \tilde{p}(y|x) \, f_i \tag{70}$$

заменом вредности добијене на основу (69)

$\log Z(\Lambda) = -\big(\log(p_*^{\lambda}) - \sum_{i=1}^{n} \lambda_i \quad f_i\big)$

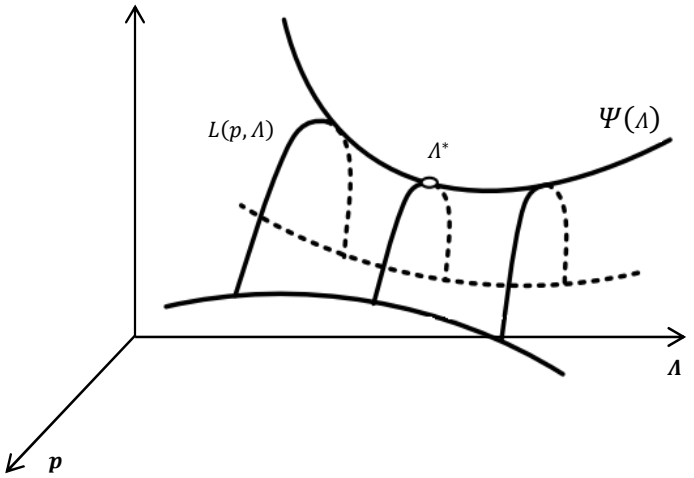
у (70), чиме добијамо дуалну Лагранжову функцију:

$$\Psi(\Lambda) \equiv L\big(p_*^{\lambda}, \Lambda\big) = \sum_{x,y} \tilde{\boldsymbol{p}}(\boldsymbol{x})\, \boldsymbol{p}_*^{\boldsymbol{\lambda}} \;\; \mathbf{log}\, \boldsymbol{Z}(\Lambda) - \sum_{x,y} \sum_{i=1}^{n} \boldsymbol{\lambda_i}\, \tilde{\boldsymbol{p}}(\boldsymbol{y}|\boldsymbol{x})\, \boldsymbol{f_i} \qquad (71)$$

Оптимални Лагранжови множитељи минимизују дуалну функцију $\Psi(\Lambda)$, односно важи

$$\Lambda^* \equiv \underset{\Lambda}{argmin}\, \Psi(\Lambda) = \min_{\Lambda} \max_{\boldsymbol{p}} L(p, \Lambda)\,.$$

Другим речима, примарни проблем проналажења оне дистрибуције вероватноћа p* којом се максимизује функција ентропије $H\big(p(y|x)\big)$ под задатим ограничењима $E(f_i) = \tilde{E}(f_i)$, еквивалентан је проблему налажења скупа параметара $\Lambda^* = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ којим се постиже минимум дуалне функције $\Psi(\Lambda)$ (слика 2.8).



Слика 2.8 Однос Лагранжове и дуалне функције

Полазећи од дуалне функције дате изразом (71), уколико узмемо у обзир (59) на основу којег важи апроксимација $\tilde{p}(x, y) \approx \tilde{p}(x)p(y|x)$, биће:

$$\Psi(\Lambda) \approx \sum_{x,y} \tilde{p}(y|x) \;\; \log Z(\Lambda) - \sum_{x,y} \sum_{i=1}^{n} \lambda_i\, \tilde{p}(y|x)\, f_i$$

$$= \sum_{x,y} \tilde{p}(y|x) \left(\log Z(\Lambda) - \sum_{i=1}^{n} \lambda_i f_i \right)$$

$$= \sum_{x,y} \tilde{p}(y|x) \left(\log Z(\Lambda) - log \left(exp \sum_{i=1}^{n} \lambda_i f_i \right) \right)$$

$$= - \sum_{x,y} \tilde{p}(y|x) \ log \left(\frac{exp \sum_{i=1}^{n} \lambda_i f_i}{Z(\Lambda)} \right)$$

$$= - \sum_{x,y} \tilde{p}(y|x) \ log\big(p(y|x)\big)$$

Може се уочити да функција $\Psi(\Lambda)$ представља негативну функцију максималне веродостојности

$\Psi(\Lambda) = -L(\Lambda) = -\sum_{x,y} \tilde{p}(y|x) \ log\big(p(y|x)\big)$

одакле се може закључити да скуп оптималних параметара $\Lambda^* = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ којим се постиже минимум дуалне функције истовремено представља решење којим се максимизује функција максималне веродостојности $L(\Lambda)$, односно да је модел полиномијалне логистичке регресије који постиже максималну веродостојност еквивалентан моделу са максималном ентропијом који поштује ограничења дефинисана скупом за учење.

2.6.2 Процес учења у моделу максималне ентропије

Док учење модела логистичке регресије представља процес налажења вектора тежина $\vec{w}$ који даје оцену максималне веродостојности над скупом за учење, што се може изразити и као

$$\hat{w} = \underset{w}{argmax} \sum_{\mathrm{j}=1}^{m} \tilde{p}(y|x) \ \log P\big(y^{(j)} \mid x^{(j)}\big)$$

процес учења у моделу максималне ентропије представља итеративни поступак налажења оног скупа Лагранжових множитеља $\Lambda^* = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ којим се постиже минимум дуалне функције $\Psi(\Lambda) = -\sum_{x,y} \tilde{p}(y|x) \ log\big(p(y|x)\big)$.

Алгоритми који се користе за налажење оптималног скупа Λ^* могу бити различити: Њутн-Рафсонова метода, L-BFGS, метода узастопних градијената, коњугованих градијената и различите методе скалирања – генерализовано итеративно скалирање (GIS) (Darroch & Ratcliff, 1972) унапређено итеративно скалирање (IIS) (Della Pietra, Della Pietra & Lafferty, 1997), итд. Како је у реализацији класификатора на основу поларитета осећања који је саставни део ове тезе коришћена метода IIS, у наставку је приказан псеудо-код датог итеративног процеса (Алгоритам 2.1). IIS спада у групу метода који у једном кораку рачуна допринос једног параметра. Укупан допринос свих параметара рачуна се након m сукцесивних корака, помоћу

$$L(\Lambda + \Delta) - L(\Lambda) = -\sum_{x,y} \tilde{p}(y|x)\ log\big(p_{\Lambda+\Delta}(y|x)\big) + \sum_{x,y} \tilde{p}(y|x)\ log\big(p_{\Lambda}(y|x)\big)$$

где је функција

$$\frac{\partial B}{\partial \delta_i} = -\sum_{x,y} \tilde{p}(x,y) f_i(x,y) - \sum_{x} \tilde{p}(x) \sum_{y} p_{\Lambda}(y|x)\, f_i(x,y) \exp(\delta_i f^{\#}(x,y))$$

и где $f^{\#}(x,y) = \sum_{i=1}^{n} f_i(x,y)$.

Алгоритам 2.1: Унапређено итеративно скалирање (IIS)
Улаз: Обележени скуп за учење дефинисан скупом предиктора $f_i, i = 1,2,\dots m$
Додатни улазни подаци: за сваки предиктор f_i дати оцену очекиване вредности на основу скупа за учење
Иницијализација: $\lambda_i = 0,\ i = 1,2,\dots m$ Итеративно: Израчунати оцену класе y_{est} за сваки од n докумената са тренутним вредностима параметара $y_{est} = p_\lambda(y\|x)$ За сваки од m параметара λ_i Поставити $\frac{\partial B}{\partial \delta_i}$=0 и решити по δ_i Поставити $\lambda_i = \lambda_i + \delta_i$
Излаз: Оптималан скуп $\Lambda^* = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$

MaxEnt метода даје дистрибуцију условних вероватноћа за све c_i класе из скупа C. Класификација је максимална вредност условних вероватноћа, тј.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c \mid x)$$

У бинарној логистичкој регресији, класификација значи проналажење једног линеарног модела односно једне раздвајајуће хиперравни. Класификација *MaxEnt* методом подразумева да се линеарни модел генерише за сваку c_i класу којим се утврђује шта је вероватније: да посматрани документ припада или не припада датој класи. Стога, може се истаћи да су добре особине ове методе:

- вишекласни систем класификације
- могућност да сваку класу описује другачији скуп предиктора
- обука скупа за учење је проблем конвексне оптимизације за који увек постоји јединствено решење.

Принцип максималне ентропије не бави се директно процесом селекције предиктора. Он једино даје најоптималнији модел од понуђених предиктора уз задовољење датог скупа ограничења. Међутим, могућност нарастања броја предиктора на више стотина или више хиљада у проблемима рачунарске лингвистике даје велику важност методама селекције предиктора, о чему ће бити речи у поглављу 5.

2.7 Оцена модела класификације

Када се класификациони модел генерише на основу датог скупа за учење, потребно је оценити његову ефикасност. Оцена успешности (евалуација) једног модела класификације састоји се у поређењу његових резултата и стварних резултата неког независног скупа података (Jurafsky & Martin, 2009). Независан скуп података може бити:

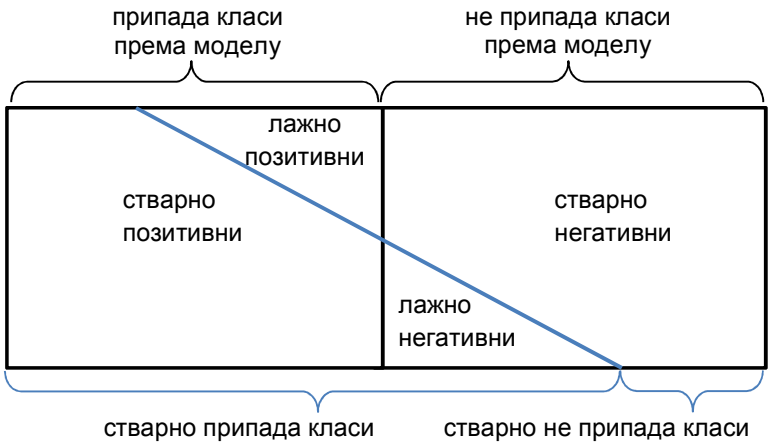
- скуп за проверу (валидациони скуп)
- скуп за тестирање
- скуп *n*-тоструке унакрсне провере.

Некада је потребно да се, пре коначне оцене класификатора, оцене поједини параметри модела, како би се он прилагодио (нпр. поткресивање стабала одлучивања, дефинисање степена регуларизације предиктора, итд.). У таквим случајевима се део скупа за учење користи као валидациони скуп (Elkan, 2012). Међутим, стварна оцена класификатора не може бити извршена на основу валидационог скупа. У ту сврху морају се користити скуп за тестирање и/или унакрсна валидација.

Када се у процесу оцене класификатора користи скуп за тестирање, онда се ради о колекцији података која није коришћена у процесу учења (тзв. unseen data). У случају унакрсне валидације, скуп за учење се подели на *n* једнаких делова. Обука и тестирање се изврше *n* пута и резултати се упросече. У сваком циклусу обука се изврши над *n-1* делова, који се тада посматрају као јединствен скуп, а тестирање над преосталим *n*-тим делом. Уобичајене вредности за *n* у системима класификације текстова су 3, 5 и 10. Како сваки класификациони модел (па и класификатор текстуалних докумената) представља апроксимацију стварних параметара неке реалне колекције, то као резултат сваког класификационог поступка имамо четири могућности:

- да модел обележи документ посматраном класом, а да јој он стварно и припада (тзв. стварно позитивни – true positives TP);
- да модел не обележи документ посматраном класом , а да јој он стварно припада (тзв. тзв. лажно негативни – false negatives FN);
- да модел обележи документ посматраном класом, а да јој он стварно не припада (тзв. лажно позитивни – false positives FP);
- да модел не обележи документ посматраном класом, а да јој он стварно не припада (тзв. стварно негативни – true negatives TP).

На слици 2.9 приказан је однос резултата класификације једног скупа за тестирање и одговарајућег класификационог модела, када се посматра једна класа.



Слика 2.9 Однос резултата класификације помоћу модела и стварне класификације

Однос класификације моделом и стварне класификације често се даје у облику матрице конфузије (енг. confusion matrix) приказане табелом 2.2 где су TP_i – укупан број докумената који припадају класи c_i, а тако су и класификовани, FN_i – укупан број докумената који припадају класи c_i али нису тако класификовани, FP_i – укупан број докумената који не припадају класи c_i али су тако класификовани и TN_i – укупан број докумената који не припадају класи c_i и тако су и класификовани.

Табела 2.2 Матрица конфузије

Класа c_i		Стварно припада класи	
		да	не
Припада класи према моделу	да	TP_i	FP_i
	не	FN_i	TN_i

Параметри који се уводе као метричке оцене ефикасности модела класификације докумената иначе се користе као метрика у проналажењу информација (Manning, Raghavan & Schütze, 2008) и засновани су на матрици конфузије. То су:

- прецизност (енг. precision) P
- одзив (енг. recall) R

- F-мера (енг. F-score, F-measure) F
- тачност (енг. accuracy) acc

Ако, на основу дефиниције 2.4, означимо непознату циљну функцију скупа докумената и скупа класа са: $\breve{\Phi}: D \times C \rightarrow \{true, false\}$, а њену апроксимацију представимо моделом класификације Φ, тада можемо дефинисати прецизност класификације P_{ij} у односу на класу c_i као вероватноћу да је случајан документ d_j класификован у класи c_i под условом да он то стварно и јесте, што се може означити условном вероватноћом (72).

$$P_{ij} = p(\breve{\Phi}(d_j, c_i) = true | \Phi(d_j, c_i) = true) \quad (72)$$

Одзив класификације у ознаци R_{ij} документа d_j у односу на класу c_i дефинишемо као вероватноћу да случајан документ d_j припада класи c_i ако га је класификатор тамо већ сврстао, што се може означити условном вероватноћом (73).

$$R_{ij} = p(\Phi(d_j, c_i) = true | \breve{\Phi}(d_j, c_i) = true) \quad (73)$$

Како су условне вероватноће (72) и (73) непознате, то користимо њихове апроксимације дате са:

$$\widetilde{P_{ij}} = \frac{TP_i}{TP_i + FP_i} \quad \text{и} \quad \widetilde{R_{ij}} = \frac{TP_i}{TP_i + FN_i}$$

Оцена класификатора може се изразити упросечавањем ових параметара на нивоу свих класа и то на два начина: микро-упросечавањем μ (енг. microaveraging) и макро-упросечавањем M (енг. macroaveraging). Код микро-упросечавања израчунавају се јединствене вредности за прецизност и одзив на нивоу класификатора из јединствене матрице конфузије (за све класе истовремено):

$$\tilde{P}^{\mu} = \frac{TP}{TP + FP} = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FP_c} \quad \text{и} \quad \tilde{R}^{\mu} = \frac{TP}{TP + FN} = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FN_c}.$$

Код макро-упросечавања израчунавају се вредности за прецизност и одзив на нивоу сваке класе, а затим се налазе просечне вредности

прецизности и одзива (што је боље користити ако је скуп за тестирање такав да су класе неуједначене):

$$\tilde{P}^M = \frac{\sum_{c\in C}\tilde{P}_c}{|C|} \quad \text{и} \quad \tilde{R}^M = \frac{\sum_{c\in C}\tilde{R}_c}{|C|}.$$

Можемо уочити да је прецизност мера тачности (вероватноћа да је класификација случајног документа тачна), а одзив је мера комплетности (вероватноћа да је документ који припада некој категорији тако и класификован). Ефикасност једног класификационог модела повећава се са повећањем прецизности и одзива.

Мера која представља тежинску хармонијску средину прецизности и одзива назива се F_β -мера. Дефинише се као

$$\mathrm{F}_\beta = \frac{(\beta^2+1)\mathrm{PR}}{\beta^2\mathrm{P}+\mathrm{R}}$$

Када се узме да је $\beta = 1$ (једнака тежина прецизности и одзива) добија се тзв. $F1$ мера тј. $F_1 = \frac{2\,PR}{P+R}$

Мере које се, такође, користе у класификацији текстова су тачност класификације (енг. accuracy) и степен грешке класификације (енг. error rate) (Yang, 1999) који се дефинишу са:

- тачност класификације acc_{ij} је вероватноћа да случајан документ d_j класификован онако како он то стварно и јесте, што се може означити условном вероватноћом (74)

$$acc_{ij} = p\big(\breve{\Phi}(d_j, c_i) = true \big| \Phi(d_j, c_i) = true\big) + p\big(\breve{\Phi}(d_j, c_i) = false | \Phi(d_j, c_i) = false\big) \qquad (74)$$

- степен грешке класификације условном вероватноћом (75)

$$err_{ij} = p\big(\breve{\Phi}(d_j, c_i) = true \big| \Phi(d_j, c_i) = false\big) + p\big(\breve{\Phi}(d_j, c_i) = false | \Phi(d_j, c_i) = true\big) \qquad (75)$$

Тачност и степен грешке класификатора израчунавају се помоћу:

$$\widetilde{Acc} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{и} \quad \widetilde{Err} = \frac{FP+FN}{TP+TN+FP+FN}$$

3. Језички ресурси и алати

У задацима обраде природног језика разноврсне језичке структуре се истражују генерисањем одговарајућих модела и увођењем параметара за процену и праћење тих модела помоћу метода различитих области рачунарских наука и математике: статистике, машинског учења, истраживања података (енг. data mining), препознавања узорака (енг. pattern recognition) (Manning & Schütze, 1999; Manning, Raghavan & Schütze, 2008). Према Менингу (Manning & Schütze, 1999), у обради природног језика статистичке законитости се не изводе посматрањем огромних количина примера примене неког језика, већ се изводе из примера мањег обима какве су колекције текстова. Колекције текстова најчешће се организују у виду корпуса,[45] али могу учествовати и у генерисању других дигиталних језичких структура као што су: онтологије, семантичке мреже, тезауруси, речници, лексикони, лексичке базе и др. који се једним именом називају дигитални језички ресурси (ДЈР). Према дефиницији[46] ELRA-е[47], ДЈР су скупови језичких (писаних и говорних) података и метаподатака представљени у машински читљивим облицима како би били коришћени у изградњи, развоју и унапређењу алгоритама и система из различитих домена: обраде природног језика, електронског издаваштва, локализације софтверских производа, проучавања језика, међународне сарадње и трансакција, итд. Резултати и примена задатака обраде природног језика зависе у великој мери од квалитета и обима расположивих дигиталних ресурса посматраног природног језика као и свеобухватности и флексибилности алата којима се дигитални ресурси обрађују (Nenadić, 2004). Стога, у ширем смислу, ДЈР обухватају и софтверска средства (алате) којима са прикупљају, припремају, складиште, прилагођавају и претражују дигитални језички ресурси. То су најчешће: стемери (енг. stemmers), алати за означавање (енг. taggers), алати за поделу текста на реченице (енг. sentencers), алати за поделу реченица на

[45] од латинске речи *Corpus* (mn. corpora) која значи "тело"
[46] http://www.elra.info/en/about/what-language-resource/
[47] The European Language Resources Association (ELRA) http://www.elra.info/en/about/elra/

речи (енг. tokenizers), алати за препознавање именованих ентитета (енг. named entity recognizers), лематизери (енг. lemmatizers), алати за означавање врстама речи (енг. POS taggers), алати за анализу зависности структура и рашчлањивање (енг. parsers), итд.

3.1 Корпуси

Корпус, као колекција одабраних текстова или говорних записа једног природног језика, мора представљати репрезентативан узорак посматраног домена, како би се анализе изведене над њим могле уопштити у односу на читав репрезентујући домен (Biber & Jones, 2009). Даље, према истим ауторима, анализе текстуалних корпуса могу се поделити у три групе које се разликују по природи објеката посматрања и начинима њихове обраде:

- истраживања и описивања различитих језичких структура (јединица која се посматра је појединачна језичка структура)
- истраживања и описивања текстова и врста текстова (јединица која се посматра је сваки појединачни текст)
- истраживања и описивања самог корпуса или његових делова тј. подкорпуса (јединица која се посматра је сам корпус).

Објекати посматрања у првој групи немају квантитативне карактеристике (нпр. истраживање о врстама везника у условним реченицама), док друге две групе могу мерити квантитативна својства посматраних објеката (нпр. поредити учесталости појављивања придева у новинским и научним чланцима).

Квалитет корпуса мери се репрезентативношћу, али и нивоом претходне обраде текста (Samardžić, 2011). Под претходном обрадом текста подразумева се уношење различитих информација у облику метаподатака о тексту ради касније ефикасније анализе. То могу бити индексне информације које обезбеђују да се свака текстуална јединица касније може наћи на основу индексног упита (нпр. за сваку реч се дефинише у којим текстовима се појављује), информације о структури текста (дефинишу се границе текстова, пасуса, реченица), морфолошке информације о врсти речи, о припадајућој

леми, информације о фреквенцији појављивања речи, информације о пореклу и врсти текста и др. Корпус који садржи неки ниво претходне обраде текста назива се етикетирани корпус. Уколико се корпус складишти у формату који је рачунарски читљив, реч је о електронском (дигиталном) корпусу (Utvić, 2014a) .

Српски језик представљају два репрезентативна електронска етикетирана корпуса:

- Корпус савременог српског језика (Obradović, Popović & Pavlović-Lažetić, 2003; Krstev & Vitas, 2005; Krstev, 2008; Utvić, 2011; Vitas, Krstev, Utvić, 2014a)
- Дијахрони корпус српског језика (Ševa & Kostić, 2003; Kostić, 2014).

Рад на корпусу савременог српског језика (Krstev & Vitas, 2005) *СрпКор* започео је Душко Витас 1981. године, а рад је наставила Група за обраду природног језика Математичког факултета у Београду. 1981. Најновија верзија корпуса, *СрпКор2013*, представља етикетирану колекцију текстова величине 122 милиона корпусних речи. Корпус је сачињен највећим делом од чланака дневних новина (више од 50%), чланака недељних и месечних новинских издања, литерарних текстова, културних додатака, кратких есеја, монографија и осталих текстова[48]. Корпусни текстови носе библиографске и морфолошке ознаке. Заједно са неколико мањих доменских корпуса јавно је доступан[49] од 2002. године и могуће га је претраживати.

Дијахрони корпус српског језика настао је на Филозофском факултету у Београду, у периоду 1957-62. Садржи око 11 милиона речи текстова насталих у периоду од XII века до данас и ручно је морфолошки означен. Електронска верзија генерисана је 2002. али није јавно доступна.

Сем корпуса опште намене, могу постојати доменски, компаративни, паралелни, педагошки и др. Свеобухватан хронолошки приказ развоја корпуса на тлу балканских земаља може се наћи у раду (Dobrić, 2012).

[48] Прецизан састав и графички приказ структуре СрпКор-а може се наћи у раду (Krstev, 2008)
[49] http://www.korpus.matf.bg.ac.rs/prezentacija/korpus.html

Најзад, значај корпуса као дигиталног језичког ресурса није одређен само резултатима у области лингвистике. Он се све више користи као извор информација у психологији, социологији, когнитивној лингвистици, економији, маркетингу, а корпусна лингвистика се, као дисциплина која се бави изградњом, проучавањем и применом корпуса у истраживањима, убрзано развија.

3.2 Електронски речници

Традиционални речници се, у условима интензивног развоја инфраструктуре интернета, све чешће дигитализују и користе интерактивно. Структуру једног речника чине лева и десна страна (Dragićević, 2010). Лева страна речника је одредница (лема или лексичка јединица) и она је предмет дефиниције која се даје на десној страни речника. Структура десне стране може бити различита, али се најчешће као први податак наводи показатељ граматичке категорије (рода, броја, врсте речи, глаголског вида и сл.). Као остали подаци наводе се: дефиниција, синоними (ако постоје), етимологија, квалификатори употребе, примери употребе и др. Иако леву страну речника обично представљају једночлане лексичке јединице, односно леме, оне могу бити и вишечлане. Према (Dragićević, 2010), вишечланим лексичким јединицама сматрају се фразеологизми (изрази), терминолошке синтагме (синтагме) и устаљене фразе. Ипак, одреднице речника могу бити само прва два типа, али не и устаљене фразе. Изрази (нпр. „бела рада“, „мајчина душица“) и синтагме (нпр. „Богу иза ногу“, „ватромет боја“) имају јединствено значење и функцију, синтаксну и семантичку самосталност, па се, стога, у речнику третирају на исти начин као леме. Нпр. израз „до голе коже“ има јединствено значење (потпуно) и врши функцију прилошке одредбе за начин, те одговара леми *потпуно*. С друге стране, устаљене фразе по својој природи јесу реченице (најчешће изреке, пословице, афоризми) и због тога немају јединствено значење, те не могу бити одреднице неког речника.

Поред дигиталних верзија традиционалних речника, креирају се и развијају електронски речници чији је задатак да буду машински читљиви и флексибилни при употреби у апликацијама широке примене. Структура електронских речника се формира тако да се могу применити у решавању проблема синтаксне (нпр. контрола исправности текста и корекција), семантичке (препознавање лингвистичких објеката и њихових хијерархијских структура) и морфолошке природе (означавање врстама речи) (Vitas & Krstev, 2009).

Електронски морфолошки речници су један од ефикасних начина описа флективних парадигми у морфолошки богатим језицима. Електронске речнике српског језика (Vitas, Krstev, Pavlović-Lažetić & Nenadić, 2000; Vitas, Krstev, Obradović, Popović & Pavlović-Lažetić, 2003; Krstev, 2008; Vitas & Krstev, 2009) чини систем речника у коме су све одреднице у потпуности морфолошки дефинисане. Развој овог система заснован је на принципима електронских речника за француски језик чији је развој отпочео 1989. године и приказан је у радовима (Gross, 1989, 1990; Courtois & Silberztein, 1990). На основу тог система развијен је у оквиру мреже RELEX[50] систем морфолошких речника за српски језик, SrpMD[51]. Основне компоненте овог система су: DELAS – речник једночланих лексичких јединица (лема), DELAC – речник вишечланих лексичких јединица, DELAF – речник свих флективних облика речи одредница DELAS речника, DELACF – речник флективних облика вишечланих лексичких јединица (Krstev, Stanković, Obradović, Vitas & Utvić, 2010) и скуп коначних трансдуктора којима се дефинишу и генеришу сви флективни облици у речницима. Систем SrpMD тренутно садржи 148.000 лема и преко 1000 коначних трансдуктора којима се генерише више од 5 милиона DELAF одредница.

Структура DELAS-а (Vitas, Krstev, Obradović, Popović & Pavlović-Lažetić, 2003) је облика:

$$W_{lex}, K_n + [SynSem]$$

[50] http://infolingu.univ-mlv.fr/english/Relex/Relex.html
[51] http://korpus.matf.bg.ac.rs/prezentacija/recnici.html

где је W_{lex} – лексичка јединица (лема), K_n – „врсте речи и кôд флективне класе“ који указује на коначни трансдуктор којим се једнозначно описују особине флективне класе којој припада дата лема W_{lex}, а $SynSem$ скуп синтаксних, семантичких и других ознака леме.

Структура DELAF-a (Vitas, Krstev, Obradović, Popović & Pavlović-Lažetić, 2003; Krstev, 2008) је облика:

$$W_{text}, W_{lex}.K_n[+SynSem]^* : [K_{kat}]^*$$

где су: одредница W_{text} – флективни облик леме W_{lex}, добијен применом правила дефинисаних коначним трансдуктором са симболичком ознаком K_n, скуп синтаксних и семантичких ознака $SynSem$ наслеђен од леме W_{lex} и опис морфосинтаксних категорија K_{kat}.

На пример, за одредницу $ljubav, N696$ која је у речнику DELAS и односи се на лему *ljubav*, у речнику DELAF се, применом коначног трансдуктора $N696$, генерише 15 флективних форми, односно 4 различита облика речи:

$ljubavima, ljubav.N696: fp3q: fd6q: fp7q$
$ljubavlju, ljubav.N696: fs6q$
$ljubavi, ljubav.N696: fs2q: fs3q: fs5q: fs6q: fs7q: fp1q: fp2q: fp4q: fp5q$
$ljubav, ljubav.N696: fs1q: fs4q$

На сличан начин дефинишу се и вишечлане лексичке јединице у речницима DELAC и DELACF. Пример генерисања флективних форми вишечланих лексичких јединица дат је у одељку 5.3.2.

3.3 Семантичке мреже

Семантичка мрежа представља један од видова репрезентације знања. Сам назив указује на мрежну (графовску) структуру чији су елементи концепти, објекти или догађаји повезани релацијама на основу семантичких правила. Усвојени ниво семантике одређује комплексност мреже, а релације између елемената генеришу се на основу доменских знања. Прве семантичке мреже развијане су да би се имплементирале у задацима примене вештачке интелигенције (Sowa, 1992). Постоји више врста семантичких мрежа, али заједничке особине су им да представљају начин репрезентације знања у

облику графова и да налазе примену у системима за аутоматско закључивање у домену који репрезентују. Једна од најразвијенијих и најкоришћенијих семантичких мрежа данас је Принстонски ворднет.

3.3.1 Принстонски ворднет

Рад на развоју Принстонског ворднета[52] (Princeton WordNet - PWN) је започет 1985. на иницијативу Џорџа Милера[53] (Miller, 1995; Fellbaum, 1998) са Универзитета Принстон, САД. Прва јавно доступна верзија, PWN 1.0, објављена је 1991. године, а тренутно актуелна верзија 3.1 може се претраживати помоћу онлајн алата[54] или преузети.[55] То је дигитални ресурс који се најчешће посматра као лексичко-семантичка мрежа. Изграђен је уз поштовање три принципа:

1. принципа одвојивости (separability) који подразумева да се сваки лексички елемент може посматрати и проучавати одвојено од других;
2. принципа примене образаца (patterning) што значи да се лексичко знање не формира посматрањем појединачних речи, већ се на основу системских образаца који постоје у језику креирају концепти и утврђују релације међу њима;
3. принципа свеобухватности (comprehensiveness) који се заснива на неопходности изградње таквог дигиталног лексичког ресурса који ће својом величином задовољити потребе различитих задатака обраде природног језика.

На самом почетку развоја PWN, Милер је креирао мали скуп од 45 именица које је повезао релацијама и тако је настала прва мрежа ове врсте. У даљем развоју је коришћен Браунов корпус (*The Brown Corpus of Standard American English*),[56] различити речници синонима и антонима, а коришћен је

[52] https://wordnet.princeton.edu/
[53] George A. Miller
[54] http://wordnetweb.princeton.edu/perl/webwn
[55] https://wordnet.princeton.edu/wordnet/download/current-version/
[56] www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private /brown/ brown.html

и лексикон COMLEX (Computational Lexicon)[57] величине преко 38 хиљада речи чији су дизајн и изградња приказани у раду (Grishman et al., 1994). Према Милеру (Miller et al., 1993, pp. 3) најважније својство ворднета као речника је организација лексичких информација у смислу значења, а не облика речи. За разлику од класичног речника, ворднет не пружа информације о изговору, односно акцентовању, етимологији, флективним и деривационим облицима и начинима употребе дефинисаних појмова. Основна јединица у ворднету је синсет (енг. synonimous set – synset) и њоме се дефинише један концепт. Сваком концепту се придружују његове могуће лексикализације – синонимске леме или литерали. Стога, ворднет представља колекцију оних литерала који носе семантичку информацију – именица, глагола, придева и прилога. Осим синонимских лема, синсет садржи и информацију о врсти речи којом се описује концепт, његову дефиницију и примере употребе у природном језику. Ворднет је и семантички лексикон јер садржи и описује односе (везе или релације) између синсетова тј. концепата. Такође, он садржи и информације о семантичким везама хипонимије/хиперонимије између синсетова, па се може посматрати и као тезаурус, односно хијерархијски уређен речник синонима и асоцијативних појмова. Релације које се успостављају у ворднету могу бити семантичке и оне постоје између синсетова и могу бити лексичке када се њима повезују лексичке јединице тј. литерали. Зависно од врсте речи дефинишу се различите релације. У прилогу 3.15 дат је табеларни приказан свих релација у PWN у коме је свака релација описана врстом везе коју представља, симболичком ознаком релације, синтаксним категоријама над којима се дефинише као и скупом еквивалентних релација у ресурсима сличне намене – *EuroWordNet*-у (EWN) и Српском ворднету (SWN).[58] Неке од најважнијих семантичких релација граде се између синсетова именица и то су релације: синонимије, антонимије, хипонимије/хиперонимије и меронимије/холонимије. Код глагола то су: тропонимија, импликација и каузалност. Међу придевима и прилозима

[57] http://nlp.cs.nyu.edu/comlex/
[58] О Српском ворднету биће више речи у одељку 3.3.3.

најчешће се користе релације између лексичких репрезентација синсетова као што је релација деривације.

Релација хипернимије (*hypernymy*) креира се између два синсета који припадају синтаксној категорији именица, од којих први представља специјализацију другог концепта. Релација се успоставља између скупова синонимских лема оба концепта, односно важи за све њихове синонимске леме, као у следећим примерима:

{*leopard*, *Panthera_pardus*}@{*big_cat*, *cat*}

{*big_cat*, *cat*}@{*feline*, *felid*}

где је @ ознака релације хипернимије.

Релација хипернимије има инверзну релацију, релацију хипонимије (*hyponymy*), којом се дефинише веза између два именичка синсета од којих први представља генерализацију другог концепта. Ознака ове релације је ~, а примери употребе су:

{*big_cat*, *cat*}~{*leopard*, *Panthera_pardus*}

{*feline*, *felid*}~{*big_cat*, *cat*}.

Релацијама хипонимије/хиперонимије генеришу се структуре стабала које у процесу генерализације воде до синсетова који немају надређених хиперонимских синсетова и који се зову почетни синсетови или *unique beginners*. У тренутно актуелној верзији PWN-а постоји 25 именских почетних синсетова којима се описују најопштији концепти.

Другу групу хијерархијских именских релација чине релације меронимије/холонимије којима се описује релације дела и целине, члана и групе, састојка и структуре. Релација холонимије успоставља се између два именска синсета од којих први означава концепт који је део другог концепта. Зависно од природе те везе, у PWN-у се дефинишу три подврсте холонимије где концепт, дефинисан посматраним синсетом, може бити:

1. део неке функционалне целине (PART OF) са симболичком ознаком *%p* као у примерима: {*limb*, *tree branch*}*%p*{*tree*}; {*finger*}*%p*{*hand*, *manus*, *mitt*, *paw*};
2. члан неког скупа (MEMBER OF) чија је симболичка ознака у PWN-у дата са *%m*, као у примеру {*tree*}*%m*{*forest, wood, woods*};

3. градивни састојак нечега (SUBSTANCE OF) чија је симболичка ознака у PWN-у дата са *%s*, као у примеру {*flour*}*%s*{*bread, breadstuff, staff of life*}.

Релација холонимије може имати инверзну релацију, релацију меронимије, али не безусловно. На пример, док за тврђење „хлеб се састоји од брашна“, важи и обратно — „брашно је састојак хлеба“, дотле за тврђење „шума се састоји од стабала“, не мора важити обратно, тј. не мора свако дрво бити део шуме. Уколико постоји, релација меронимије се успоставља између два именска синсета од којих први означава концепт који је целина којој на неки начин припада други концепт. Аналогно холонимији, дефинишу се три подврсте меронимије и обележавају се симболичким ознакама: *#p, #m, #s*.

Између одређених именичких синсетова може бити дефинисана и релација антонимије чија је симболичка ознака знак узвика (!), а примена дата у следећим примерима: {*man*}*!*{*woman*}; {*beauty*}*!*{*ugliness*}. Како је релација антонимије симетрична, важи и обрнуто: {*woman*}*!*{*man*}; {*ugliness*}*!*{*beauty*}.

Релације тропонимије (*troponymy*) са ознаком @, импликације (*entailment*) са ознаком * и каузалности (*causes*) са ознаком > дефинишу се као семантичке релације између синсетова глагола. Релација тропонимије еквивалентна је релацији хиперонимије код именских синсетова, па им је симболичка ознака иста. Хијерархијска стабла која граде синсетови глагола релацијом тропонимије обично су знатно краћа од оних која граде синсетови именица. Релација импликације (*entailment*) може бити двојака. У једном случају ради се о истовремености дешавања (*temporal inclusion*) и тада релација импликације представља релацију између два глаголска синсета од којих први означава радњу које се дешава само уколико се у истом временском интервалу дешава и радња коју означава други синсет, као у примеру {*snore*, *saw wood*, *saw logs*}*{*sleep*, *kip*, *slumber*, *log Z's*, *catch some Z's*}, где је јасно да хркање није могуће без истовременог спавања. У другом случају ради се о временској редоследности дешавања или тзв. *temporal exclusion* и тада релација импликације представља релацију између два глаголска синсета од којих први означава радњу које се дешава само уколико се у претходно десила радња коју означава други синсет, као у примеру

{*win*}*{*compete, vie, contend*}, где побеђивању претходи такмичење. Посебан случај релације импликације је релација каузалности (*causes*) између два глаголска синсета од којих други означава радњу које се дешава само уколико се десила радња коју означава први синсет, као у примеру {*show*}>{*see*}.

Кад је реч о придевима, они се у PWN деле у две групе: дескриптивне (*descriptive*) и релационе (*relational*). Дескриптивни придеви се међусобно могу повезивати релацијама антонимије. Антонимија код придева може бити директна и тада се синсетови чије лексичке репрезентације имају директне антониме називају главним (*head*) синсетовима. Антонимија може бити и индиректна и тада се остварује посредно. Синсетови придева чији литерали немају директне антониме чине посебну групу тзв. придева-сателита (*satellite adjectives*) тј. пратећих придева. Такви синсетови се групишу у кластере синонимских синсетова тако што се сваки од њих релацијом типа *similar* повеже са синонимским главним (*head*) синсетом који, с друге стране, релацијом антонимије остварује везу са антонимским кластером. Осим међусобно, литерали синсетова дескриптивних придева повезују се релацијом *attribute* са литералима именских синсетова. Релациони придеви повезују се релацијама типа *derived* са лексичким репрезентацијама именских синсетова од којих настају деривационим правилима. На пример у релацији {generational}\{generation} придев „генерацијски“ (generational) настао је деривацијом из именице „поколење“ (generation). Прилози се повезују са придевима релацијом деривације, а међусобно релацијом антонимије.

Године 1989. развојни тим PWN-а одлучио је да учини помак у смислу морфолошке доградње овог ресурса информацијама које су се односиле на препознавање флективних форми. Иако PWN садржи само основне облике речи (леме), претрага овог ресурса се може вршити упитом који садржи било који флективни облик. Програмски модул *morphy* дефинисан је као скуп морфолошких функција којима се упит обрађује тако што се, применом дефинисаних морфолошких правила и њихових изузетака долази до

основног облика жељене лексичке јединице. Модул *morphy* користи два типа додатних ресурса:

1. листу суфикса и правила њихове трансформације како би се извео основни облик лексичке јединице; листа се дефинише за сваку синтактичку категорију посебно;
2. листу изузетака од правила дефинисаних у листи суфикса где се сваки изузетак дефинише понаособ; листа се дефинише за сваку синтактичку категорију посебно и носи екстензију *exc* (*noun.exc, verb.exc,...*).

Нека од тих правила која се налазе у листама суфикса дата су у табели 3.1, док су у табели 3.2 дати примери изузетака.[59]

Табела 3.1 Примери синтаксних правила свођења флективних облика именица, глагола и придева на основни облик речи, који се могу наћи у одговарајућим листама суфикса

Синтаксна категорија	**Суфикс**	**Правило трансформације**
Noun	“s”	“”
Noun	“ses”	“s”
Noun	“xes”	“x”
Verb	“ies”	“y”
Verb	“ing”	“e”
Adj	“er”	“”
Adj	“est”	“”

Табела 3.2 Примери изузетака од синтаксних правила свођења флективних облика именица, глагола и придева на основни облик речи који се налазе у одговарајућим листама изузетака

Синтаксна категорија	**Флективни облик**	**Основни облик лексичке јединице**
Noun	abaci	abacus
Noun	fora	forum
Noun	wives	wife
Verb	held	hold
Verb	ran	run
Adj	bigger	big
Adv	best	well

[59] Иако се примери односе на просте речи, иста правила примењују се и над вишечланим лексичким јединицама које имају улогу синонимских лема у PWN.

3.3.1.1 Структура Принстонског ворднета

Структура[60] *PWN* је таква да су синтаксне категорије физички одвојене, односно синсетови чији су литерали исте синтаксне категорије налазе се у једној датотеци. Унутар физичких датотека груписање је остварено и логички - у лексикографским датотекама и унутар њих - у синсетовима. Синсет физички преставља један ред текстуалне датотеке, а логички преставља најмању семантичку целину која означава неки појам или концепт.

PWN базу сачињавају:

- 8 текстуалних датотека - *index.noun, data.noun, index.verb, data.verb, index.adj, data.adj, index.adv, data.adv*
- 4 листе морфолошких изузетака - *noun.exc, verb.exc, adj.exc, adv.exc*
- 2 додатне текстуалне датотеке које садрже додатна објашњења употребе неких глагола - *sentidx.vrb, sents.vrb*.

Свака од синтаксних категорија представљена је помоћу две датотеке: *index.kategorija* и *data.kategorija* где *kategorija* може бити именица(*noun*), глагол(*verb*), придев(*adj*) или прилог(*adv*). Остале помоћне датотеке користе се у функцијама претраживања. Свака *index.kategorija* датотека представља алфабетски уређену листу свих појмова дате категорије. У сваком реду датотеке се исписују редом: реч у форми леме и листа оних синсетова у којима се лема налази.

Датотеке података или лексикографске датотеке (има их 45), представљају логичке целине у које се групишу синсетови. Назив лексикографске датотеке је у формату

pos.suffix

где су:

pos –синтаксна категорија (noun, verb, adj, adv)

[60] Структура PWN коју разматрамо и користимо у овој тези односи се на верзију PWN 3.0 https://wordnet.princeton.edu/wordnet/download/

suffix - логичка категорија у коју природно може да се смести дата реч (animal, plant, event,...).

Свака лексикографска датотека има јединствени идентификатор, тзв. *FileNumber*. Табела свих лексикографских датотека и категорија којима припадају дата је у прилогу 3.1.

3.3.1.2 Формат датотека типа Index

Свака датотека типа *Index* почиње информацијама о правима коришћења, броју верзије и текстом лиценце. Софтвер за претраживање прескаче ове редове, имајући у виду да почињу са 2 бланко знака и редним бројем дате линије. Сви остали редови у индексној датотеци имају истоветан формат и односе се на појмове садржане у аналогној датотеци типа *Data*. Формат, односно структура једног реда индексне датотеке је:

lemma pos synset_cnt p_cnt [ptr_symbol...] sense_cnt tagsense_cnt synset_offset [synset_offset...]

Пример:

arm n 6 5 @ ~ #p %p + 6 4 05563770 02737833 04565375 02737660 08401248 04236377

где су:

lemma – реч или фраза записана малим словима у форми леме (фразе које садрже више од једне речи користе карактер „_“ за повезивање). У примеру arm

pos - синтаксна категорија, односно ознака физичке датотеке којој реч припада по својој природи: *n* именице, *v* глаголи, *a* придеви, *r* прилози. У примеру n

synset_cnt - број синсетова у којима се реч налази. То је истовремено и број различитих значења која дата реч има. У примеру 6

p_cnt - укупан број различитих симбола показивача (поинтера), односно релација које дата реч има са другим синсетовима, узимајући у обзир све синсетове у којима се налази. У примеру 5

ptr_*symbol* - листа симбола показивача,[61] раздвојених бланко карактером, димензије *p_cnt*. Уколико дата реч нема ниједан показивач, ни у једном синсету, онда се ово поље изоставља, а вредност поља *p_cnt* је 0.

[61] Табела симбола показивача дата је у Прилогу 3.2, а улога показивача детаљније објашњена у одељку 3.3.1.4

У примеру `@ ~ #p %p +`

sense_cnt – вредност је иста као *synset_cnt*. Ова вредност се непотребно понавља, али је укључена због компатабилности с претходном верзијама. У примеру `6`

tagsense_cnt – број значења дате речи добијених према њиховој фреквенцији појављивања[62] у текстовима са ознакама конкорданци. У примеру `4`

synset_offset – листа ознака свих синсетова (*synset_offset*) у датотеци *data.kategorija* који садрже дату реч. Свака ознака синсета је 8-цифрени број са укљученим водећим нулама и има јединствену вредност у посматраној датотеци. У примеру `05563770 02737833 04565375 02737660 08401248 04236377`

Реч *arm* налази се и у синтаксној категорији глагола, односно у датотеци *index.verb* и описана је помоћу:

`arm v 2 4 ! @ ~ + 2 2 01087197 02334867`

На сајту Друштва за језичке ресурсе и технологије[63] постоји систем за упоредну претрагу Принстонског и Српског ворднета[64]. У прилогу 3.3 дат је пример претраге речи *arm* у PWN где су, као резултат претраге, приказани сви синсетови који садрже дату реч, груписани по синтаксним категоријама.

3.3.1.3 Формат датотека типа *Data*

Свака датотека типа *Data* почиње истим информацијама као и *Index*. Сви остали редови имају истоветан формат и односе се на концепте односно синсетове. Формат, односно структура једног реда датотеке података је:

synset_offset lex_filenum ss_type w_cnt word lex_id [word lex_id...] p_cnt [ptr...] [frames...] | gloss

Пример:

05563770 08 n 01 arm 0 015 @ 05560244 n 0000 #p 05216365 n 0000 #p 02472293 n 0000 + 02739427 n 0101 %p 05338614 n 0000 %p 05361123 n 0000 %p 05564323 n 0000 %p 05564590 n 0000 %p 05568767 n 0000 %p 05579436 n 0000 %p 05579753 n 0000 %p 05579944 n 0000 %p 05584928 n 0000 %p 05593017 n 0000 %p 05593181 n 0000 | a human limb; technically the part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb

[62] У *PWN* значења (senses) једне речи се опадајуће сортирају према фреквенцији појављивања, па најфреквентнија реч добија вредност 1
[63] http://www.jerteh.rs/
[64] Веб апликација за претрагу је креирана у склопу активности и у току рада на овој тези.

synset_offset – ознака синсета као 8-цифрени број са водећим нулама. У примеру 05563770

lex_filenum – двоцифрен број који представља јединствени *FileNumber* лексикографске датотеке у којој је садржан дати синсет. У примеру 08

ss_type – ознака синтаксне категорије којој припада дати синсет: n именица, v глагол, a придев, r прилог, s пратећи придев. У примеру n

w_cnt - двоцифрен хексадецимални број који означава укупан број речи или фраза које међусобно представљају синониме у текућем синсету. У примеру 01

word[65] - листа речи или фраза које међусобно представљају синониме и чији је број дефинисан параметром *w_cnt*, написаних у ASCII формату и одвојених бланко знаком, сем речи које заједно чине фразу и које се у том случају међусобно повезују знаком _ . Речи се исписују малим и великим словима уз поштовање синтаксних правила (Gresham's_Law) за разлику од индексне датотеке где се исписују искључиво малим словима (gresham's_law). У *data.adj* датотеци, иза речи или фразе може уследити синтаксни маркер. Синтаксни маркер[66] је опциона ознака и ако се додаје, пише се унутар малих заграда непосредно иза речи и без размака у односу на исту. У примеру arm

lex_id – једноцифрени хексадецимални број (декадно од 0 до 15) који, када се додаје речи (леми), једнозначно идентификује њено значење на основу припадности истој лексикографској датотеци. Бројеви *lex_id* обично почињу од 0 и повећавају се за 1, како се датој речи додају значења унутар исте лексикографске датотеке, мада то није обавезно, већ је важно да се вредности не понављају за дату реч. Вредност 0 је подразумевана (default) што значи да реч има само једно значење у једној лексикографској датотеци. У примеру 0

p_cnt – Троцифрени цео број написан са водећим нулама који означава укупан број показивача датог синсета на друге синсетове. Вредност 000 указује да дати синсет нема показиваче. У примеру 015

ptr – Показивач са датог на други синсет. Поље *ptr* се даје у формату:

pointer_symbol synset_offset pos source/target

где су:

pointer_symbol [67]– (в. *ptr_symbol* у одељку 3.3.1.2)

synset_offset - јединствени идентификатор синсета (тзв. *byte offset*) на кога указује показивач

[65] Синтакса поља *Word* дата је у одељку 3.3.1.5
[66] Синтаксни маркери описани су у одељку 3.3.1.6
[67] Табела симбола показивача дата је у Прилогу 3.2

pos – ознака синтаксне категорије којој припада синсет на кога се указује (в. *ss_type* у истом одељку)

source/target - поље које указује на разлику између лексичких и семантичких показивача, односно којим се дефинише природа релације као семантичке или лексичке. Састоји се од два хексадецимална броја од којих први означава редни број речи у текућем синсету (рачунајући слева удесно и почев од 1), а други редни број речи у одредишном синсету на који показивач указује. Вредност 0000 указује да показивач представља семантичку везу између датих синсетова. У датом примеру постоји 15 показивача, чиме је дефинисано 15 различитих релација између синсета датог у примеру и синсетова наведених унутар поља показивача. Овде их дајемо у табели 3.3. Сви показивачи су семантички осим показивача с редним бројем 4 (деривациони показивач).

Табела 3.3 Показивачи[68] са синсета *arm* на друге синсетове PWN

Рбр	Показивач *ptr*
1	@ 05560244 n 0000
2	#p 05216365 n 0000
3	#p 02472293 n 0000
4	+ 02739427 n 0101
5	%p 05338614 n 0000
6	%p 05361123 n 0000
7	%p 05564323 n 0000
8	%p 05564590 n 0000
9	%p 05568767 n 0000
10	%p 05579436 n 0000
11	%p 05579753 n 0000
12	%p 05579944 n 0000
13	%p 05584928 n 0000
14	%p 05593017 n 0000
15	%p 05593181 n 0000

frames – користи се само у датотеци *data.verb* и представља листу генеричких оквира[69] и веза сваког од оквира са речима датог синсета.

gloss – Сваки синсет садржи текстуално појашњење концепта који описује. Почиње знаком | иза кога следи текст са структуром реченица природног језика до краја датог реда. Реченице се међусобно одвајају знаком тачка-зарез (;) као у примеру a human limb; technically the part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb

[68] Уочити да је у овом примеру само релација под ред. бројем 4 дефинисана као лексичка
[69] Структура генеричких оквира (*frames*) дата је у одељку 3.3.1.7

3.3.1.4 Показивачи (pointers)

Сврха показивача је да генерише релацију између два синсета (концептуална релација) или између речи које припадају различитим синсетовима (лексичка релација). Лексичке релације се исказују показивачима: *Antonym, Pertainym, Participle, Also See, Derivationally Related*. Остали показивачи односе се на концептуалне релације. Неке од наведених релација су симетричне – ако један синсет има показивач на други синсет, у случају симетричности релације, мора и други синсет да има одговарајући показивач на први. Листа парова симетричних релација дата је у прилогу 3.4.

Најважнија лексичка релација је релација синонимије (*synonimy*) при чему се исказује имплицитно, односно све речи које су међусобно повезане овом релацијом су чланови истог синсета {*big*, *large*} (велики, пространа), {*car*, *auto*, *automobile*} (кола, ауто, аутомобил) и сл. Друга важна лексичка релација је релација антонимије (*antonymy*) и представља однос два појма супротног значења {*open*, *close*} (отворити, затворити), {*dry*, *wet*} (сув, мокар), {*hot*, *cold*} (вруће, хладно) и сл. Концептуалне релације могу се представити хијерархијски - структуром стабла. Листа свих релација установљених у PWN дата је у прилогу 3.2, а описи у прилогу 3.5.

3.3.1.5 Речи (words)

Сваки синсет мора садржати најмање једну лексикографску репрезентацију концепта који описује. Уколико постоји више синонима, они се исписују у формату *речи (word)* као листа синонимских лема које се наводе сукцесивно и са бланко размаком. Синтакса поља *речи (word)* дата је форматом:

word [(marker)] [lex_id] [word[(marker)][lex_id]]*

као у примеру

```
00142622 00 s 02 equipped 0 weaponed 0 001 & 00142407 a 0000 | carrying weapons
```

где је *word* обавезно поље, а *marker* и *lex_id* опциона. У датом примеру синсет са јединственим идентификатором синсета 00142622 из датотеке са бројем 00 који припада категорији пратећих придева (*adjective satellites*) s садржи 02 придева equipped и weaponed.

Реч *word* може бити унета коришћењем и малих и великих слова осим у скупу (кластеру) придева[70]. Речи једне фразе међусобно се повезују карактером „_“. Могу се уносити и бројеви самостално или као делови речи. Уколико се употреби, вредност поља *lex_id* је између 1 и 15. Тиме се даје могућност да једна реч има до 15 различитих значења унутар исте лексикографске датотеке. Вредност се одређује ручно и додељује је лексикограф према фреквенцији појављивања (в. *tagsense_cnt* у одељку 3.3.1.2 и *lex_id* у одељку 3.3.1.3). Подразумевана вредност је 0 и то значи да дата реч има јединствено значење у датој лексикографској датотеци.

3.3.1.6 Маркери (markers)

Речи које представљају придеве могу бити означене синтаксним маркером (*syntactic marker*) који представља ограничење које придев може имати у релацији са именицом коју ближе одређује. У том случају се маркер ставља непосредно иза речи-придева. Вредности синтаксних маркера могу бити:

(p) predicate position (предикатска улога придева)
(a) prenominal (attributive) position (атрибутска улога придева)
(ip) immediately postnominal position (атрибутска улога придева на позицији иза речи коју модификује).

Пример употребе маркера (p) у синсету са два придевска литерала који се користе у улози предиката:

```
00026196 00 s 02 used_to(p) 0 wont_to(p) 0 001 & 00025994 a 0000 | in the habit; "I am used to hitchhiking"; "you'll get used to the idea"; "...was wont to complain that this is a cold world"- Henry David Thoreau
```

[70] Синтакса пратећих придева дата је у одељку 3.3.1.3 у опису формата поља *word.*

Најзад, размотримо комплетну синтаксу којом се описује синсет придева *able* "способан" :

00001740 **00** a **01** able **0** **005** = 05138679 n 0000 = 05546715 n 0000 + 05546715 n **0101** + 05138679 n **0101** ! 00002098 a **0101** | (usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project"

Синсет 00001740, смештен у лексикографској датотеци 00, физичке датотеке *data.adj*, описан једном 01 речи, где реч *able* има једно значење 0, 005 релација, од којих су: 4 према именицама n (2 су семантичке, означене са = и представљају релацију „атрибут", а друге 2 су лексичке-деривационе и означене са + и успостављене између првих речи 0101 полазног синсета 00001740 и одредишних синсетова 05546715 и 05138679 респективно) и једна према придевима a (лексичка-антоним означена са ! где се релација успоставља између првих речи 0101 полазног синсета 00001740 и одредишног синсета 00002098).

3.3.1.7 Глаголски оквири (verb frames)

Сваки глаголски синсет у PWN садржи листу општих правила употребе датог глагола. Свако правило означено је редним бројем и дефинише један оквир употребе као и пример употребе. На пример, за глагол *see* (видети) и његово четврто значење важе два оквира (08 и 26) за које су дати и примери употребе:

see, visualize, visualise, envision, project, fancy, figure, picture, image -- (imagine; conceive of; see in one's mind; "I can't see him on horseback!"; "I can see what will happen"; "I can see a risk in this strategy")
08 Somebody ----s something
26 Somebody ----s that CLAUSE

Глаголски оквир даје се у формату:

f_cnt + f_num w_num [+ f_num w_num...]

где је *f_cnt* двоцифрени цео број који означава укупан број генеричких глаголских оквира на које се ослања дати синсет, *f_num* двоцифрен цео број

који означава редни број генеричког оквира, *w_num* двоцифрен цео број који означава редни број речи у синсету (рачунајући слева удесно и почев од броја 1) на који се генерички оквир односи. Ако *w_num* има вредност 00, онда се дати генерички оквир односи на све речи текућег синсета. Сваком пару *f_num, w_num* претходи карактер + .

Пример:

01087197 33 v 04 arm 0 build_up 0 fortify 0 gird 0 008 + 05635624 n 0301 + 03420559 n 0302 + 08197742 n 0101 + 01156899 n 0102 + 04565375 n 0102 ! 01087835 v 0101 ~ 01087559 v 0000 ~ 01087729 v 0000 01 + 02 00 | prepare oneself for a military confrontation; "The U.S. is girding for a conflict in the Middle East"; "troops are building up on the Iraqi border"

У датом примеру, синсет $\{arm, build\_up, fortify, gird\}$ са значењем „наоружати“ у пољу за генерички оквир има вредност + 02 00 што се интерпретира тако да се све речи синсета (*arm, build_up, fortify, gird*) ослањају на оквир са редним бројем 02 (в. листу генеричких глаголских оквира у прилогу 3.6).

3.3.2 Проширења Принстонског ворднета

PWN је широко коришћен и значајан језички ресурс и данас за њега постоји велики број додатака, мањих ресурса насталих из истраживачких пројеката у различитим областима[71]. Заједничке карактеристике оваквих додатака су да представљају доменска проширења PWN која су поравната на нивоу синсетова. Један од првих ресурса те врсте је *WordNet-Domains*[72] којим се сваки PWN синсет означава[73] бар једном ознаком из хијерархије семантичких домена који садржи 164 лабеле. Овај јавно доступан[74] ресурс (Bentivogli, Forner, Magnini & Pianta, 2004) креиран је на основу Џуијеве децималне класификације (*Dewey Decimal Classification* - DDC) која представља општеприхваћени систем за класификацију знања у библиотечким системима.

[71] https://wordnet.princeton.edu/wordnet/related-projects/
[72] http://wndomains.fbk.eu/
[73] Према (Magnini & Cavaglià, 2000) 96% свих PWN синсетова је означено лабелом домена.
[74] http://wndomains.fbk.eu/download.html

GeoWordNet (Giunchiglia, Maltese, Farazi & Dutta, 2010) је геолокацијски, јавно доступан[75] доменски ресурс заснован на PWN који укључује знања сервиса *Geonames*[76] за генерисање информације о географској ширини, дужини, надморској висини и алтернативним називима на различитим природним језицима, за сваки литерал који је по својој природи географски појам.

TempoWordNet (Dias, Hasanuzzaman, Ferrari & Mathet, 2014) је, такође, јавно доступан[77], лингвистички ресурс који најпре оцењује могућност сваког литерала да искаже неки вид временске информације додељивањем етикете *atemporal* сваком синсету, а затим оцењује степен изражавања те информације у погледу времена вршења радње додељујући 3 додатне етикете: *past, present, future*.

Интеграцијом PWN и формализованих знања из области психологије и психолингвистике развијени су ресурси попут *Sentiwordnet*[78] и *WordNet-Affect*[79], о чему ће више речи бити у поглављу 4.

Најзад, PWN се на различите начине интегрише и са формалним онтологијама различите намене и свеобухватности. Једна од широко примењиваних и свеобухватних је онтологија SUMO (*Suggested Upper Merged Ontology*), о чему ће бити више речи у одељку 3.4.2.

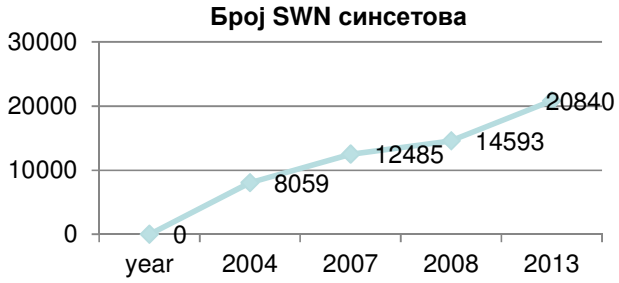
3.3.3 Српски ворднет

Принстонски ворднет извршио је снажан подстицај изградњи лексичко-семантичких мрежа и за друге језике. Шта више, у неким пројектима се радило на развоју вишејезичних и синхронизованих мрежа овог типа. Еуроворднет[80] (EWN) (Vossen, 1998a, 1998b) је развијан као вишејезични ворднет у оквиру европског пројекта за развој језичких ресурса од 1996. до 1999. године и обухватао је седам европских језика чију међусобну

[75] http://datahub.io/dataset/geowordnet
[76] www.geonames.org
[77] https://tempowordnet.greyc.fr/
[78] http://sentiwordnet.isti.cnr.it/
[79] http://wndomains.fbk.eu/wnaffect.html
[80] EuroWordNet http://www.illc.uva.nl/EuroWordNet/

повезаност обезбеђује међујезички индекс - *Interlingual Index* (ILI). Иако се ослања на PWN, EWN користи XML формат и уводи низ нових релација*: xpos_near_synonym, xpos_antonym, role, involved, be_in_state, near_antonym*, и др. (Vossen, 1997). На темељу достигнућа семантичке мреже Еуроворднет, у склопу пројекта Балканет (Stamou et al., 2002) који је трајао у периоду 2001-2004, развијена је вишејезична лексичко-семантичка ворднет мрежа (BWN) језика 5 балканских земаља (Бугарске, Грчке, Румуније, Србије и Турске) и Чешке. Балканет мрежа заснована је на EWN XML структури података. Иако се све појединачне ворднет мреже Балканета данас развијају одвојено, основна структура је иста, усвојена у време зачетка пројекта.

Српски ворднет (SWN) (Krstev, Pavlović-Lažetić, Vitas & Obradović, 2004; Krstev, 2008; Krstev et al., 2008) започео је свој развој у склопу пројекта Балканет и на његовом завршетку 2004. достигао је величину од око 8000 синсетова. Након тога настављен је његов развој, нарочито у доменима биологије, биомедицине, психолингвистике, гастрономије и др. (слика 3.1).

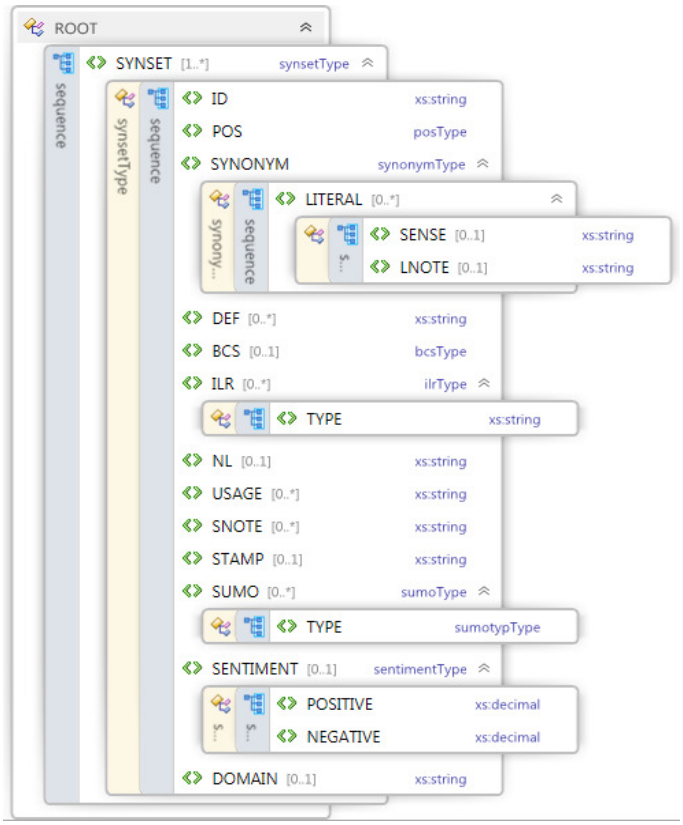


Слика 3.1 Динамика развоја Српског ворднета

Као лексички ресурс SWN је примењен у истраживањима вишечланих лексичких јединица (Krstev, Stanković, Obradović, Vitas & Utvić, 2010; Mitrović, Mladenović & Krstev, 2015), класификацији текста (Pavlović-Lažetić & Tomašević, 2010), претрази вишејезичних дигиталних база података (Stanković, Krstev, Obradović, Trtovac & Utvić, 2012), препознавању реторичких фигура (Mitrović, 2014), анализи осећања израженог у тексту (Mitrović, Mladenović, Krstev & Vitas, 2015) и др.

Детаљан приказ структуре и фаза развоја SWN може се наћи у радовима (Krstev, Pavlović-Lažetić, Vitas & Obradović, 2004; Krstev, 2008; Koeva, Krstev &

Vitas, 2008). Прва XSD схема дефинисана за SWN XML структуру приказана је у раду Крстев и сарадника (Krstev, Vitas, Stanković, Obradović & Pavlović-Lažetić, 2004). Користио ју је софтверски алат *LeXimir* (претходни назив *ILReMat*) који је повезивао *VisDic*[81] (алат коришћен за развој свих ворднет мреже Балканета) са електронским морфолошким речницима српског језика. Међутим, *VisDic* није користио XSD схему у процесу валидације XML структуре, што је отклоњено тек креирањем нових валидационих алата и нове веб апликације (Mladenović, Mitrović & Krstev 2014) за развој SWN мреже (в. одељак 3.5). На слици 3.2 дата је графичка репрезентација актуелне верзије SWN XSD схеме, а у прилогу 3.7 и њен опис.



Слика 3.2 XSD схема SWN XML

Српски ворднет је повезан са неколико дигиталних језичких ресурса што даје већи значај овом ресурсу у задацима обраде природног језика, о чему ће бити више речи у одељку 3.5 и поглављима 5 и 6.

[81] http://nlp.fi.muni.cz/projects/visdic/

Везу са електронским морфолошким речницима SWN постиже елементом *LNOTE* који је саставни део елемента *LITERAL* и чији садржај је идентичан вредности K_n - коду флективне класе оне одреднице у *DELAS* речнику (в. одељак 3.2) која је једнака вредности елемента *LITERAL*, односно синонимској леми посматраног синсета. Може се рећи да је *литерал*, садржан у елементу *LITERAL,* лема која представља лексичку репрезентацију концепта у коме се налази. Детаљнијије информације о природи и реализацији ове везе могу се наћи у радовима (Mladenović, Mitrović & Krstev 2014; Mladenović & Mitrović, 2014).

С обзиром да је SWN поравнат са PWN на нивоу ILI, сви ресурси настали из корелације са PWN могу се користити на исти начин, а могу се и физички интегрисати у структуру SWN, што је и учињено са: *WordNet-Domains, Sentiwordnet* и SUMO, о чему ће детаљније бити речи у одељку 3.5.

3.4 Формалне онтологије

Онтологија је вид репрезентације знања. Уколико је онтологија дата на неком формалном језику и складишти се у формату који је рачунарски читљив реч је о формалној онтологији. Позната и општеприхваћена Груберова[82] (Gruber, 1993) дефиниција онтологије као „формалне спецификације дељене концептуализације“ указује на онај вид знања које је могуће пренети, размењити и употребити. Према Девеџићу (Devedžić, 2006), главна сврха онтологија у информатици није да буду коришћене као речници или таксономије, већ да учествују у дељењу и вишеструкој употреби знања од стране различитих интелигентних агената и апликација. Зависно од тога који део стварности описују, онтологије могу бити:

- онтологије највишег нивоа (енг. top level ontologies) - када моделују опште концепте, а знање које репрезентују је свеобухватно, систематизовано и применљиво у великом броју апликација. Примери

[82] Thomas R. Gruber

ове класе онтологија су: *Cyc,* GOLD, DOLCE, SUMO и др. Неке од најважнијих су детаљније представљене у одељку 3.4.2;

- доменске онтологије (енг. domain ontologies) – када се знања која репрезентују тичу једног домена или класе проблема. Један од значајних примера доменских онтологија су оне које се баве концептима и релацијама домена простор-време[83]. Једна од таквих доменских онтологија развија се на Универзитету Бремен (Bateman & Farrar, 2004). Због комплексности домена, упоредо се ради и на развоју специјализованог формалног језика *Geographical Markup Language* (GML)[84], кроз активности конзорцијума за геолокацијска истраживања и стандарде *Open Geospatial Consortium* (OGC)[85]. Доменске онтологије у области лингвистике могу бити различите намене и структуре. У радовима (Harris & Di Marco, 2009; Kelly et al., 2010) представљени су принципи развоја доменских онтологија реторичких фигура за енглески језик. Један пример ове класе, *RetFig* – дескриптивна онтологија реторичких фигура српског језика, развијен у току рада на овој тези (Mladenović & Mitrović, 2013), биће приказан у поглављу 6;
- онтологије задатака (енг. task ontologies)[86] (Chandrasekaran, Josephson & Benjamins, 1998) или апликацијске онтологије (енг. application ontologies) – садрже само она знања неопходна за извршавање дате класе задатака. Примери ове класе су CBT (Computer Based Training task ontology) онтологија задатака развијена као мета-модел система који помаже у развоју модела за обуку уз помоћ рачунара (Ikeda, Seta & Mizoguchi, 1997) и онтологије задатака система SWBE (Semantic Web-Based Education Systems) (Devedžić, 2006). У овој тези биће речи о *SemRetFig* онтологији задатака једне класе семантичких реторичких фигура српског језика (поглавље 6).

[83] http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/
[84] http://www.opengeospatial.org/standards/gml
[85] http://www.opengeospatial.org
[86] http://ksi.cpsc.ucalgary.ca/KAW/KAW98/chandra/index.html

Онтологије имају структуру која не зависи од њихове сложености, а њу чине:

- концепти или класе (енг. concepts, classes)
- инстанце класа или индивидуе (енг. instances, individuals)
- релације између класа (енг. relations, properties)
- атрибути (енг. attributes, slots)
- формална правила (енг. axioms)

Концепти генеришу таксономију ступајући међусобно у релације. Индивидуе представљају инстанце, односно конкретне примерке класа. Релације могу постојати између класа, између индивидуа и између индивидуа и класа. И концепти и индивидуе могу се описати атрибутима. Формалним правилима се исказују знања која нису дата експлицитно, односно релације међу концептима које се могу извести на основу знања о стварности коју описују. На пример, уколико постоје концепти (класе) *A, B* и *C* такви да међу њима важе и две релације

$$sestra(A,B) \quad \text{и} \quad roditelj(B,C)$$

тада се може дефинисати формално правило[87] које описује семантичке односе у стварном свету, а тиче се постојања концепта тетке, помоћу кога се закључује:

$$sestra(?\,a,?\,b), roditelj(?\,b,?\,c) \rightarrow tetka(?\,a,?\,c)$$

где су a, b и c инстанце одговарајућих класа.

Онтолошка знања описују се формалном логиком. То може бити дескриптивна логика на којој се заснива изражајност семантичког веба и која представља подскуп предикатске логике првог реда, али то може бити и формална логика првог реда (нпр. код онтологија SUMO и *Cyc*). Препорука W3C[88] за формални стандардни језик представљања онтологија на семантичком вебу је *Web Ontology Language* (OWL). OWL језик се може користити у три облика које разликује ниво изражајности језика. Најчешће

[87] Правило је задато у језику SWRL (в. одељак 6.3)
[88] http://www.w3.org/

се користи OWL DL (*Description Logic*) који, с једне стране, обезбеђује максималну изражајност, а са друге комплетност и одлучивост. Под комплетношћу се подразумева да се све релације могу разрешити у коначном времену, а под одлучивошћу да се у коначном времену може закључити да ли је онтологија конзистентна (нема контрадикција) и да ли је свако правило логичка последица других правила. Постоји више језика којима се могу представити онтолошка правила: *RuleML, Semantic Web Rule Language* (SWRL), *Knowledge Interchange Format* (KIF), *CycL* и др. Алати који користе онтолошка правила за извођење нових знања зову се расуђивачи (енг. reasoners), а најчешће се користе: *FACT*[89], *Pellet*[90], *Jena*[91], *RaserPro*[92], *Sesame*[93], *Euler*[94] и др.

3.4.1 Ворднет као онтологија

Ворднет, због своје свеобухватне и разгранате таксономије, успешно се користи у разним задацима као што су: оцена семантичке сличности, лексичко-семантичка анотација, класификација и сумаризација текстова, анализа осећања, аутоматско генерисање система за одговарање на питања и др. Међутим, иако је знање представљено ворднетом формализовано, он се не може применити у системима који могу расуђивати, односно изводити чињенице које нису експлицитно дате. Због тога је велики број истраживања био усмерен ка проналажењу метода и система аутоматске и једнозначне трансформације мрежне структуре ворднета у структуру формалне онтологије.

Један од првих таквих пројеката објавио је Брикли[95] који је предложио аутоматско генерисање RDFS структуре у којој учествују само именице ворднета и релација хипонимије (Brickley, 1999). Затим су Декер и Мелник[96]

[89] http://www.cs.man.ac.uk/~horrocks/FaCT/
[90] https://github.com/complexible/pellet
[91] https://jena.apache.org/documentation/javadoc/jena/com/hp/hpl/jena/reasoner/Reasoner.html
[92] http://franz.com/agraph/racer/
[93] http://rdf4j.org/
[94] https://www.w3.org/2001/sw/wiki/Euler
[95] Dan Brickley
[96] Stefan Decker and Sergey Melnik

дали RDF репрезентацију ворднета која садржи све синсетове, али ограничен скуп релација. Синонимске леме једног синсета биле су третиране као лабеле инстанце синсета, што значи да се значење речи и синонимска лема у оваквој онтологији не могу адресирати, јер они немају свој URI (Melnik & Decker, 2001). Свеобухватнија метода конверзије, која је касније била полазна тачка за истраживања Хјуа[97] и његових сарадника (Hu, Du, Liu & Ouyang, 2006), Хуанга и Зоуа[98] (Huang & Zhou, 2007), Гравеса и Гутиереза[99] (Graves & Gutierrez, 2006), предложио је ван Асем са сарадницима 2004. Овај свеобухватни алгоритам аутоматске конверзије неког тезауруса дат је у раду ван Асема[100] и његових сарадника (van Assem, Menken, Schreiber, Wielemaker & Wielinga, 2004), а примењен је и оцењен на примеру Принстонског ворднета. Алгоритам садржи четири корака: припремну фазу, синтаксну конверзију, семантичку конверзију и стандардизацију.

Ван Асем је са сарадницима две године касније представио унапређени алгоритам конверзије у раду (van Assem, Gangemi & Schreiber, 2006) који је постао званична препорука W3C за конверзију Принстонског ворднета у RDF/OWL формат, у коме се значења речи и све синонимске леме могу адресирати, а у моделу учествују и све ворднетом дефинисане лексичко-семантичке релације.

На комплексност целокупног поступка трансформације ворднета, као семантичке мреже, у формалну онтологију указали су у раду (Gangemi, Guarino, Masolo, Oltramari & Schneider, 2003) Гангеми и Гуарино[101] са сарадницима, трагајући за конзистентним системом транслације релације хипонимије/хипернимије и синсетова који су њоме повезани. У том смислу, они су размотрили неколико структуралних проблема са којима се ворднет сусреће када је у питању формална репрезентација знања и дали предлоге за њихово превазилажење. Пре свега, у структури ворднета, постоји конфузија у случајевима када су два синсета, од којих један има семантику концепта, а

[97] H.Hu
[98] Huang Xiao-xi, Zhou Chang-le
[99] Alvaro Graves, Claudio Gutierrez
[100] Mark van Assem
[101] Aldo Gangemi, Nicola Guarino

други преставља инстанцу (индивидуу) концепта, у релацији *is-a (*релација хипонимије*)* са истим синсетом. Аутори наводе пример синсета *composer* (композитор) који представља надређени концепт синсетовима *songwriter* (текстописац), који се сматра концептом, и синсетовима *Bach* (Бах) и *Beethoven* (Бетовен) који се сматрају инстанцама (индивидуама) концепта композитор. Аутори предлажу превазилажење ове врсте проблема увођењем релације INSTANCE-OF и преуређењем хијерархијске структуре релације хипонимије/хипернимије како би се јасно одвојиле везе концепт-концепт од веза концепт-индивидуа. Друга врста проблема последица је постојања синсетова, од којих један има семантику концепта, а други има семантику мета-податка који се налазе у релацији *is-a* (релација хипонимије) са истим синсетом. У том случају, у истој равни, односно истом нивоу апстракције у односу на надређени концепт стоје податак и метаподатак. На пример, концепт *Person* (особа) има два надређена концепта: концепт *Organism* (организам) који се интерпретира као тип и концепт *Causal_Agent* који представља улогу, па не би требало да буде надређен концепту *Person* (особа) који се интерпретира као тип. Трећа врста проблема се односи на различите нивое генерализације синсетова који се налазе на истом нивоу репрезентације у хијерархији семантичког подстабла неког надређеног концепта. На пример, у Српском ворднету, синсетови *дивљач* и *ларва* су у *is-a* релацији са синсетом *животиња*, мада је јасно да је ларва само један стадијум у развоју животиње. У оба случаја аутори предлажу да се при додавању новог концепта провере све постојеће хијерархијски организоване релације, не само према родитељима, већ и према концептима који су истог или нижег семантичког нивоа. Остали проблеми који се тичу проналажења полисемије и неадекватне позиције концепата у хијерархијској структури *is-a* релације такође се могу решавати реорганизацијом таксономије ворднета.

Пример такве реорганизације је пресликавање основне таксономије (*backbone taxonomy*) ворднета у онтологију DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*) (Gangemi, Guarino, Masolo, Oltramari & Schneider, 2003; Gangemi, Guarino, Masolo & Oltramari, 2003; Gangemi, Guarino, Masolo & Oltramari, 2010).

3.4.2 Формалне лингвистичке онтологије

Развој онтологије је комплексан и временски захтеван посао. Идеја о аутоматизацији тог посла тј. процесу „обуке онтологије“ (енг. ontology learning) подразумева развој техника и алата за екстракцију, анотацију и интеграцију информација у постојећу структуру једне онтологије (Devedžić, 2006). Знање које поседује једна онтологија може се генерисати и допуњавати на више различитих начина. Када је реч о лингвистичким знањима, процес учења онтологије може се интегрисати са процесима над различитим језичким ресурсима. Према (Prévot, Borgo & Oltramari, 2005), та сарадња може унапредити обе стране у процесу или може довести до генерисања квалитативно бољег ресурса у сврху примене у сложеним задацима обраде природног језика. Прево и сарадници истичу три начина сарадње онтологије и језичких ресурса:

1. реструктуирање језичког ресурса на основу принципа вођених знањем онтологије
2. увећање знања онтологије лексичким информацијама
3. поравнање језичког и онтолошког ресурса на жељеном нивоу.

Пример првог вида сарадње је реорганизација основне таксономије Принстонског ворднета пресликавањем у онтологију DOLCE (в. претходни одељак). Када је реч о увођењу лексичких информација и увећању лексичких знања једне онтологије, Баутелар[102] (Buitelaar, 2005) предлаже изградњу модела лексикона *LingInfo*[103] који би био намењен допуњавању онтологија информацијама о: врсти речи, морфолошкој и синтаксној декомпозицији речи, граматичким (граматичка правила) и контекстним моделима (контекстна репрезентација у *n*-грамским моделима) и др. За дати лексикон се најпре дефинише метакласа *ClassWithLingInfo* и метарелација *PropertyWithLingInfo*, а затим се дефинишу инстанце класе *LingInfo* која има релације *term*, *lang* и *morphoSyntacticDecomposition.* Релација

[102] Paul Buitelaar
[103] http://olp.dfki.de/LingInfo

morphoSyntacticDecomposition повезује инстанце класе *LingInfo* са инстанцама класе *PhraseOrWordForm.* Сама класа *PhraseOrWordForm* садржи подкласе: *WordForm* (дефинише лему и морфолошку структуру речи помоћу релација *DataProperties* које означавају род, број , падеж и сл.) , *stem* (дефинише корен речи), *Phrase* (изрази у којима се појављује реч), *InflectedWordForm* (дефинишу се правила генерисања флективних облика). Речник се генерише повезивањем са семантичком мрежом ворднет и селекцијом доменски одређеног скупа синсетова чији литерали постају речничке одреднице.

Најзад, интеграцијом онтологија и лексичко-семантичких мрежа могу настати и нови ресурси. Гангеми и његови сарадници (Gangemi, Guarino, Masolo & Oltramari, 2005) развили су *OntoWordNet*[104] ресурс који представља интеграцију Принстонског ворднета и онтологије DOLCE. DOLCE[105] припада класи онтологија највишег нивоа јер се њоме моделује опште људско знање. Развијена је у склопу пројекта *WonderWeb*[106] као део *Foundational Ontologies Library* (WFOL). Према ауторима (Gangemi, Guarino, Masolo, Oltramari & Schneider, 2003; Gangemi, Guarino, Masolo & Oltramari, 2010) намена овог ресурса није да он буде универзална стандардна онтологија, већ да помогне у откривању скривених лингвистичких веза и значења у другим врстама лингвистичких ресурса као што су семантичке мреже попут ворднета. Основне претпоставке изградње ове онтологије су да постоји експлицитна разлика између општег (universals) и посебног (particulars), између објеката који су просторно-временски константни (endurant) као што су то бића и предмети, објеката који су просторно-временски привремени и повремени (perdurant) као што су то догађаји и стања, као и објеката који представљају апстрактне концепте (Gangemi, Guarino, Masolo & Oltramari, 2003). Генерисањем *OntoWordNet-a* добијен је свеобухватни семантички ресурс са могућношћу расуђивања.

Други значајан пример интеграције Принстонског ворднета са онтологијом, у циљу генерисања ресурса са могућношћу резоновања на

[104] http://www.loa.istc.cnr.it/old/ontologies/OWN/OWN.owl
[105] http://www.loa.istc.cnr.it/old/DOLCE.html
[106] http://cordis.europa.eu/result/rcn/41438_en.html

принципима формалне логике првог реда, предложен је у виду поравнања SUMO онтологије и PWN-а на нивоу концепта. *Suggested Upper Merged Ontology* (SUMO) је онтологија највишег нивоа изражена помоћу SUO-KIF, варијанте KIF језика. Развијена је 2000. захваљујући Пизу[107] (Pease, 2011) као прва формална онтологија поравната са PWN. SUMO садржи релативно мали број концепата (око 1000), али велики број формалних тврђења (око 4000) и правила (око 800) што омогућава да се овај ресурс користи у системима расуђивања, проналажења информација и другим задацима обраде природног језика. Важна компонента у развоју овог ресурса је *Sigma Knowledge Engineering*[108], радно окружење које омогућава да се корисничке онтологије интегришу са SUMO као и да користе систем за аутоматско доказивање теорема *Vampire*. У овом развојном окружењу корисник може креирати и сопствени систем тврђења и правила и интегрисати га у постојећи систем расуђивања заснован на формалној логици првог реда.

На достигнућима онтологије SUMO, на Универзитету Аризона, 2000. је креирана прва доменска онтологија која описује лингвистичка знања - *Electronic Metastructure for Endangered Languages Data* (E-MELD). Она садржи формализоване описе лингвистичких метаподатака о појмовима као што су: време, простор, људски односи, функције и др. Концепти које описује ова онтологија односе се на знање о језику уопште - граматику, дијалект, говорно и писано изражавање језика. Унапређењем E-MELD онтологије настала је свеобухватнија лингвистичка онтологија - *General Ontology for Linguistic Description* (GOLD). Ова онтологија је првенствено намењена апликацијама које укључују дескриптивну лингвистику. Према (Farrar & Langendoen, 2003), GOLD је креирана како би омогућила аутоматско расуђивање о разним граматичким феноменима и теоријама. Заснована је на принципима онтолошког инжењеринга, односно скупа принципа, процеса, активности и системских методологија који омогућавају развој и примену онтологије током њеног природног циклуса: пројектовања, имплементације,

[107] Adam Pease
[108] http://sigmakee.sourceforge.net/

евалуације, валидације, одржавања, дистрибуције, пресликавања, интеграције, дељења и поновног коришћења (Gašević, Djurić & Devedžić, 2006). Сви концепти организовани су у четири главне категорије: изрази, граматика, конструкције података и метаконцепти. Два основна циља којима тежи су: опис свих граматичких концепата и њихових инстанци (облика појављивања) и могућност анализе значења израза у природним језицима у чему јој помаже веза са SUMO онтологијом. GOLD обезбеђује и прилагођавање функција језичким профилима што чини да буде применљива и на другим језицима осим енглеског. Спада у групу јавно доступних лингвистичких ресурса[109], а постоји и алат за онлајн претрагу[110].

3.5 Језички алати

Језички алати су софтверски производи који обезбеђују и олакшавају развој и одржавање језичких ресурса са једне стране, а са друге њихову међусобну интеграцију ради ефикасније примене у задацима обраде природног језика (Krstev, Stanković, Vitas & Obradović, 2006; Obradović & Stanković, 2008).

Један од основних језичких алата имплементира методе којима се све лексичке јединице посматраног текста препознају и означавају појединачно и назива се токенизација (енг. tokenization). Токеном се, у општем случају, сматра она језичка целина која је омеђена празнинама, односно бланко знацима. Ово правило важи за већину европских језика, али не важи, на пример, за језике источне Азије где се речи не одвајају бланко знацима. Одступања од основног правила дефинишу се изузецима који зависе и од језика и од конкретне примене. На пример, у одређеним случајевима токеном се сматрају и лексичке јединице које у себи садрже празнине, ако самостално немају смисла (приказ датума, разне врсте нумерација, именовани ентитети као што је Нови Сад и сл.). Токеном се могу сматрати и вишечлане лексичке јединице добијене повезивањем речи знаком цртице (нпр. данас-сутра, индо-

[109] http://linguistics-ontology.org/gold-2010.owl
[110] http://linguistics-ontology.org/gold/

европски, призренско-јужноморавски итд.) или другим знацима (нпр. број 23.456,78, новчани износ 234€ итд.).

Други важан језички алат имплементира методе којима се препознају и означавају границе реченица у тексту на природном језику (постоји више појмова којима се на енглеском назива овај алат – *sentencer*, *sentence splitter* и одговарајући процес означавања – *sentence segmentation*, *sentence boundary disambiguation*). Подела текста на реченице представља једну од основних процедура којима се неки текст припрема за дубљу језичку анализу. Сам алат може бити заснован на моделу добијеном неким алгоритмом машинског учења или на моделу који представља скуп правила. Без обзира на начин реализације, уобичајен начин означавања почетка и краја реченице је, рецимо, етикетама *<S>* и *</S>*.

Два, по својој природи слична, језичка алата односе се на примену морфолошких правила у циљу сажимања различитих морфолошких варијанти једне речи, било флективних или деривационих, на један јединствени облик. Један од тих алата је стемер, а други је лематизер. Иако представљају различите морфолошке алате, њихов је задатак да трансформацијом облика речи генеришу одговарајући полазни облик, али уз очување основног значења речи.

Кореновање или стеминг (енг. stemming), односно свођење речи на њен корен подразумева уклањање афикса (најчешће суфикса и неких префикса који не мењају значење речи) из посматране речи како би се добио њен корен. Корен свих флективних и деривационих облика насталих из једне леме је исти.

Аналогно процесу кореновања може се спровести и процес лематизације. Лематизација (енг. lemmatization) је поступак свођења речи на њен основни облик - лему. Лематизер, као језички алат, мора омогућити да се свака флективна класа једнозначно препозна, чиме се сви флективни облици једне леме своде увек на њу. За разлику од кореновања, резултат лематизације је увек лингвистички исправан облик. Лематизацијом се обухватају само флективни облици, а кореновањем и деривациони.

Означавање врстама речи или морфосинтаксно означавање (Part-of-speech - POS tagging) је процес придруживања граматичке категорије свакој речи у тексту, при чему је саставни део тог процеса и препознавање хомографије, односно разрешавање проблема вишезначности речи (нпр. у реченици *Пошто је вредно радио, купио је радио* реч *радио* у првом делу реченице представља глагол, а у другом именицу).

Језички алат који, у општем случају, анализира неку хијерархијску или сложену језичку структуру рашчлањивањем њених саставних делова назива се парсер (енг. parser[111]). У рачунарској лингвистици, парсирање је препознавање синтаксне структуре реченице на основу познатих правила којима се гради реченична структура једног природног језика. Иако се заснива на неким општим језичким правилима (правила постојања субјекта, објекта или радње која се врши), парсер као алат није језички независан, јер имплементира особине и правила оног природног језика коме је намењен.

Један од најсвеобухватнијих језичких алата за српски језик је *LeXimir*[112] – модуларно, интегрисано окружење за рад са језичким ресурсима на српском језику, развијено у склопу активности Групе за језичке технологије Универзитета у Београду (Krstev, Stanković, Vitas & Obradović, 2006). *LeXimir* интегрише семантичку мрежу Српски ворднет и морфолошке речнике српског језика и садржи низ функција за конверзију формата, конверзију кодних распореда, обраду и визуелизацију паралелизованих текстова и др. *LeXimir* је преузео и проширио функционалности апликације *VisDic* којом је развијан Српски ворднет од свог настанка. Међутим, *VisDic* је и поред својих добрих особина које подразумевају: истовремени приказ већег броја ворднетова, копирање синсетова из једног у други ворднет, графички приказ хијерархије семантичких релација, имао и низ недостатака:

- непостојање валидације XML документа, што може довести до неконзистентне структуре ворднета;
- немогућност проширења ворднета скупом нових ознака;

[111] од латинског израза „*pars orationis*“ што значи „врсте речи“
[112] Назив претходне верзије овог алата је Workstation for Lexical Resources – WS4LR

- немогућ вишекориснички рад, односно непостојање подршке за рад на мрежи;
- непостојање поступка конверзије и серијализације у друге формате: OWL, RDF, CSV и др.
- недовољна сарадња са SUMO и DOMAIN (не постоји могућност упита на основу SUMO или DOMAIN концепата) и другим проширењима ворднета (в. одељак 3.3.2).

Детаљна анализа функционалности *VisDic*-а и предлози унапређења алата за развој Српског ворднета дати су у раду (Mladenović, Mitrović & Krstev 2014).

Рад на унапређењу Српског ворднета и обезбеђењу што функционалнијег и свеобухватнијег софтверског алата за даљи развој и посебно примену ове семантичке мреже у задацима анализе осећања састојао се из неколико корака. Прво је било неопходно креирати алат *SWNUpgradeUtil* (прилог 3.8) за једнозначан и безбедан прелаз верзије SWN, компатабилне са PWN 2.0, на верзију компатабилну са PWN 3.0 за шта постоји јавно доступан ресурс који садржи функције пресликавања[113] синсетова међу верзијама PWN. Подизање верзије SWN омогућило је интеграцију са онтологијама и лексичким ресурсима (SUMO, *SentiWordNet*) који нису кореспондирали са старијим верзијама PWN. При том су решаване две врсте проблема: појава поделе синсета из верзије 2.0 на више различитих синсетова у верзији 3.0 и појава недостатка синсета у верзији 3.0 који би представљао резултат пресликавања синсета из верзије 2.0. У првом случају софтвер је нотирао 26 таквих синсетова, а у другом 147. Оба проблема решавана су ручно. Синсетови са ознакама BILI и SRP који представљају специфичности балканских језика и српског језика, остали су исти. *SWNUpgradeUtil* је омогућио и превођење *Аурора* записа (Vitas, 1980) ASCII кодне стране на *Unicode*.

Други корак даљег развоја односио се на креирање и имплементацију XSD схеме XML документа који представља SWN. Током рада са *VisDic*

[113] http://www.talp.upc.edu/index.php/technology/resources/multilingual-lexicons-and-machine-translation-resources/multilingual-lexicons/98-wordnet-mappings?highlight=WyJtYXBwaW5nIl0=

апликацијом било је могуће обезбедити добро формиран документ код кога се постизала јединственост *ID* вредности синсетова као и литерала. Међутим, појаве недозвољених вредности за *BCS* и *LNOTE* етикете, могућност различитог редоследа појављивања ознака у различитим синсетовима, вишеструког броја појављивања истих ознака унутар истог синсета (нпр. POS ознаке), недозвољеног садржаја или неадекватно означеног садржаја појединих ознака, указивали су на неопходност увођења валидације SWN XML документа у циљу отклањања постојећих недоследности и спречавања њиховог поновног појављивања у даљем развоју овог ресурса. У току имплементације овог корака креирана је веб апликација са функционалношћу валидатора XML документа (*SWNvalidateUtil*) (прилог 3.9) у односу на дату XSD схему која омогућава тимски рад на проналажењу и отклањању проблема ове врсте насталих у периоду пре укључења обавезне валидације ресурса. Сама XSD схема креирана је на основу првобитне верзије схеме дефинисане за SWN XML структуру (Krstev, Vitas, Stanković, Obradović & Pavlović-Lažetić, 2004).[114]

У трећем кораку, SWN ресурс допуњен је SUMO ознакама (Mladenović & Mitrović, 2014). Рад на овом проширењу условљен је потребама истраживача да се SWN употреби као ресурс од кога ће се поћи у креирању базе знања у области нутриционизма као и у области истраживања фигуративног језика (реторичких фигура) - о чему ће детаљно бити речи у поглављу 6. Проширење су креирали и објавили Нилс и Пиз[115] (Niles & Pease, 2003), а састоји се у дефинисању једнозначног пресликавања[116] свих синсетова PWN 3.0 у концепте (класе) SUMO онтологије. Приликом пресликавања SWN у концепте SUMO онтологије, на основу датих пресликавања PWN у SUMO, остало је непресликано 876 синсетова. Од тога 706 синсетова се односило на BILI и SRP синсетове, а осталих 161 на синсетове добијене на основу PWN 3.0. Овај проблем решен је имплементацијом логике која за сваки неповезани синсет испитује постојање везе ка SUMO концептима оних синсетова који су

[114] в. одељак 3.3.3
[115] Ian Niles and Adam Pease
[116] http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/

са неповезаним синсетом у релацији која зависи од његове POS ознаке: *hypernym* за именице, *derivative* и *hypernym* за глаголе, *derived* и *particle* за придеве, *derived* за прилоге. Уколико веза ових синсетова ка SUMO концептима постоји, онда се иста веза аутоматски генерише и за синсет који је неповезан са SUMO. Након тога, преостали неповезани синсетови пресликани су ручно. Неповезаност је могла бити последица постојања циркуларне релације холонимије двају значења (нпр. синсет са идентификатором ENG30-01605119-n којим се дефинише концепт *porodica Accipitridae*), последица вишеструких веза хиперонимије, последица деривације због чега се није могао једнозначно дефинисати SUMO концепт за пресликавање. У примеру глагола *кретати се* синсет са ознаком ENG30-00014549-v

```
<SYNSET><ID>ENG30-00014549-v</ID><POS>v</POS>
<SYNONYM><LITERAL>kretati
se<SENSE>2x</SENSE><LNOTE>V524+Imperf+It+Iref+Ref</LNOTE></LITERAL>
</SYNONYM><DEF>Biti u stanju akcije.</DEF>
...
<ILR>ENG30-00165942-n<TYPE>eng_derivative</TYPE></ILR>
<ILR>ENG30-14004317-n<TYPE>eng_derivative</TYPE></ILR>
...
<SUMO>BodyMotion<TYPE>=</TYPE></SUMO>
...
</SYNSET>
```

садржи две деривационе везе које указују на синсетове повезане са различитим SUMO концептима (*Deciding* и *Motion*), како је то приказано у наставку.

```
 <SYNSET><ID>ENG30-00165942-n</ID><POS>n</POS>
<SYNONYM><LITERAL>potez<SENSE>6</SENSE><LNOTE>N1</LNOTE></LITERAL></SYNON
YM><DEF>Čin odluke da se uradi nešto.</DEF>
...
<ILR>ENG30-00162632-n<TYPE>hypernym</TYPE></ILR>
<ILR>ENG30-00014549-v<TYPE>eng_derivative</TYPE></ILR>
<ILR>ENG30-02367363-v<TYPE>eng_derivative</TYPE></ILR>
<ILR>ENG30-00168237-n<TYPE>hyponym</TYPE></ILR>
...
<SUMO>Deciding<TYPE>=</TYPE></SUMO>...</SYNSET>
---------------------------------------------------------------------------------------------------------------------------
<SYNSET><ID>ENG30-14004317-n</ID><POS>n</POS>
<SYNONYM><LITERAL>kretanje<SENSE>1</SENSE><LNOTE>N300+VN</LNOTE></LITERAL>
<LITERAL>kretnja<SENSE>2</SENSE><LNOTE>N600</LNOTE></LITERAL></SYNONYM>
<DEF>stanje promene</DEF>
...
<ILR>ENG30-00024720-n<TYPE>hypernym</TYPE></ILR>
<ILR>ENG30-00014549-v<TYPE>eng_derivative</TYPE></ILR>
...
<SUMO>Motion<TYPE>=</TYPE></SUMO>...</SYNSET>.
```

У четвртом кораку SWN је проширен тако што је одређеним синсетовима додат елемент којим се дефинише поларитет и интензитет осећања синсета, како би Српски ворднет могао бити коришћен у задацима анализе осећања. Иако постоји више специјализованих лексикона сентименталних речи и израза као и додатака Принстонском ворднету којима са генеришу информације о поларизацији и интензитету осећања представљених синсетовима, у раду (Mladenović & Mitrović, 2014) описано је проширење SWN употребом додатака *SentiWordNet* чији су аутори Есули и Себастиани (Baccianella, Esuli & Sebastiani, 2010). *SentiWordNet* дефинише јачину негативног и позитивног поларитета осећања синсета и исказује је децималним бројем из опсега [0,1]. Аутори су генерисали и јавно доступни ресурс[117] који је за сваки синсет дефинисао ове вредности. Структура тог ресурса дата је на примеру синсета чији је идентификатор 01461292 и односи се на придев *љубазан* (енг. likeable):

```
a 01461292 0.75 0.25 likeable#2 likable#2 easy to like; agreeable; "an
attractive and likable young man"
```

где су: вредности 0.75 и 0.25 јачине позитивног и негативног поларитета осећања синсета који је означен ID вредношћу 01461292 и представља придев

[117] http://sentiwordnet.isti.cnr.it/download.php

a са значењем #2. Након проширења SWN ресурса, структура синсета којим се описује концепт придева *љубазан* (енг. likeable) дата је са:

```
<SRPWN>
<SYNSET>
 <ID>ENG30-01461292-a</ID>
 <POS>a</POS>
 <SYNONYM>
   <LITERAL>ljubazan<SENSE>2</SENSE><LNOTE /></LITERAL>
   <LITERAL>prijatan<SENSE>2</SENSE><LNOTE /></LITERAL>
 </SYNONYM>
 <DEF>koji je dopadljiv; odgovarajući</DEF>
 <NL>yes</NL>
 <STAMP>jeca 18/09/2012 00:00:00</STAMP>
 <SUMO>SubjectiveAssessmentAttribute<TYPE>+</TYPE></SUMO>
 <SENTIMENT>
   <POSITIVE>0.75000</POSITIVE>
   <NEGATIVE>0.25000</NEGATIVE>
 </SENTIMENT>
 <DOMAIN>factotum</DOMAIN>
</SYNSET>
</SRPWN>.
```

У петом кораку креирана је веб апликација SWNE (Serbian WordNet Editor)[118] за даљи групни рад на развоју и употреби SWN ресурса. Сврха апликације (Mladenović & Mitrović, 2014; Mladenović, Mitrović & Krstev, 2014) је да обједини добре стране досадашњег алата, нове захтеве корисника српске семантичке мреже и савремене технике развоја софтвера како би се омогућио безбедан, функционалан, вишекориснички, модуларан и лако проширљив систем развоја семантичких ресурса на српском језику са увидом у динамику доградње ефикасним системом извештавања и свеобухватним системом претраге. Безбедност апликације одржава се помоћу улога (*roles*) и нивоа (*levels*) унутар датих улога. Дефинисане улоге су: неауторизовани корисници, ворднет корисници, администратори и корисници осталих ресурса (*SimNet*, *RetFig*, *GradAdj* и др.). Неауторизованим корисницима је дозвољен елементарни упит над мрежом, постављање сложених логичких филтера и статистичко извештавање о везама и значењима унутар саме мреже. Улогама се обезбеђује приступ појединим ресурсима, а у оквиру сваке улоге се дефинишу нивои - обични корисници могу уносити и мењати само

[118] http://sm.jerteh.rs/

записе које су они дефинисали и модератори који имају контролу над комплетним ресурсом коме приступају на основу улоге. Улогом *WordNetMember* кориснику се дају привилегије уноса нових синсетова и измене само оних синсетова које је корисник унео. Виши ниво унутар ове улоге је *Moderator* и он омогућава измену свих синсетова без обзира на ауторство. Улогом *RetFig* приступа се ресурсу реторичких фигура, а улогом *Adjectives* осталим семантичким ресурсима који се развијају у склопу ове апликације (*SimNet*, *GradAdj* и *Antonyms*). Апликација даје могућност рада са жељеном XSD схемом и SWN XML документом. Уграђена је валидација и аутоматска корекција документа. Задржана је синхронизација са претрагом PWN и могућност преписивања PWN синсетова у SWN коју је нудио *VisDic*. Међутим, апликација унапређује систем претраге имплементацијом модула упита над PWN ресурсом. У ту сврху коришћен је јавно доступан модул *WordNetEngine*[119] (Gerber, 2013) који је био основа за пројектовање и имплементирање одговарајућег модула *SWNEngine,* након чега су ова два модула интегрисана могућношћу копирања одабраног синсета из PWN у Српски ворднет. Апликација омогућава једноставну измену и допуну SWN синсетова, контролу вишеструког уноса описа, ID-а, литерала и значења, олакшава избор морфолошких, SUMO и *SentiWordNet* ознака, нуди свеобухватну претрагу, добар хијерархијски приказ резултата претраге и серијализацију у формате TXT, CSV, RDF и XML који се постиже одабраном XSL трансформацијом (XSLT) комплетног ресурса или делова добијених на основу задатог услова претраге.

Апликација тренутно садржи пет функционалних целина:

- онлајн едитор Српског ворднета (прилог 3.10),
- сегмент за статистичка извештавања о SWN XML (прилог 3.11),
- модул за претрагу SWN XML документа аутоматским генерисањем XPATH упита (прилог 3.12),

[119]https://code.google.com/p/opensim4opencog/source/browse/trunk/lib/LAIR.ResourceAPIs/?r=1607#LAIR.ResourceAPIs%2FWordNet%2FWordNet

- јединицу за формирање сложених логичких XPATH упита над SWN XML ресурсом (прилог 3.13),
- модул за паралелну претрагу SWN и PWN на основу литерала, дефиниције, примера употребе или домена. (прилог 3.14).

4. Методе класификације текста на основу осећања

Анализа људских осећања, ставова, мишљења, процена вредности и препорука исказаних на разне начине (покретом, говором, изразом лица, писаним путем или њиховим комбинацијама) може постати предмет анализе осећања онога тренутка када се информација о тако исказаним осећањима дигитализује. Иако предмет анализе осећања могу бити слика (дигитализовани израз лица) (Poria, Cambria, Hussain & Hunag, 2015) и говор (дигитализовани говор) (Poria, Cambria, Hussain & Hunag, 2015), предмет интересовања ове тезе је анализа осећања исказаних писаним путем, односно текстом. Извор текстова ове врсте анализа обично су садржаји: друштвених мрежа попут Фејсбука (Facebook) (Ortigosa, Martín & Carro, 2014) и Твитера (Twitter) (Pak & Paroubek, 2010; Perera, Anand, Subbalakshmi & Chandramouli, 2010), блогова (Yano & Smith, 2010), електронске поште (Mohammad & Yang, 2011), дискусионих група, система за оцену квалитета производа (Hagenau, Liebmann & Neumann, 2013), система за коментарисање вести (Hu, Bose, Koh & Liu, 2012), али то могу бити и литерарни текстови (Mohammad, 2011). У уводном поглављу истакли смо да је анализа на основу осећања по својој природи класификациони процес. Зависно од структуре текста који је предмет класификације, можемо говорити о три нивоа анализе осећања: анализи на нивоу целог документа, анализи на нивоу реченице и анализи на нивоу атрибута неког сложеног концепта. Сам класификациони процес предмете класификације може делити на позитивно, негативно и неутрално оријентисане на основу осећања, а може се увести и скала градације. Резултат сваке анализе осећања је генерисање семантичког знања о поларитету осећања и евентуално о степену тог поларитета посматраног предмета класификације. Када је реч о методама које се примењују, оне се заснивају на семантичким правилима помоћу којих се групишу или издвајају чланови класа, на статистичким методама којима се моделују особине и границе класа или на хибридним методама које комбинују обе врсте метода. У наредним одељцима овог поглавља биће приказане актуелне методе и алгоритми који се примењују у анализи текстова на основу осећања, а

упоредне карактеристике и категоризације алгоритама, метода и апликација анализе осећања могу се наћи и у свеобухватним прегледним радовима (ChandraKala & Sindhu, 2012; Hajmohammadi, Ibrahim & Ali Othman, 2012; Seerat & Azam, 2012; Buche, Chandak & Zadgaonkar, 2013; Medhat, Hassan & Korashy, 2014).

4.1 Нивои класификације текста на основу осећања

Ниво класификације у анализи осећања, пре свега, зависи од структуре текста који се анализира и циљева класификације. Уколико се ради о класификацији садржаја дужих форми какви су постови на блогу, прикази филмова, анализе на политичким профилима или су у питању литерарни текстови, онда се анализа углавном ради на нивоу докумената, а резултати класификације дају поларизацију свеукупних осећања посматраног документа. У случају обраде текстова кратких форми (микро-блогинг поруке) или анализе електронске поште (утврђивање постојања спама), класификација је резултат анализе осећања исказаних у појединачним реченицама. У случајевима анализе тржишта употребом система за препоруке или дискусионих група специјализованих за оцењивање квалитета одређених класа производа, користе се методе за оцену одређених, унапред дефинисаних атрибута посматраних производа, а затим њихову сумаризацију, на основу чега се изводе опште оцене о испитиваним производима.

4.1.1 Методе класификације на нивоу документа

У анализи осећања на нивоу докумената уводи се апстракција којом се идентификује један извор осећања и он је усмерен ка једном објекту. Заједничка карактеристика свих метода анализе осећања на нивоу документа је избор релевантног скупа предиктора којима се може описати сваки појединачни документ, а такође и избор релевантног алгоритма за оцену поларитета осећања тако одабраног скупа предиктора. Постоји велики број истраживања у овој области која експериментишу са различитим

структурама предиктора, различитим методама класификације и различитим алгоритмима за оцену предиктора којима се врши предикција поларитета исказаних осећања, а овде ћемо приказати само неке од значајнијих.

Једну од првих метода ове врсте увео је Турни[120] 2002. у раду (Turney, 2002). Као предмет класификације коришћени су текстови који представљају четири врсте приказа и оцена: аутомобила, филмова, банака и туристичких дестинација са сајта за онлајн оцењивање Епинионс[121], а тестови за оцену класификатора на основу осећања вршени су за сваку врсту понаособ. Класификатор који је Турни предложио заснивао се на анализи придева и прилога у тексту. Посебним алгоритмом из класе PMI (*Pointwise Mutual Information*) прво је утврђена семантичка оријентација (SO) (в. одељак 1.4) придева и прилога. Овај алгоритам мери вероватноћу да се посматрани придев, прилог или фраза у којој се неки од њих налази, појави у непосредној близини речи „excellent" у односу на вероватноћу да се нађе у непосредној близини речи „poor" у документима који су добијени као резултат машине за претрагу Алтависта[122].

$$\mathrm{SO}(\text{phrase}) = \log_2\left(\frac{\text{hits(phrase NEAR "excellent") hits("poor")}}{\text{hits(phrase NEAR "poor") hits("excellent")}}\right)$$

Као NEAR оператор коришћен је „AltaVista NEAR" оператор који је претрагу ограничавао само на оне документе који садрже дату фразу и једну од речи „*excellent*" и „*poor*" на међусобној удаљености од највише 10 речи. Уз претпоставку да придеви и прилози чија је семантичка оријентација таква да су „ближи" речи „*excellent*" носе позитиван поларитет осећања, а они који су „ближи" речи „*poor*" негативан, резултујући поларитет осећања у тексту одређиван је на основу просечне семантичке оријентације поларитета свих таквих речи и фраза.

[120] Peter Turney
[121] http://www.epinions.com
[122] http://en.wikipedia.org/wiki/AltaVista

Други приступ анализи осећања на нивоу докумената дао је Панг[123] са сарадницима (Pang, Lee & Vaithyanathan, 2002) анализирајући текстове који представљају онлајн оцене филмова. Методе предложене у овом раду су методе машинског учења: Наивни Бајесов алгоритам (*NB*), метода максималне ентропије (*MaxEnt Classification*) и метода потпорних вектора (*SVM*). Сви документи се посматрају као скупови предиктора на основу којих се класификациони систем моделује тако да може да израчуна највероватнију класу припадности посматраног документа.

У раду (Nigam, Lafferty & McCallum, 1999), користе се метода максималне ентропије и Наивни Бајесов алгоритам, али над три различита скупа текстова: садржајем студентских преписки са универзитетског веб сајта, садржајем компанијских веб страна и садржајем *UseNet* дискусионих група.

Метода потпорних вектора у анализи осећања текстова на нивоу докумената примењена је у радовима (Pang & Lee, 2005) и (Mullen & Collier, 2004). Панг и сарадници уводе ову методу машинског учења у процес вишекласне класификације текста на основу осећања, на основу унапред задате скале јачине осећања којом се дефинише припадност датој класи. Коришћени су садржаји сајтова где се квалитет филма може оцењивати бројевима од 1 до 5 или навођењем броја звездица (такође од 1 до 5). У раду је показано да се јако мали број инстанци налази у рубним класама, па је тестирање вршено са могућношћу класификације у три или четири класе. Предиктори класа су дефинисани тако што је одабрано по n ($n \geq 20$) предиктора којима је описана свака класа. Нпр. речи "meaningless" и "disgusting" били су коришћени као предиктори класе 0, "pleasant" и "uneven" као предиктори класе 1, а у класи 2 су се нашли и предиктори као "straightforward" и "oscar". Мулен и Колиер[124] су такође користили методу потпорних вектора у бинарној класификацији текстова из IMDB[125] базе текстова о филмовима на „позитивне" и „негативне" експериментишући са различитим особинама предиктора.

[123] Bo Pang
[124] Collier
[125] http://www.imdb.com/

4.1.2 Методе класификације на нивоу реченице

Анализа осећања на нивоу реченице подразумева да се сваки текстуални документ претходно подели на реченице (коришћењем неког од алгоритама за ту намену).[126] Реченица постаје предмет класификације на основу осећања, а подразумевани ниво апстракције идентификује један извор осећања који је усмерен ка једном објекту. Класификатори на основу осећања могу бити или бинарни или вишекласни. Бинарни класификатори могу бити или идентификатори субјективности (енг. subjectivity identification) (Esuli & Sebastiani, 2006) који реченице деле на оне које носе емоционални садржај и оне које имају искључиво чињенични садржај, или класификатори на основу осећања (енг. sentiment classification) тј. идентификатори поларитета осећања (енг. polarity identification) (Turney & Littman, 2003) којима се реченице које носе емоционални садржај деле на оне са позитивним и оне са негативним емоционалним садржајем. У случају примене вишекласног класификатора, истовремено се врши идентификација субјективности и идентификација поларитета осећања, јер се реченице обично сврставају у три групе: „неутралне“, „позитивне“ и „негативне“. Сви алгоритми и методе који се користе у класификацији на нивоу докумената, могу бити примењени и у овом случају (Liu, 2012). Значај анализе осећања на нивоу реченице посматрамо у њеној комбинацији са анализом на нивоу докумената. Пре свега, утврђивањем емоционално неутралних реченица, могуће их је изоставити из даљег разматрања и тиме повећати ефикасност класификатора поларитета. С друге стране, уклањањем емоционално неутралних реченица смањује се шум у подацима приликом класификације осећања на нивоу докумената. Панг и Ли[127] (Pang & Lee, 2004) су на овај начин побољшали резултате класификације у односу на основне методе класификације докумената.

Истраживања у овој области ослањају се на употребу језичких ресурса, правила синтаксе језика, алата за лексичку анализу, а користе се и

[126] в.слику 4.5 у одељку 4.6
[127] Bo Pang and Lillian Lee

алгоритми за оцену семантичке сличности. Класификатори се моделују помоћу метода надгледаног и ненадгледаног учења.

Једно од првих истраживања ове врсте објавили су Хаџивасилоглу и Виби[128] (Hatzivassiloglou & Wiebe, 2000). Они су у свом раду показали да постоји статистички значајан утицај семантички оријентисаних и градабилних[129] придева (Dragićević, 2010) на класификаторе субјективности на нивоу реченица. Ју и Хаџивасилоглу[130] (Yu & Hatzivassiloglou, 2003) извели су истраживање у коме су користили методе надгледаног машинског учења у циљу класификације реченица према субјективности. Користили су методе: наивни Бајес, вишеструки наивни Бајес као и методу семантичке сличности, где су претпоставили да је произвољна реченица која носи субјективно значење усмерено ка једном објекту семантички сличнија оној реченици која такође носи субјективно значење усмерено ка том објекту него реченици која не носи субјективно значење. Мерењем семантичке сличности између реченица познатог субјективитета и осталих, могуће је обучити класификатор да препознаје субјективне реченице у односу на дату тему или објекат посматрања.

Многе методе класификације на нивоу реченице експериментишу са различитим врстама предиктора. Сем утицаја придева и градабилних придева (Hatzivassiloglou & Wiebe, 2000), може се мерити и семантички утицај речи у непосредној близини речи за које постоје претходна знања о њиховој субјективној оријентацији или поларитету осећања. Рилоф[131] је са сарадницама (Riloff, Wiebe & Wilson, 2003) предложила употребу образаца за препознавање и екстракцију именица које носе субјективност полазећи од претпоставке да је субјективност у корелацији са поларитетом придева у њиховој близини. На пример, за образац „*expressed* <dobj>“ , екстраховане су именице: *condolences, hope, grief, views, worries, recognition*, итд. Тако генерисан

[128] Vasileios Hatzivassiloglou and Janyce M.Wiebe
[129] Градабилност је семантичка особина речи која јој омогућава да учествује у поредбеним конструкцијама и да у сарадњи са модификаторима семантичког значења делује као модулатор (појачивач или ослабљивач) основног значења речи (Wiebe et al., 2004; Dragićević, 2010)
[130] Hong Yu and Vasileios Hatzivassiloglou
[131] Ellen Riloff

скуп образаца коришћен је за унапређење класификатора субјективности реченица. Ким и Хови[132] (Kim & Hovy, 2004) су у истраживање увели и NLP алате:[133] за препознавање именованих ентитета (енг. Named Entity Recognition - NER) и означавање врстама речи (енг. POS tagger) којима се проширује скуп атрибута у реченици за које се испитује и утврђују субјективност и поларитет осећања.

Методе анализе осећања на нивоу реченице посебно су актуелне у анализи микро-блогинг система попут Твитера и система кратких порука где величина комуникационе јединице у просеку одговара једној реченици.[134] Према истраживању из 2013.[135] (Krikorian, 2013) укупан број генерисаних Твитер порука по дану у августу 2013. достизао је 500 милиона твитова. Као специфичан и динамичан медијум, Твитер је посебно интересантан ресурс у анализи осећања. Скуп предиктора на основу којих се истражује његов утицај проширен је скуповима симбола познатих под називом „емотикони" и специјализованих симбола као што су хештагови и знакова интерпункције. Ефикасне методе и предиктори у анализи твитова могу се наћи у радовима (Go, Bhayani & Huang, 2009; Davidov, Tsur & Rappoport, 2010; Agarwal, Xie, Vovsha, Rambow & Passonneau, 2011; Batista & Ribeiro, 2013).

Најзад, на нивоу реченице може се истраживати постојање фигуративног говора. Улога реторичких фигура може значајно утицати на поларитет осећања реченице, па и дужих структура документа. Примери фигура које мењају поларитет осећања реченице или њеног дела су иронија и сарказам, а постоји и низ других чијим деловањем значење посматраног дела текста може бити измењено. Према (Liu, 2012), саркастичне реченице нису присутне често у оценама производа и услуга, али су веома честе у онлајн дискусионим групама и блоговима са политичким темама. Гонзалез-Ибанез[136] (González-Ibáñez, Muresan & Wacholder, 2011) истражује утицај сарказма на твитове. У раду (Davidov, Tsur & Rappoport, 2010), сем твитова, користе се и оцене

[132] Soo-Min Kim and Eduard Hovy
[133] Алати обраде природног језика (NLP-Natural Language Processing)
[134] Максимална дужина Твитер поруке је 140 знакова.
[135] http://www.internetlivestats.com/twitter-statistics/#trend
[136] Roberto González-Ibánez

производа са сајта Амазон[137], а Барбиери и Сагион[138] моделују систем за препознавање ироније у твитовима (Barbieri & Saggion, 2014). О утицају реторичких фигура на анализу осећања у тексту и моделима њиховог проналажења у текстовима на српском језику биће више речи у поглављу 6.

4.1.3 Методе класификације на нивоу атрибута

Класификација осећања на нивоу атрибута (Bunescu & Wayne, 2003; Hu & Liu, 2004b; Yi, Nasukawa, Popescu & Etzioni, 2005; Zhang & Zhu, 2013) представља поступак којим се оцењује или ентитет који има сложену структуру, па се посматра као сумарна оцена својих појединачних елемената, или, уколико узмемо у обзир дефиницију 4.3 из одељка 2.2, постоји више ентитета за које, такође, тражимо сумарну оцену. Ова врста класификације је синтеза поступка екстракције (атрибута), класификације (поларитета) и сумаризације (поларитета) позната и под називом "*feature-based opinion mining and summarization*" (Hu & Liu, 2004b). Проблем одабира ефикасног и репрезентативног скупа атрибута предмет је многобројних истраживања, почев од Хјуа и Лиуа[139] (Hu & Liu, 2004b), Јиа[140] са сарадницима (Yi, Nasukawa, Bunescu & Wayne, 2003), Попеску и Ецонија[141] (Popescu & Etzioni, 2005) до новијих Занга[142] (Zhang & Zhu, 2013), Заиа[143] (Zhai, Liu, Xu & Jia, 2011), Заоа[144] (Zhao, Li & Wang, 2013) и др. Посебан значај у овим истраживањима је дат развоју метода аутоматске селекције репрезентативног скупа атрибута. У раду (Hu & Liu, 2004b) дефинисана су четири разлога због којих је важно генерисати ефикасан алгоритам за предвиђање оптималног скупа атрибута производа који су предмет продаје на неком сајту електронске трговине: (1) када продавац у понуди има велики број производа различитих произвођача

[137] www.amazon.com
[138] Francesco Barbieri and Horacio Saggion
[139] Minqing Hu and Bing Liu
[140] Jeonghee Yi
[141] Ana-Maria Popescu and Oreon Etzioni
[142] Yu Zhang and Weixiang Zhu
[143] Zhongwu Zhai and Bing Liu, and Hua Xu and Peifa Jia
[144] Peng Zhao and Xue Li and Ke Wang

и садржај листе се често мења, одржавање листе атрибута тих производа је дуготрајан и захтеван задатак; (2) речи којима купци описују производ могу се разликовати од речи које су произвођачи утврдили као оне којима дефинишу атрибуте; (3) купци могу желети да оцене и оне атрибуте које произвођач није уврстио у листу; (4) у извесним случајевима произвођачи могу желети да намерно изоставе неке атрибуте производа желећи да релативизују њихов значај. Због ових и других разлога (ефикасних метода груписања, оцењивања семантичке сличности међу речима, увођења фраза у методу оцењивања и др.) аутоматизацијом избора репрезентативног скупа атрибута долази се до аутономних механизама који се могу примењивати и кад је реч о другим предметима посматрања, а не само о предметима електронске трговине. Тако, на пример, Ји[145] са сарадницима у раду (Yi, Nasukawa, Bunescu & Wayne, 2003) примењује јединствени алгоритам над различитим скуповима: електронских уређаја, веб докумената општег типа и музичких албума. Други важан корак у алгоритмима класификације на нивоу атрибута је поступак одређивања поларитета осећања скупа атрибута до кога се дошло у претходном поступку. У том случају, најчешће се користе семантичке мреже попут *WordNet*-а[146] (Yi, Nasukawa, Bunescu & Wayne, 2003; Hu & Liu, 2004a; Zhai, Liu, Xu & Jia, 2011; Zhao, Li & Wang, 2013) и *ConceptNet*-а[147] (Poria, Cambria, Hussain & Hunag, 2015; Cambria & Hussain, 2015), лексикони сентименталних речи и израза (Yi, Nasukawa, Bunescu & Wayne, 2003), базе података са примерима који носе поларитет осећања (Yi, Nasukawa, Bunescu & Wayne, 2003) и алгоритми за оцену семантичке сличности (Popescu & Etzioni, 2005; Zhang & Zhu, 2013). У трећем кораку приступа се сумаризацији сентимената. Према Ким[148] (Kim, Ganesan, Sondhi & Zhai, 2011), истраживања у области аутоматизације сумирања оцена о исказаним осећањима представљају синтезу знања области класификације на основу осећања и сумаризације текстова, при чему су истраживања у

[145] Jeonghee Yi
[146] https://wordnet.princeton.edu/
[147] http://sentic.net/about/
[148] Hyun Duk Kim

области сумаризације текстова фокусирана на методологије екстракције истакнутих реченица из текста и њихове кохерентне организације у смислу изградње резимеа целог текста. Међутим, Лиу (Liu, 2010) истиче да технике сумаризације текста нису квантитативне природе, па их није могуће применити у техникама сумаризације осећања када се жели исказати и степен поларитета осећања (нпр. када би уместо реченице која представљала резултат алгоритма сумаризације текстова „Већини корисника се допада овај производ“ требало добити сумаризацију у форми реченице „Овај производ је позитивно оценило 72% корисника.“). Због тога се развијају специфичне методе сумаризације, познате под називом „*Aspect-based Opinion Summarization*“ (Kim, Ganesan, Sondhi & Zhai, 2011), а резултати сумаризације, осим текстуално, могу бити приказани нумерички (статистичким табелама) и графички (Mei, Ling, Wondra, Su & Zhai, 2007).

4.2 Методе класификације према субјективности

Класификација субјективности (Esuli & Sebastiani, 2006), као поступак којим се идентификује текст који носи субјективно мишљење у односу на онај који носи искључиво чињеничне информације, може бити примењен на свим нивоима анализе осећања: документа, реченице, фразе и речи. Утврђивање субјективне оријентације речи и фраза може се спроводити генерисањем и применом лексикона сентименталних речи и израза (Riloff, Wiebe & Wilson, 2003; Mohammad & Turney, 2013), изградњом и применом семантичких мрежа (Esuli & Sebastiani, 2006; Strapparava & Valitutti, 2004) са концептима који се односе на емотивна понашања као што су *Sentiwordnet*[149], *WordNet-Affect*[150]. У изградњи ових мрежа (в. одељке 4.4.3 и 4.5.1) користе се правила за утврђивање субјективности речи и фраза (Riloff, Wiebe & Wilson, 2003; Riloff & Wiebe, 2003), алгоритми груписања (Hatzivassiloglou & McKeown, 1997; Wiebe, 2000) и алгоритама за утврђивање семантичке сличности субјективних речи и фраза (Hatzivassiloglou & Wiebe, 2000; Yu &

[149] http://sentiwordnet.isti.cnr.it/
[150] http://wndomains.fbk.eu/wnaffect.html

Hatzivassiloglou, 2003) и др. Прва истраживања у области аутоматског откривања субјективности могу се наћи у радовима Виби (Bruce & Wiebe, 1999; Wiebe, Bruce & O'Hara, 1999; Wiebe, 2000; Hatzivassiloglou & Wiebe, 2000; Riloff, Wiebe & Wilson, 2003; Riloff & Wiebe, 2003; Wiebe et al. 2004), Хацивасилоглуа (Hatzivassiloglou & McKeown, 1997; Hatzivassiloglou & Wiebe, 2000; Yu & Hatzivassiloglou, 2003) Мекеонове (Hatzivassiloglou & McKeown, 1997), Брусове (Bruce & Wiebe, 1999; Wiebe, Bruce & O'Hara, 1999; Wiebe et al. 2004;), Рилофове (Riloff, Wiebe & Wilson, 2003; Riloff & Wiebe, 2003) и касније, у радовима (Pang & Lee, 2004), (Esuli & Sebastiani, 2006) и др. Према Виби (Wiebe, 2000), субјективност придева и градабилност придева статистички значајно утичу на субјективност реченица, али и субјективност речи које се појављују само једном у посматраном корпусу (*hapax legomena*) као и субјективност колокација (*n*-грама) (Wiebe et al., 2004). Рилоф је истраживала утицај именица уз помоћ семантичких речника сачињених од именица које носе субјективно значење. Речници су генерисани помоћу основног скупа од 850 именица добијених из претходно анотираног скупа за учење (Riloff, Wiebe & Wilson, 2003) тако што су примењени алгоритми за екстракцију семантички сличних речи речима из основног скупа. На тај начин је генерисано око 4000 именица помоћу којих је обучаван Наивни Бајесов класификатор идентификације субјективитета реченица са прецизношћу 77% и одзивом 64%. Есули и Себастијани су за проширење скупа придева посматрали и семантичке релације синонимије и антонимије, а такође су користили и WordNet, уз претпоставку да литерали[151] са истим поларитетом осећања имају и дефиниције (*glosses*) са истим поларитетом осећања.

Алгоритми класификације субјективности могу представљати први корак у процесу бинарне класификације осећања према поларитету. Пенг и Ли су показали (Pang & Lee, 2004) да се тачност класификације осећања може значајно побољшати, ако се из скупова за учење претходно елиминишу реченице које не носе субјективно значење.

[151] в. одељке 3.3.1 и 3.3.3

4.3 Методе класификације према поларитету осећања

Поступак којим се утврђује позитиван или негативан поларитет осећања у тексту спада у класу проблема класификације осећања према поларитету. Обично се користи након примене алгоритма класификације према субјективности или се користи самостално у случајевима када је скуп над којим се примењује већ обележен као субјективан или када алгоритам класификације није бинарни већ укључује и класу „неутралних" текстова. Као и код идентификације субјективности, може бити примењен на свим нивоима анализе осећања: документа (Pang, Lee & Vaithyanathan, 2002), реченице (Yu & Hatzivassiloglou, 2003), фразе (Wilson, Wiebe & Hoffmann, 2009) и речи (Popescu & Etzioni, 2005; Esuli & Sebastiani, 2006). Исто тако, могу се користити лексикони сентименталних речи и израза (Yi, Nasukawa, Bunescu & Wayne, 2003; Mohammad & Turney, 2013), семантичке мреже као што су *Sentiwordnet*[152] (Esuli & Sebastiani, 2006) и *WordNet-Affect*[153] (Strapparava & Valitutti, 2004), правила за утврђивање субјективности речи и фраза (Riloff, Wiebe & Wilson, 2003; Riloff & Wiebe, 2003), алгоритми сегрегације (Hatzivassiloglou & McKeown, 1997; Wiebe, 2000) и алгоритми за утврђивање семантичке сличности субјективних речи и фраза (Hatzivassiloglou & Wiebe, 2000; Yu & Hatzivassiloglou, 2003), методе машинског учења (Pang, Lee & Vaithyanathan, 2002; Mullen & Collier, 2004; Pang & Lee, 2005; Mullen & Malouf, 2006; Zhang, Huang & Wu, 2008; Alsharif, Alshamaa & Ghneim, 2013) као и различите врсте хибридних метода.

Значај метода класификације осећања лежи у њиховој применљивости у различитим доменима. Милен и Малуф (Mullen & Malouf, 2006) су применили методу машинског учења, Наивни Бајесов алгоритам, као и варијанту Тарнијеве методе из групе LSO метода, тзв. PMI-IR методу, на скупу текстова који су добијени парсирањем постова са сајтова који се баве политичким темама (Mullen & Malouf, 2008). Сомаундаран је са сарадницима (Somasundaran, Ruppenhofer & Wiebe, 2007) истраживао могућност

[152] http://sentiwordnet.isti.cnr.it/
[153] http://wndomains.fbk.eu/wnaffect.html

идентификације осећања и ставова изражених у текстовима који чине корпус транскрипата дебата забележених на видеу (AMI[154] *Meeting Corpus*). Пенг и сарадници (Pang, Lee & Vaithyanathan, 2002) су показали да се методама машинског учења успешно могу класификовати рецензије филмова. Ји је са сарадницима (Yi, Nasukawa, Bunescu & Wayne, 2003) примењивао LSO методе у циљу утврђивања оцена производа (дигитални фото-апарати и музички дискови) који су предмет електронске трговине. Тарни (Turney, 2002) је, пак, у својој анализи ненадгледаних метода класификације на основу осећања користио Епинионс колекцију (в. одељак 4.1.1) рецензија различитих врста производа и услуга. Мохамад и Јанг[155] (Mohammad & Yang, 2011) су формирали корпус писама из више извора: скуп онлајн[156] љубавних писама (*Love Letters Corpus* -LLC), скуп порука корисника Миленијум[157] пројекта о стварима које не воле (hate mail corpus -HMC) (Mohammad, 2011), скуп опроштајних[158] порука (*Suicide Notes Corpus* -SNC) и Енрон[159] корпус електронске поште (*The Enron Email Corpus*) како би анализирали дистрибуцију речи са поларитетом осећања у разним врстама писама и како би утврдили разлике у изражавању осећања мушкараца и жена. Најзад, има и истраживања чији су изазов литерарни текстови, попут поезије (Alsharif, Alshamaa & Ghneim, 2013) и бајки (Mohammad, 2011; Alm, Roth & Sproat, 2005). У свом раду (Mohammad, 2011), Мохамад је користио *Fairy Tale Corpus* (FTC)[160] да анализира бајке (браћа Грим, Андерсен, Портер) и утврди дистрибуцију речи којима се изражавају осећања, а Алм (Alm, Roth & Sproat, 2005) се са сарадницима бавила класификацијом бајки браће Грим методом машинског учења (метода потпорних вектора).

[154] http://groups.inf.ed.ac.uk/ami/download/
[155] Saif M. Mohammad and Tony (Wenda) Yang
[156] lovingyou.com
[157] http://www.ratbags.com/
[158] http://www.well.com/art/suicidenotes.html?w
[159] https://foundationdb.com/key-value-store/documentation/enron.html
[160] https://www.l2f.inesc-id.pt/wiki/index.php/Fairy tale corpus

4.4 Методе класификације према јачини осећања

Унапређење метода класификације осећања може се одвијати у правцима:

- постизања веће тачности класификације
- постизања свеобухватности, тј. применљивости на различитом доменима (*cross-domain sentiment analysis*) (Liu, 2012)
- применљивости на различитим језицима (*cross-language sentiment analysis*) (Balahur & Turchi, 2014; Balahur & Perea-Ortega, 2015)
- унапређења техничких карактеристика (повећања скалабилности како хардвера, тако и софтверских компоненти, а посебно база података) како би се могао процесирати велики, комплексан скуп неструктуираних и разнородних података (*big data sentiment analysis*) (Kaushik & Mishra, 2014; Cambria, Rajagopal, Olsher & Das, 2013)
- постизања већег степена интерпретације људских осећања препознавањем различитих емоционалних стања, степена њихове изражености и међусобне корелације. (*sentiment rating prediction*, (Liu, 2012) или *fine-grained sentiment analisys* (Strapparava & Mihalcea, 2008))

Препознавање степена израженост осећања у квантитативном (интензитет) или квалитативном (врсте емоционалних стања) облику познато је под називом *fine-grained sentiment analisys* и може се остварити:

- увођењем емоционалних категорија које се изводе из дискретних и димензионалних когнитивних теорија осећања као што су Екманова, Изардова, Томпкинсонова, Плутчикова, итд.
- израчунавањем јачина израженост осећања и њиховом нумеричком репрезентацијом (квантитативна оцена израженост осећања)
- утврђивањем већег броја класа које представљају уређени низ јачина осећања („слабо-умерено-јако“)
- симболичким обележавањем текстова и аутоматским проналажењем осећања у тексту на основу додељених симбола.

У наредним одељцима биће више речи о методама које су засноване на моделима когнитивних теорија и о методама које се баве препознавањем и оценом степена израженості осећања.

4.4.1 Емоционалне категорије изведене из дискретних когнитивних теорија

Теорије дискретних емоција (Strongman, 2003) заснивају се на претпоставци да постоји мали број основних осећања, али много речи којима се описују, како оне, тако и њихове комбинације (Hobbs & Gordon, 2011). Основни скуп осећања чине оне које су биолошки детерминисане и које су заједничке свим људима. Њихов број варира од 2 до 26, зависно од теорије и аутора. Истакнути теоретичари ове групе су: Арнолд, Екман, Фријда, Грej, Изард, Томкинс, Вотсон[161] (Beck, 2000; Hobbs & Gordon, 2011) и др. Табела 4.1 приказује скупове основних емоционалних категорија којима су наведени аутори дефинисали основна осећања (емоције) и скупове њихових лексичких репрезентација.

[161] Magda Arnold, Paul Ekman, Nico Frijda, Jeffrey Alan Gray, Carroll Ellis Izard, Silvan Tomkins, John Watson

Табела 4.1 Скупови основних осећања аутора теорија дискретних емоција

Аутор	Основне емоције (осећања)	Број емоција
Arnold	anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	11
Ekman	anger, disgust, fear, joy, sadness, surprise	6
Frijda	desire, happiness, interest, surprise, wonder, sorrow	6
Gray	rage, terror, anxiety, joy	4
Izard	anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	10
Plutchik	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	8
Tomkins	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	9
Weiner	happiness, sadness	2
Watson	fear, love, rage	3
Matsumoto	joy, anticipation, anger, disgust, sadness, surprise, fear, acceptance, shy, pride, appreciate, calmness, admire, contempt, love, happiness, exciting, regret, ease, discomfort, respect, like	22

Екманов (Scherer & Ekman, 1984; Ekman, 1992) скуп шест основних осећања (љутња–*anger*, гађење–*disgust*, страх–*fear*, радост–*joy*, туга–*sadness* и изненађење-*surprise*), често је коришћен[162] у вишекласној анализи осећања, над различитим скуповима података и применом различитих метода. Инкпен са сарадницама (Inkpen, Keshtkar & Ghazi, 2009) генерише скуп од 7 класа тако што Екмановом скупу од 6 класа додаје и неутралну (*non-emotion*), класификујући скуп наслова новинских чланака, који представљају податке за задатак SemEval 2007 Task 14[163] и скуп постова са блога који су анотирани осећањима. SemEval 2007 Task 14 постављен је као задатак у организацији ACL (The Association for Computational Linguistics), међународног удружења истраживача у области рачунарске лингвистике, у циљу истраживања веза између емоција и лексичке семантике. Задатак је учесницима обезбедио скуп од 250 анотираних наслова и 1000 наслова у форми скупа за тестирање модела који су развијани. Скуп је коришћен у вишекласној анализи осећања у истраживању Инкпен и њених сарадника који су увели механизам оцене према интензитету користећи скалу опсега (-100,100) над комплетним

[162] Посебно се користи у методама аутоматског препознавања осећања изражених на лицу http://www.paulekman.com/category/facial-recognition-technology/
[163] http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml

скупом од 1250 наслова као скупом за учење, применом десетоструке унакрсне валидације, методом потпорних вектора у радном окружењу *Weka*.[164] У другом експерименту, да би применили исту методологију, аутори су користили корпус постова са блога сачињен од 2090 реченица екстрахованих из 173 поста и ручно анотираних осећањима.

SemEval 2007 Task 14 скуп (1250 наслова новинских чланака) користи се и у раду (Strapparava & Mihalcea, 2008) у изградњи ручно анотираног скупа за учење помоћу Екмановог скупа етикета. Анотацију је извршило шест испитаника тако што је сваком наслову додељена одговарајућа реч (из скупа основних Екманових речи) којом се описује осећање садржано у наслову, као и степен изражености тог осећања у опсегу [0,100], где 0 представља одсуство осећања, а 100 њено максимално присуство. Анотирани корпус наслова је подвргнут вишекласној анализи осећања тако што је методом латентне семантичке анализе (*Latent Semantic Analysis*)[165] за сваки од наслова утврђена семантичка сличности између текста наслова и посматране класе, а добијена вредност упоређена је са ручно додељеном ознаком. У другом експерименту истог истраживања, Страпарава и Михалчеа су употребили корпус који је садржао 8761 постова[166] са блога, анотираних Екмановим скупом осећања у поступку вишекласне анализе осећања Наивном Бајесовом методом.

Алм је са сарадницима (Alm, Roth & Sproat, 2005), такође, користила Екманов скуп у поступку ручне анотације 22 бајке браће Грим[167] (в. одељак 4.3) тако што је свака од 1580 реченица анотирана једном од емоционалних категорија из датог скупа. Тако анотиран скуп коришћен је касније у класификационим методама машинског учења. Шулц са сарадницама (Schulz, Thanh, Paulheim & Schweizer, 2013), користи скуп од 7 класа тако што Екмановом скупу од 6 класа додаје и неутралну у поступку ручне анотације скупа за учење од 200 твитова на енглеском језику. Анотирање је извршено

[164] http://www.cs.waikato.ac.nz/ml/weka/
[165] в. одељак 4.5.2
[166] са сајта LiveJournal.com
[167] Fairy Tale Corpus

путем јавне анкете и коначан скуп (одабрани су само они твитови који су означени на идентичан начин од стране више од 50% испитаника) који је бројао 114 твитова коришћен је у десетострукој унакрсној валидацији три методе машинског учења које су међусобно упоређене: метода која користи наивни Бајесов бинарни модел[168], метода која користи наивни Бајесов вишезначни модел[169] и метода потпорних вектора[170]. У раду (Neviarouskaya, Prendinger & Ishizuka, 2009) користи се Изардов скуп од 9 осећања (*Anger, Disgust, Fear, Guilt, Interest, Joy, Sadness, Shame, Surprise*) на основу кога се изграђује свеобухватна база података емоционалних атрибута (*Affect database*) која садржи табеле за: емотиконе, скраћенице, придеве, прилоге, именице, афективне знакове интерпункције и модификаторе поларитета. Сви чланови табела означени су емоционалном категоријом и интензитетом из опсега [0,1]. Означавање су извршила ручно три независна анотатора. Аутори ову базу користе како би непознату реченицу изразили вектором емоционалних атрибута садржаних у бази, што омогућава једноставну калкулацију поларитета осећања реченице као и његовог интензитета. *Affect database* је оцењен и на колекцији од 700 ручно анотираних реченица из постова са блога.[171] Значајна истраживања заснована на Екмановој категоризацији могу се наћи и у радовима (Aman & Szpakowicz, 2007), (Smith & Lee, 2012), а Шатлс и Ајд[172] (Suttles & Ide 2013) су над изузетно великим скупом твитова применили методе машинског учења (наивни Бајес и максималну ентропију) генеришући вишекласне класификационе моделе засноване на Плутчиковом моделу (Plutchik & Hope, 1997) од 8 основних осећања.

[168] Naïve Bayes Binary Model
[169] Naïve Bayes Multinomial Model
[170] Support Vector Machine
[171] http://www.nielsenbuzzmetrics.com
[172] Jared Suttles and Nancy Ide

4.4.2 Емоционалне категорије изведене из димензионалних когнитивних теорија

Плутчик[173] је представник димензионалних теорија (Strongman, 2003) који осећање посматра кроз његов интензитет, поларитет и степен сличности са другим осећањима. Модел (Plutchik & Hope, 1997) који је он истраживао и предложио, често се назива „Плутчиковим точком емоција“ (слика 4.1)[174] и представља три нивоа интензитета осећања у облику концентричних кругова: слаб, средњи и јак.

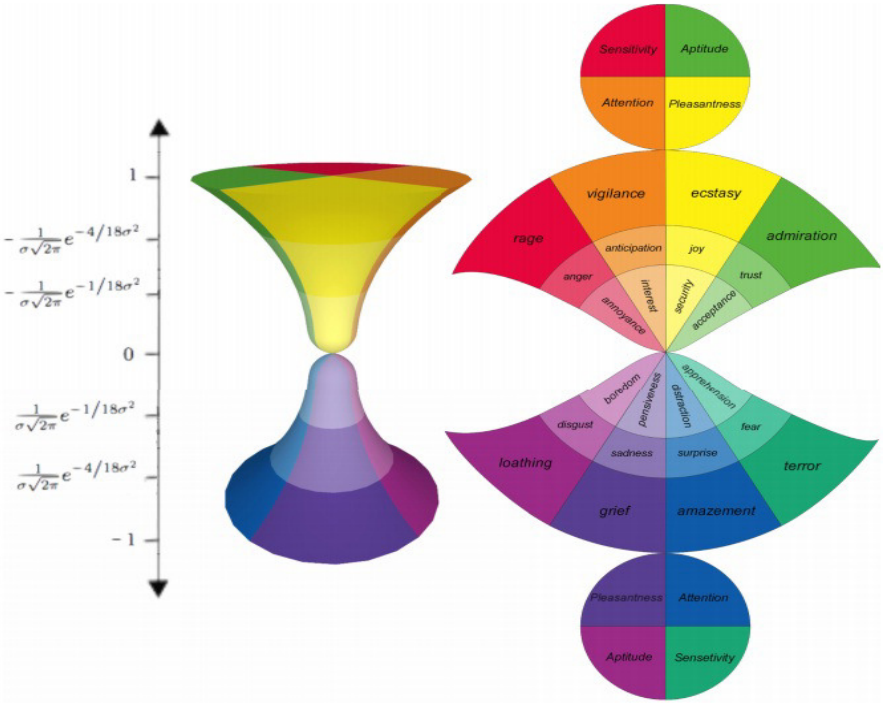
Слика 4.1 Плутчиков „точак емоција“

На сваком нивоу дефинисано је 4 пара биполарних емоција. Парови међусобно граде опружен угао, а области између суседних осећања дефинишу сложенија емоционална стања која представљају њихове комбинације. Иако је овакав модел критикован (TenHouten, 2007, pp. 22-24) због своје непотпуности и апстракције проблема изражавања комплексних емоција које настају и трају као последице интеракције других емоција које им претходе, у области рачунарске лингвистике је наишао на широку примену, посебно у методама аутоматске и полуаутоматске изградње

[173] Robert Plutchik
[174] слика преузета са: http://contentinsights.co/wp-content/uploads/2013/04/Plutchiks-Wheel-of-Emotion.jpg

семантичких ресурса као што су лексикони сентименталних речи и израза (*EmoLex*[175]) (Mohammad & Turney, 2013), доменске онтологије (HEO[176]) (Cambria & Hussain, 2015) и други ресурси (*SenticNet*[177]) (Cambria, Olsher & Rajagopal, 2014) који се користе у задацима анализе осећања. Плутчиков „точак емоција" послужио је као инспирација Камбрији[178] који је са сарадницима (Cambria, Livingstone & Hussain, 2012) дефинисао модел емоција (осећања) у форми пешчаног сата (слика 4.2).[179] Главни мотив изградње овог модела је интеграција афективних информација са текстом као изражајним средством. „Пешчаник емоција" (*The Hourglass of Emotions*) је модел који редефинише Плутчиков најпре тиме што уводи 4, међусобно независне, афективне димензије (пријатност - *Pleasantness*, заинтересованост - *Attention*, осетљивост - *Sensitivity*, способност - *Aptitude*) дуж којих се протеже 6 нивоа који одређују интензитет афективности (аутори их називају нивоима активације) и који се могу представити и нумерички функцијом $G(x) = -\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}$ чије су вредности у опсегу [-1,1]. Нивои активације одређени су вредностима функције $G(x)$ у тачкама 1, 1/2, 2/3, -1/3 , -2/3 и -1 (табела 4.2).

[175] http://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[176] Human emotion ontology (HEO)
ftp://ftp.uk.freesbie.org/sites/downloads.sourceforge.net/h/he/heronto/HEO%20ONTOLOGY%20V%2030.09.2012.owl
[177] SenticNet http://sentic.net/
[178] Erik Cambria
[179] слика преузета из књиге (Cambria & Hussain, 2012)



Слика 4.2 Модел „пешчани сат емоција“

Сваки од сегмената у моделу је идентификован јединственом емоцијом, а заједно представљају скуп од 24 основне емоције, па модел може бити посматран и као димензионалан и као дискретан.

Табела 4.2 Шест нивоа активације у моделу „пешчани сат емоција“ одређених вредностима функције $G(x)$

Interval	Pleasantness	Attention	Sensitivity	Aptitude
[G(1), G(2/3)]	ecstasy	vigilance	rage	admiration
[G(2/3), G(1/3)]	joy	anticipation	anger	trust
[G(1/3), G(0)]	serenity	interest	annoyance	acceptance
[G(0), G(-1/3)]	pensiveness	distraction	apprehension	boredom
[G(-1/3), G(-2/3)]	sadness	surprise	fear	disgust
[G(-2/3), G(-1)]	grief	amazement	terror	loathing

На основу модела „пешчани сат емоција“, Камбрија (Cambria, Livingstone & Hussain, 2012) формира правила формалног описивања емоција вишег нивоа као комбинација основних. У табели 4.3[180] приказани су услови формирања

[180] Табеле 4.2 и 4.3 преузете из књиге (Cambria & Hussain, 2012).

скупа емоција другог нивоа на основу модела „пешчани сат емоција“. На сличан начин могу се формирати и правила за генерисање емоција трећег нивоа (нпр. *jealousy= joy+trust+anger*).

Табела 4.3 Емоције другог нивоа на основу модела „пешчани сат емоција“.

	Attention>0	**Attention<0**	**Aptitude>0**	**Aptitude<0**
Pleasantness>0	optimism	frivolity	love	gloat
Pleasantness<0	frustration	disapproval	envy	remorse
Sensitivity>0	aggressiveness	rejection	rivalry	contempt
Sensitivity<0	anxiety	awe	submission	coercion

На основу модела „пешчани сат емоција“, интензитет емоције вишег нивоа од основног, према (Cambria, Livingstone & Hussain, 2012), одређен је интензитетом сваке од афективних димензија, а поларитет је алгебарска сума (76) поларитета појединачних емоција нижег нивоа које учествују у њеном формирању.

$$p = \sum_{i=1}^{N} \frac{\text{Pleasantness}(c_i) + |\text{Attention}(c_i)| - |\text{Sensitivity}(c_i)| + \text{Aptitude}(c_i)}{3N} \quad (76)$$

где је c_i концепт емоције, N је укупан број концепата, 3- нормализациони фактор. Модел „пешчани сат емоција“ користи се у задацима класификације осећања, нпр. Камбрија га је са сарадницима (Cambria, Hussain & Havasi, 2010) употребио у процесу класификације мишљења и искустава пацијената изнетих преко онлајн платформе UK *National Health Service* (NHS) као и при класификацији 5000 постова са блога.[181] Такође, коришћен је у класификацији на основу осећања 40 прича за децу, применом методе потпорних вектора (Lertsuksakda, Pasupa & Netisopakul, 2015).

4.4.3 Методе квантитативне оцене израженості осећања

Једно од првих истраживања могућности квантитативног оцењивања степена израженості осећања дали су Пенг и Ли (Pang & Lee, 2005). Они су

[181] са сајта LiveJournal.com

пошли од претпоставке да се скуп за учење може формирати тако да садржи текстове који представљају препоруке са веб сајтова за оцену производа и услуга. У свом раду, најпре су формирали скуп који су сачињавале рецензије филмова које су оценили онлајн корисници бројем звездица (од 1 до 5) што је омогућавало да се свака рецензија класификује у једну од 5 класа, зависно од просечног броја звездица, а затим су експериментализали са 3, 4 и 5-то класним скуповима за учење и три класификатора: регресионом анализом, мултикласном методом потпорних вектора и једном од метода оцене семантичке сличности, тзв. метричким обележавањем (енг. metric labeling), чиме су добили информације о „лошим", „осредњим" и „добрим" филмовима. Лонг са сарадницима (Long, Zhang & Zhut, 2010) унапређује методу Пенг и Ли тако што интуитивно уводи претпоставку да побољшање селекције текстова који чине скуп за учење може довести до повећања тачности алгоритма који се обучава таквим скупом. Испитивање своје претпоставке спровели су над текстовима рецензија сајта о туристичким понудама,[182] а као меру информатичког доприноса текстова користили су комплексност Колмогорова[183] како би селектовали оне текстове код којих је укупно информатичко растојање између текста и одабраног скупа предиктора[184] минимално, тј. оне текстове који садрже „свеобухватну анализу". Најзад, аутори су извели серију експеримената како би показали успешност побољшане верзије алгоритма Пенг и Ли (Pang & Lee, 2005) када се као скуп за учење користи његов подскуп добијен предложеном методом селекције.

Проблему нумеричке оцене поларитета осећања у тексту могуће је приступити на више начина. Један од првих алгоритама (1993. год.) за квантитативну анализу текстова, *Linguistic Inquiry and WordCount* (LIWC[185]) (Pennebaker, Booth & Francis, 2007) развила је на Универзитету у Тексасу група истраживача предвођена Пенабејкером[186]. LIWC користи стратегију фреквентности речи како би анализирао и садржину и стил писања неког

[182] http://www.tripadvisor.com/
[183] Kolmogorov complexity
[184] о предикторима је више било речи у одељцима 1.3.2 и 2
[185] http://www.liwc.net/
[186] James W. Pennebaker, Roger J. Booth, Martha E. Francis

текста, па га примењују психолози да истраже везу између лингвистике и психологије (Tausczik & Pennebaker, 2010). Најважнију улогу има речник који за енглески језик садржи око 4500 речи и корена речи дефинисаних помоћу једне или више категорија (параметара). На пример, реч „плакао“ (*cried*) у речнику је означена помоћу 5 параметара: туга (*sadness*), негативна емоција (*negative emotion*), укупан утисак (*overall affect*), глагол (*verb*), глагол прошлог времена (*past tense verb*). Када се ова реч нађе у тексту, свака од скала ових 5 параметара се повећа за 1. Приликом анализе текста, разматра се око 80 различитих параметара подељених у неколико група: групу од 4 параметра општег типа (укупан број речи у тексту, просечан број речи по реченици, проценат речи нађених у речнику уграђеном у софтвер, проценат речи дужих од шест слова); групу од 22 лингвистичка параметра (нпр. проценат врста речи у реченици); 32 параметра који се односе на психолошке конструкције (емоција, спознаја, мишљење, биолошки процес); 7 параметара који карактеришу личност (посао, забава, кућа, итд.); 12 параметара који означавају интерпункцијске знаке (зарез, тачка, узвичник,...) и др. Алгоритам секвенцијално анализира речи у тексту и израчунава фреквенције појављивања свих 80 параметара. Анализом кумулативних процената параметара који се односе на афективна понашања могу се добити квантитативне оцене поларитета осећања у текстовима. Софтвер се успешно примењује код дужих текстова, а резултате приказује у облику табела које се директно могу користити у статистичким алатима. Верзија LIWC2007 омогућава имплементацију речника и на другим језицима. Развој LIWC речника за српски језик, *LIWCser*, предмет је истраживања приказаних у радовима (Bjekić, Lazarević, Erić, Stojimirović & Đokić, 2012; Bjekić, Lazarević, Živanović & Knežević, 2014). *LIWCser* садржи 12103 одредница, разврстаних у 65 категорија, од којих је 4328 описано афективним, а 2444 когнитивним категоријама.

Такође значајан алгоритам за квантитативну оцену степена изражености осећања је *SentiStrength* (Thelwall, Buckley, Paltoglou, Cai & Kappas, 2010),[187]

[187] http://sentistrength.wlv.ac.uk/

лексички алат за вишекласну анализу осећања на нивоу реченице којим се јачина осећања може утврдити на скали опсега [1,5] у оба смера (позитивном и негативном). И овај алгоритам користи речник (грађен по узору на LIWC) чије одреднице имају дефинисан поларитетом осећања и нумеричку ознаку јачине тог поларитета. Поред тога, користи низ примера нестандардне употребе правописа за изражавање осећања (емотикони, вишеструки знакови интерпункције, скраћенице попут „lol“, „btw“), али и друга лингвистичка средства којима се мења интензитет осећања као што су: модулатори јачине, попут „веома“, „незнатно“, „потпуно“, речи и фразе које врше обрнуту поларизацију као у примерима „нимало леп“, „ограничених способности“, речи које представљају уобичајени начин негације оних са којима се налазе у пару, итд.). *SentiStrength* је развијен помоћу скупа од 2.600 ручно класификованих коментара преузетих са друштвене мреже *MySpace*,[188] а оцењен је на случајном узорку од 1.041 коментара са исте мреже. Његове главне карактеристике су да:

- користи методологију машинског учења за оптимизацију јачина осећања одредница у речнику (алгоритам почиње од степена поларитета одредница на почетку добијених ручном анотацијом, да би се, затим, за сваку одредницу, сукцесивно, утврдило да ли повећање или смањење степена њеног поларитета за 1 утиче на повећање свеукупне тачности алгоритма класификације);
- користи специфичне методе за проналажење степена поларитета на основу присуства интерпункције (нпр. јачи позитивни поларитет има израз „постигнут је погодак!!!“ од израза „постигнут је погодак“, а постојање узвичника у фрази „он је човек!!!“ мења поларитет речи човек са неутралног у позитиван);
- користи методе за корекцију правописа, с обзиром да су друштвене мреже, у највећем броју случајева, извор текстова који се анализирају, а аутори тих текстова не воде превише рачуна о правопису.

[188] https://myspace.com/

Унапређена верзија, *SentiStrength2,* примењена је у раду (Thelwall, Buckley & Paltoglou, 2012), у оцени јачине осећања у текстовима узетих са 6 различитих друштвених мрежа (*MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums*) који су ручно анотирани и чија је поузданост анотације утврђена применом Крипендорфовог[189] α теста. Добијени резултати анотације према осећањима су упоређени са резултатима алгоритама из обе групе метода машинског учења. Од метода надгледаног учења, за поређење су коришћени: *SVM, SLOG, ADA Boost, SVM Regression, Decision Table, Naïve Bayes, J48* и *JRip*. Сваки од алгоритама тестиран је 30 пута помоћу десетоструке унакрсне валидације употребом десет различитих скупова предиктора (100,200,...,1000). Методе машинског учења су показале статистички значајно боље резултате од *SentiStrength2* методе само у случају два скупа за тестирање: *BBC Forums* и *Twitter*-а. У случају ненадгледане *SentiStrength2* ML методе такође показују нешто боље резултате класификације.

Имплементацију *SentiStrength* речника за турски језик и примену овог ресурса у анализи осећања заснованој на лексикону користили су у свом раду (Vural, Cambazoglu, Senkul & Tokgoz, 2012) Вурал и његови сарадници. Оцену система који су развили спровели су помоћу скупа рецензија филмова на турском језику.

Један занимљив и користан начин примене *SentiStrength* методе дат је у раду Корнелије Карагеа[190] и њених колега (Caragea, Squicciarini, Stehle, Neppalli & Tapia, 2014). Они су истраживали гео-локацијску дистрибуцију поларитета осећања у твитовима објављеним за време природних несрећа, на примеру скупа твитер порука насталих у време налета урагана Сенди. Добијене резултате су приказали визуелно – на географској мапи и помоћу графикона. Истраживање је показало да је јачина осећања реципрочна у односу на два параметра: време протекло од главног удара несреће и просторну удаљеност извора твитер поруке у односу на центар удара.

[189] Klaus Krippendorf
[190] Cornelia Caragea

Лексички ресурс *SentiWordNet*[191] (Baccianella, Esuli & Sebastiani, 2010), развијен да семантичкој мрежи *WordNet* обезбеди информације о типу и интензитету осећања сваког синсета, користи се у многим истраживањима, самостално и у спрези са мрежом *WordNet* из које је настао, у процесима изградње ресурса за анализу осећања, модела за класификацију, метода за екстракцију речи са одређеним степеном поларитета осећања итд. У раду (Kirtsis, Tzekou, Besharat & Stamou, 2013) *SentiWordNet* је коришћен за израчунавање јачине осећања ($Polarity\ Strength$) изражених у чланцима Википедије помоћу:

$$Polarity\ Strength(a_i) = \frac{|\#polarized\ opinion\ phrases\ in\ (a_i)\ |}{|\#opinion\ phrases\ in\ (a_i)|}$$

па се поларитет осећања ($Polarity\ Value$) у чланку (a_i) може израчунати помоћу:

$$Polarity\ Value\ (a_i) = Polarity\ Distribution(a_i) \cdot\ Polarity\ Strength(a_i)$$

где је

$$Polarity\ Distribution(a_i) = \frac{|\#polarized\ segments\ (a_i)\ |}{|\#segments\ (a_i)|}$$

Осим метода које користе претходно анотиране лексиконе и друге ресурсе у задацима оцене јачине осећања, има истраживања која оцену те јачине граде на основу калкулација семантичке зависности међу одабраним елементима. У свом раду, Лиу и Сенеф[192] (Liu & Seneff, 2009) посматрају семантичку везу облика „прилог-придев-именица“ (нпр. “how good the food”, “how bad the service”, “quite good caesar”, итд.) која се добија као резултат парсирања реченице у хијерархијску структуру (тзв. *linguistic frame*). Скуп за развој састојао се од докумената који представљају рецензије услуга у ресторанима у девет градова који су оцењени и рангирани од онлајн корисника. Аутори су нумеричку вредност рецензије сваког докумената

[191] http://sentiwordnet.isti.cnr.it/
[192] Jingjing Liu, Stephanie Seneff

додељивали и свим речима које су се нашле у документу, а градиле су структуру „прилог-придев-именица“. Метода је оцењена на скупу од 1000 фраза облика „придев-именица“ који је анотиран ручно скалом опсега [1,5] како би се испитао утицај додавања прилога таквим структурама.

У раду (Qu, Ifrim & Weikum, 2010) користи се комбинација оцена поларитета осећања речи из лексикона сентименталних речи и израза и оцена које се израчунавају предложеним алгоритмом из докумената скупа за развој. Аутори предлажу и разматрају нову структуру, тзв. „врећу мишљења“ (енг. bag-of-opinions), скуп уређених тројки састављених од одговарајуће сентименталних речи, модулатора или модификатора сентимента и речи које представљају негацију сентимента. Сваки документ скупа за учење репрезентује се помоћу скупа оваквих тројки, генерисаних из самог текста, тако што се открива постојање ових елемената у дефинисаном оквиру речи. Оквир се формира када се пронађе сентиментална реч, а дужина оквира укључује четири речи лево и три десно од пронађене речи. Постојање модификатора и негације се испитује на левој страни оквира, а на десној се тражи модификатор. Интензитет осећања израчунава се на основу почетног поларитета осећања речи добијеног из лексикона, који може затим бити појачан или ослабљен модификаторима, док се негативним елементима интензитет може само смањити или потпуно изменити. За утврђивање тежине, односно утицаја свих елемената модификације и негације, аутори су користили модел линеарне регресије којим су рангирали документе и претпоставили да и припадајуће тројке имају исти начин рангирања. Пренаученост модела је избегнута применом алгоритма L2-регуларизације (L2-norm) којим се минимизира сума квадрата резидуала (в. одељак 2.4.1), али је као скуп за учење коришћен доменски неутралан скуп[193] оцењених текстова, како би се добио довољно велики скуп оцењених „тројки“, такав да

[193] употребљен је скуп од 350.000 хиљада оцена разних производа са сајта *Amazon.com*

се може успешно тестирати и над мањим, доменски оријентисаним[194] скуповима текстова.

4.4.4 Симболичко обележавање текстова и методе аутоматског откривања расположења у тексту на основу додељених симбола

Иако је у психологији дато много, више или мање различитих, дефиниција, може се рећи да се појам „расположење“ (енг. mood) односи на средње и дуготрајно афективно стање (Thelwall, Buckley, Paltoglou, Cai & Kappas, 2010) и разликује се од појма „емоција“ или „осећање“ (енг. emotion) којим се означава реакција краћег трајања на конкретне догађаје, агенсе или објекте (Strongman, 2003).[195] Последњих година се развија дигитална симболика којом се, у текстуалним документима на интернету, могу исказати како осећања (емоције), тако и расположења. Једна од често коришћених класа таквих симбола су „емотикони“ (Ptaszynski, Rzepka, Araki & Momouchi, 2011). У раду (Zhao, Dong, Wu & Xu, 2012) приказана је метода у којој се 95 емотикона пресликава у 4 класе расположења (љутња - *angry*, гађење - *disgusting*, радост - *joyful*, туга - *sad*), а затим се скуп за учење, формиран од 3.5 милиона анотираних твитова са *Weibo*[196] мреже која има улогу Твитера у Кини, користи у обуци Наивног Бајесовог класификатора.

Неке друштвене мреже и блог системи омогућавају корисницима да, уз текстове којима изражавају своја осећања, додају иконице којима исказују свеукупно расположење или став (енг. moods icons). Једна од таквих мрежа је *Zazie*[197], италијанска друштвена мрежа намењена пасионираним љубитељима књига и лепе књижевности. *Zazie* омогућава корисницима да помоћу највише 2 од 25 могућих иконица (тзв. moods) изразе свој став о некој књизи, као и да га поделе са осталим корисницима мреже. У свом раду

[194] коришћена су три различита доменска скупа са сајта *Amazon.com* (оцене књига, музичких дискова и филмова) од којих се сваки састојао од скупа за учење величине 8000 оцена и скупа за валидацију величине 4000 оцена
[195] в. одељак 1
[196] http://weibo.com/login.php
[197] http://www.zazie.it

(Franzoni, Poggioni & Zollo, 2013) Францони[198] и њене сараднице показале су да је аутоматска класификација књига са *Zazie* мреже, на основу додељених иконица, могућа и да је тачност класификације задовољавајућа.

Експериментима са класификацијом према расположењу постова са блога бавио се Мишне[199] у раду (Mishne, 2005). Он је предложио и оценио методу у којој користи претпоставку да су управо оне речи којима се може описати неко расположење много карактеристичније за оне текстове који су анотирани посматраним расположењем него за текстове који то нису. Уз ту претпоставку Мишне је генерисао оптималне скупове предиктора за сваку класу расположења, а затим употребио методу потпорних вектора да класификује скуп од преко 800 хиљада твитова са сајта *LiveJournal* које су корисници сами анотирали избором једне од 132 ознаке „расположења".

4.5 Лексичко-семантички оријентисане (LSO) методе

Лексичко-семантички приступ (*Lexical Semantic Orientated Approach* – LSO *Approach*) проблемима класификације на основу осећања у текстовима заснива се на откривању и оцени позитивних и негативних сентименталних речи и фраза, садржаних у текстовима који су предмет анализе и претпоставци да присуство речи и фраза са позитивним, односно негативним, поларитетом доприноси поларитету текста на идентичан начин (Turney, 2002). Заједничка карактеристика метода ове врсте је да не захтевају претходну обуку и етикетирање посматраног скупа текстова. LSO методе користе две основне врсте техника у задацима сентимент класификације:

- технике засноване на лексиконима
- технике засноване на корпусима.

[198] Valentina Franzoni
[199] Gilad Mishne

4.5.1 Методе класификације засноване на лексикону

Употребом лексикона сентименталних речи и израза могу се спровести задаци класификације према субјективности и поларитету осећања, задаци сумаризације, екстракције осећања и др. Значајан допринос овој области може се наћи у радовима: (Bruce & Wiebe, 1999), (Wiebe, Bruce & O'Hara, 1999), (Wiebe, 2000), (Hatzivassiloglou & Wiebe, 2000), (Riloff & Wiebe, 2003), (Riloff, Wiebe & Wilson, 2003), (Yu & Hatzivassiloglou, 2003), (Wiebe et al., 2004), (Hu & Liu, 2004a, 2004b), (Kim & Hovy, 2004), (Strapparava & Valitutti, 2004), (Wilson, Wiebe & Hoffmann, 2005), (Esuli & Sebastiani, 2006), (Denecke, 2008), (Ohana & Tierney, 2009), (Dang, Zhang & Chen, 2009), (Taboada et al., 2011), (Kraychev & Koychev, 2011), (de Albornoz, Plaza & Gervás, 2012), (Mohammad & Turney, 2013), (Cambria, Olsher & Rajagopal, 2014), (Cambria & Hussain, 2015) и др. У наставку овог одељка приказаћемо неке од најважнијих лексикона, њихову структуру, начин изградње и примену.

Колекција речи и фраза са дефинисаним поларитетом осећања и могућношћу дефинисања интензитета тог осећања обично се назива лексикон сентименталних речи и израза. Улога овог лексикона у свим техникама анализе осећања је значајна и може бити директна (када се користе у самом процесу класификације) и посредна (када се користе као базе знања у фазама пре самог процеса класификације). Постоје бројна истраживања и методе аутоматског генерисања ове врсте дигиталних ресурса. Најчешће је реч о методама које користе речнике (Stone, 1966), или семантичке мреже (Kim & Hovy, 2004; Strapparava & Valitutti, 2004; Esuli & Sebastiani, 2006; Kamps, Marx, Mokken & De Rijke, 2004).

Методе генерисања лексикона сентименталних речи и израза засноване на речницима полазе од малог скупа сентименталних речи (енг. seed words), добијеног или ручним путем или из малих, референтних ресурса који описују сентиментално оријентисане концепте, који се затим увећава аутоматским методама које користе правила семантичке сличности, синонимије и антонимије над дигиталним речницима, семантичким мрежама и специјализованим или доменским ресурсима.

Тако је у раду (Strapparava & Valitutti, 2004) презентован процес изградње ресурса *WordNetAffect* који представља додатак принстонском WordNet-у (в. одељак 3.3.2) у форми тзв. афективних обележја (a-labels) PWN синсетова. Развој овог додатка текао је у две фазе. У првој фази генерисано је „језгро" синсетова, тзв. база *AFFECT*, којима су додељена афективна обележја, а у другој је извршено његово проширење помоћу релација *WordNet*-а. Изградња „језгра" кренула је од полазног скупа (seed words) који је садржао 1093 ручно анотираних речи. У складу са теоријом Ортонија[200], све речи „језгра", тј. базе AFFECT, класификоване су у неколико афективних класа (*emotion, cognitive state, trait, behaviour, attitude, feeling,...*), а афективна информација, *a-label,* о сваком појму базе *AFFECT* допуњена је и на основу знања и других психолошких теорија. Када је успостављена релација између речи у бази *AFFECT* и *WordNetAffect* синсетова, тада је сваки синсет добио одговарајућу ознаку афективне класе, тј. *a-label*. У другој фази развоја овог ресурса, употребљене су оне лексичко-семантичке релације *WordNet*-а за које важи да пропагирају исти поларитет осећања ка оним синсетовима који учествују у таквим релацијама. Нпр. уколико је синсет s_i означен афективном ознаком *emotion,* биће исто обележен и синсет s_j који је у релацији *also-see* са s_i. Релације које су коришћене у проширењу „језгра" су: *antonymy, similarity, derived-from, pertains-to, attribute, also-see*. Релације које могу, али не и обавезно, пропагирати афективну ознаку су: *hyperonymy, entailment, causes, verb-group* и оне су провераване ручно. На крају овог поступка *WordNetAffect* садржао је 2874 синсетова и 4787 литерала са *a-label* ознакама, а неки од примера дати су у табели 4.4.[201]

[200] Andrew Ortony
[201] део табеле из рада (Strapparava & Valitutti, 2004)

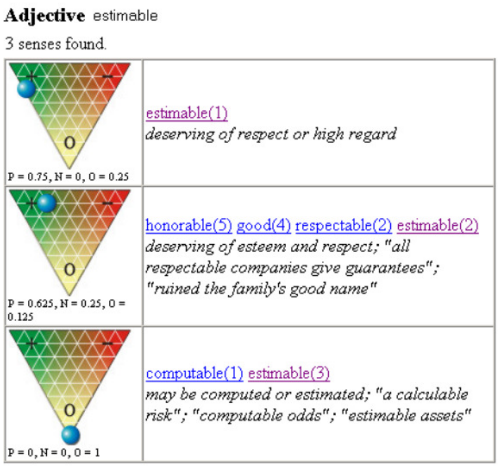
Табела 4.4 Ознаке *a-label* додељене неким *WordNetAffect* синсетовима

a-labels	***synsets***
EMOTION	noun anger#1,verb fear#1
MOOD	noun animosisy#1,adjective amiable#1
TRAIT	noun aggressiveness#1, adjective competitive#1
COGNITIVE STATE	noun confusion#2,adjective dazed#2
PHYSICAL STATE	noun illness#1,adjective all in#1
BEHAVIOUR	noun offense#1,adjective inhibited#1
ATTITUDE	noun intolerance#1,noun defensive#1
SENSATION	noun coldness#1,verb feel#3

Есули и Себастијани су у раду (Esuli & Sebastiani, 2006) изложили поступак изградње јавно доступног лексичког ресурса који се може користити у задацима анализе осећања. Реч је о ресурсу *SentiWordNet* који је заснован на *WordNet*-у и у коме је сваки синсет *s* означен помоћу три нумеричка показатеља интензитета и поларитета осећања: Obj(*s*), Pos(*s*), Neg(*s*), који показују колики је степен (изражен као реалан број из интервала [0,1]) објективности, позитивног и негативног поларитета датог синсета, односно сваког његовог литерала. Изградња *SentiWordNet*-а одвијала се у неколико корака. И у овом случају креиран је полазни скуп сентименталних речи као унија три ручно креирана скупа: *Lp, Ln* и *Lo*, који су представљали позитивно, негативно и неутрално оријентисане синсетове *WordNet*-а. Скупови *Lp* и *Ln* проширени су анализом релација синонимије и антонимије, као и лексичким релацијама (*also-see, derived-from, pertains-to, attribute*). Скуп *Lo* проширен је на другачији начин. Као проширење овог скупа узет је онај подскуп WordNet-а чији синсетови нису садржани у скуповима *Lp* и *Ln,* а садрже литерале који у лексикону GI[202] немају етикету *Positive* или *Negative*. Затим је уведена претпоставка да синсетови са сличним поларитетом имају „сличне“ дефиниције (*glosses*). Дефиниције синсетова представљене су у моделу TF.IDF векторског простора, а затим је овакав скуп вектора формирао скуп за учење два класификатора: *Rocchio* и SVM. Нормализацијом резултата оба класификатора добијен је комплетан *SentiWordNet*. У истом раду, аутори представљају и визуелни модел *SentiWordNet*-а у коме се сваки синсет

[202] General Inquirer lexicon, детаљније описан у наставку овог одељка.

представља троуглом чији углови представљају максималне вредности показатеља интензитета и поларитета осећања (слика 4.3)[203]. Лексички ресурс *SentiWordNet* коришћен је и у проширењу скупа етикета српског *WordNet*-а. Поступак проширења и даља употреба етикетираних синсетова у процесу изградње модела за класификацију осећања у текстовима на српском језику, који је предмет ове тезе, биће детаљније приказани у одељку 5.4.2.



Слика 4.3 Визуелна репрезентација у *SentiWordNet*-у ознака поларитета осећања за три синсета у којима учествује придев „*estimable*“ (цењен; проценљив)

Кампс[204] је са сарадницима користио PWN и у свом раду (Kamps, Marx, Mokken & De Rijke, 2004) предложио методу за одређивање поларитета осећања придева на основу растојања два синсета у хијерархији ворднета. У раду се растојање $d(t_1, t_2)$ између два придева t_1 и t_2 дефинише као најкраћи пут семантичких веза којим су ова два придева повезана. У том случају, семантичка оријентација посматраног придева a_t одређена је његовим релативним растојањем од синсетова придева „добар“ (*good*) и „лош“ (*bad*) и може се изразити као:

[203] Слика преузета из рада (Esuli & Sebastiani, 2006).
[204] Jaap Kamps

$$SO(a_t) = \frac{d(a_t, bad) - d(a_t, good)}{d(good, bad)}$$

одакле се закључује да придеви за које важи да је $SO(a_t) > 0$ (растојање посматраног синсета a_t до сентиментално позитивно поларисаног синсета „добар“ је мање од растојања до сентиментално негативно поларисаног синсета „лош“), имају позитиван поларитет осећања, а негативан у супротном.

Табоада са сарадницима (Taboada et al., 2011) у свом раду описује изградњу својеврсног „калкулатора“ семантичке оријентације (*Semantic Oriented Calculator* SO-CAL) засновану на две претпоставке: речи имају свој подразумевани поларитет осећања (енг. prior polarity) независан од контекста и тај се поларитет може изразити нумерички. У SO-CAL речнику, поларитет осећања одреднице дефинисан је поларитетом и јачином из интервала [-5,+5]. Речник је грађен ручном анотацијом речи узетих из три различита извора: (1) корпуса од 400 текстова који представљају рецензије производа осам различитих категорија производа са сајта *Epinions*[205]; (2) подскупа од 100 рецензија скупа од 2000 филмова које чине *Polarity Dataset* креиран и употребљен у раду (Pang & Lee, 2004); (3) позитивно и негативно означених речи из речника GI (*General Inquirer*). Речник представља унију четири подречника које карактерише да су све одреднице једног подречника речи исте врсте (именице, глаголи, придеви и прилози). И мада већину одредница чине појединачне речи, SO-CAL речник садржи и вишечлане лексичке јединице (енг. multi-word units – MWU) које имају предност над појединачним простим речима када се одређује семантичка оријентација неког текста из разлога што поларитет вишечлане лексичке јединице може бити другачији у односу на поларитет појединачних речи чланица. На пример, поларитет речи *funny* (забаван) је позитиван са интензитетом +2, а поларитет вишечлане лексичке јединице “*act funny*” (делује смешно) је -1. У табели 4.5 приказани су примери одредница у подречнику прилога.

[205] в. одељак 4.1.1.

Табела 4.5 Примери одредница и вредности њихових SO у речнику SO-CAL

Word	SO value
excruciatingly	-5
sleazy	-3
foolishly	-2
satisfactorily	1
purposefully	2
hilariously	4

Подешавања интензитета у речнику могу се вршити проналажењем модификатора сентимент интензитета – појачивача (енг. intensifiers) и ослабљивача (енг. downtoners). Модификатори се у SO-CAL речнику дефинишу процентом појачања или слабљења и неки примери дати су у табели 4.6.

Табела 4.6 Примери модификатора и интензитети њихових модулација SO у SO-CAL речнику

Intensifier	Modifier(%)
slightly	-50
somewhat	-30
pretty	-10
really	+15
very	+25
extraordinarily	+50

Узимајући у обзир примере из табела 6 и 7, колокација „somewhat sleazy“ би у SO-CAL речнику добила SO вредност: −3 × (100% − 30%) = −2.1. Поред придева и прилога, као појачивачи се користе и:

- проналажење речи исписаних само великим словима
- употреба узвичника и већег броја узвичника у низу
- употреба везника „but“ ради проналажења истакнутих информација, итд.

Речник садржи 177 различитих појачивача. Додатна подешавања интензитета у речнику могу се вршити проналажењем негационих модификатора, речи које мењају поларитет осећања оних са којима чине семантичку целину, али сами по себи не носе субјективност (*any, anything, ever, at all*).

Мохамад и Тарни[206] (Mohammad & Turney, 2013) су користили методу групне расподеле рада (енг. crowdsourcing) у поступку креирања лексикона сентименталних речи и израза *Emolex*[207] *(*познат је и под називом *The NRC Word-Emotion Association Lexicon*). Иницијални скуп који представља унију четири подскупа, добијен је комбинацијом три разнородна извора речи и израза претходно обележених ознакама поларитета осећања: (1) *EmoLex-WAL* су чинили литерали преузети из *WordNetAffect*-а, (2) EmoLex-GI су сачињавали појмови добијени из речника *General Inquirer*, (3) *EmoLex-Uni* и *EmoLex-Bi* представљали су подскупове униграма и биграма добијених из *Google* корпуса *n*-грама. Добијени скуп *Emolex* садржао је 10.170 парова појам-сентимент, али је после евалуације овог скупа методом групне расподеле рада, речник сведен на 8.883 одреднице.

Лексикон сентименталних речи и израза (назван још и афективни лексикон) *SentiSense*, заснован је на концептима семантичке мреже *WordNet* v. 2.1 и развијен у току истраживања аутора де Алборноза, Плазе и Герваса[208] (de Albornoz, Plaza & Gervás, 2012). Сам речник, величине 5.496 речи, односно 2.190 синсетова груписаних у 14 категорија које представљају различите врсте осећања, допуњен је софтверским алатима за визуелизацију речника и статистичким методама за анализу дистрибуције емоција. Речник се састоји од два XML документа: *categories.xml* којим се дефинишу саме категорије и *synsets.xml* који садржи одреднице овог речника и које су у корелацији са *WordNet* синсетовима помоћу вредности јединственог идентификатора синсета *SID*. Структура ових категорија дата је у табели 4.7.[209]

[206] Saif M. Mohammad and Peter D. Turney
[207] http://www.saifmohammad.com/WebPages/lexicons.html
[208] Jorge Carrillo de Albornoz and Laura Plaza and Pablo Gervás
[209] табела креирана на основу рада (de Albornoz, Plaza & Gervás, 2012)

Табела 4.7. Структура *SentiSense* лексикона сентименталних речи и израза

Категорије *SentiSense* лексикона *categories.xml*
<SentiSenseEmotionalCategories> <EmotionalCategoryname="joy' antonym="sadness" /> <EmotionalCategoryname="fear' antonym="calmness" /> <EmotionalCategoryname="love' antonym="hate" /> ... </SentiSenseEmotionalCategories>
Одреднице *SentiSense* лексикона - *synset.xml*
<SentiSenseCorpus> <Concept synset="SID-00152712-A" pos="adjective" gloss="lacking cordiality..." emotion="disgust"/> <Concept synset="SID-00050667-R" pos="adverb" gloss="in a joyous manner..." emotion="joy"/> <Concept synset="SID-03430539-N" pos="noun" gloss="a weapon that discharges..." emotion="fear"/> <Concept synset="SID-02571914-V" pos="verb" gloss="come upon or..." emotion="surprise"/> ... </SentiSenseCorpus>

Поред описаних речника, који се често користе у разним истраживањима и задацима обраде природног језика, постоје и бројна друга решења генерисана и употребљена у поступцима анализе осећања: Камбрија са сарадницима је представио ресурс за анализу осећања, *SenticNet*, (Cambria & Hussain, 2012; Cambria, Olsher & Rajagopal, 2014; Cambria & Hussain, 2015;) као јавно доступни лексикон и API намењен за употребу у апликацијама истраживања података на друштвеним мрежама. Лексикони: *The Arguing Lexicon*[210] (Somasundaran, Ruppenhofer & Wiebe, 2007), *The Subjectivity Lexicon*[211] (Wilson, Wiebe & Hoffmann, 2009) и *+/-Effect Lexicon*[212] (Choi & Wiebe, 2014) делови су *The* MPQA *Opinion*[213] корпуса који садржи велики број ручно анотираних новинских чланака, као и читав низ емоционалних категорија и стања (веровања, емоције, осећања, ставови, итд). Речник сентименталних речи и израза *Opinion Lexicon*[214] (познат и као *Sentiment Lexicon*), који садржи око 6800 позитивно и негативно поларисаних речи, генерисан је 2004. (Hu &

[210] http://mpqa.cs.pitt.edu/lexicons/arg_lexicon/
[211] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[212] http://mpqa.cs.pitt.edu/lexicons/effect_lexicon/
[213] http://mpqa.cs.pitt.edu/
[214] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

Liu, 2004a) и укључује информације о морфолошким облицима, могућим грешкама у писању, примере сленга и специјалне симболе који се користе на друштвеним мрежама. Лексикон сентименталних речи и израза Ортонија, Клора и Фоса[215] (Ortony, Clore & Foss, 1987)) *Affective Lexicon* (познат и као OCC *model*) настао 1987. године садржао је око 500 речи које су употребљене да се креира свеобухватна таксономија услова афективности за означавање директних емоционална стања (нпр. туга, срећа) и индиректних референци на та стања унутар одговарајућег контекста (нпр. глагол плакати).

Најзад, један од најстаријих лексикона ове врсте, *General Inquierer*[216], представља базу од око 13.000 речи, семантичких и когнитивних категорија, ручно креирану 1966. године (Stone, 1966). Уз базу, генерисан је и алат за истраживање садржаја текстова који укључује позитивну и негативну конотацију. Структура GI лексикона дата је на слици 4.4.[217]

	Entry	Positiv	Negativ	Hostile ...184 classes ..	Othtags	Defined
1	A				DET ART	...
2	ABANDON		Negativ		SUPV	
3	ABANDONMENT		Negativ		Noun	
4	ABATE		Negativ		SUPV	
5	ABATEMENT				Noun	
...						
35	ABSENT#1		Negativ		Modif	
36	ABSENT#2				SUPV	
...						
11788	ZONE				Noun	

Слика 4.4 Структура лексикона *General Inquierer*

Лексикон сентименталних речи и израза за српски језик развијен је у оквиру рада на овој тези. Одреднице лексикона су речи и колокације са ознакама поларитета осећања. О структури овог дигиталног ресурса биће више речи у одељку 5.4.1.

[215] Andrew Ortony, Gerald Clore and Mark Foss
[216] http://www.wjh.harvard.edu/~inquirer/
[217] слика преузета са http://sentiment.christopherpotts.net/lexicons.html#inquirer

4.5.2 Методе класификације засноване на корпусу

Претпостављања да речи и фразе имају свој подразумевани поларитет осећања (*prior polarity*), независан од контекста, користи се у анализи осећања која се заснива на лексиконима сентименталних речи и израза. Међутим, већ на једноставним примерима можемо уочити важност контекста у коме се подразумевани поларитет неке речи или фразе може укинути или потпуно променити. На пример, у реченици „То је мали, градски ауто“, реч „мали“ има позитиван поларитет, јер је то особина која је пожељна у контексту одабира дате врсте возила, али у реченици „То је мали, неудобан ауто“ поларитет овог придева је негативан. Такође, „лепа хаљина“ не може бити „лепа“, уколико кошта „лепу своту новца“. „Јефтин политички потез“ је непожељан параметар, али „јефтина мобилна телефонија“ јесте, итд. Методе које се баве истраживањем утицаја контекста на анализу осећања у тексту, ослањају се на корпусе из којих се анализом могу откривати лингвистичке везе и законитости које доводе до утврђивања поларитета осећања како делова, тако и читавих докумената. Корпусни приступ у анализи осећања подразумева примену корпуса у изградњи и проширењу лексикона сентименталних речи и израза и у процесима класификације према субјективности и према осећањима. Када је реч о методама изградње и проширења лексикона, према Лиу (Liu, 2012), постоје два приступа:

1) почев од датог, ограниченог скупа речи обележених ознакама поларитета осећања (seed words), помоћу скупа претпостављених правила и семантичких релација, у корпусу се проналазе друге речи или изрази за које се, са одређеним степеном тачности, утврђује њихов поларитет осећања;
2) почев од лексикона сентименталних речи и израза опште намене, применом одређених скупова лингвистичких правила, из доменски оријентисаних корпуса се генеришу специјализовани, такође доменски-оријентисани, лексикони сентименталних речи и израза.

Када се корпуси користе у задацима класификације, опсег контекста може бити једна реченица или низ сукцесивних реченица. Тада се, према Канајама и Насукави[218] (Kanayama & Nasukawa, 2006), ради о :

1) анализи конзистенције осећања унутар једне реченице (*intra-sentential*),
2) анализи конзистенције осећања између суседних реченица (*inter-sentential*).

Изградња и проширење лексикона сентименталних речи и израза употребом корпуса предмет је истраживања у раду (Hatzivassiloglou & McKeown, 1997). Аутори предлажу примену полазног скупа речи обележених ознакама поларитета осећања (seed words) и лингвистичких правила за проналажење оних придева у корпусу којима се исказују осећања. Једно од тих правила је „семантичка конјункција", тј. сукцесиван скуп придева повезаних везником „и", као у следећем примеру: „Тај ауто је леп, поуздан, рентабилан и економичан". Уколико претпоставимо да реч „леп" постоји у полазном скупу обележених речи и носи ознаку позитивног поларитета осећања, онда се применом правила „семантичке конјункције", позитиван део скупа може проширити и осталим придевима који учествују у овој релацији. Хацивасилоглу и Мекеонова[219] истражују и релације између придева које садрже и друге везнике: „али", „или", „или-или", „ни", „ни-ни" и сл.

Важна мера контекстуалне асоцијативности парова речи или фраза која се користи у методама ове класе проблема је PMI мера (*Pointwise mutual information*) која нам каже колико једна реч (*point*)[220] говори о другој, односно која мери количину информација о једној речи када је дата количина информација о другој речи. У том смислу, нека се посматрају две дискретне случајне променљиве X и Y (*points*), такве да се посматрају догађаји њихових појављивања. Ако обележимо са:

$p(X)$ – вероватноћу појављивања прве случајне променљиве X

$p(Y)$ – вероватноћу појављивања друге случајне променљиве Y

[218] Hiroshi Kanayama and Tetsuya Nasukawa
[219] Vasileios Hatzivassiloglou and Kathleen R. McKeown
[220] На основу чега метода носи назив.

p(X, Y) – вероватноћу заједничког појављивања X и Y ,
тада се може увести следећа дефиниција.

Дефиниција 7.1 Ако су x и y исходи дискретних случајних променљивих X и Y, респективно, PMI је

$$PMI(x,y) \equiv log_2 \frac{p(x,y)}{p(x)p(y)} = log_2 \frac{p(x|y)}{p(x)} = log_2 \frac{p(y|x)}{p(y)}$$

мера независности исхода x и y.

PMI, као меру контекстуалне семантичке сличности пара речи, у методе анализе осећања увео је Тарни (Turney, 2001). Његов рад се заснива на истраживањима у раду (Church & Hanks, 1990), где се као мера сличности користи математичко очекивање мере PMI, односно мера међусобне информације MI (mutual information). Уведену претпоставку да реч карактеришу речи којима је окружена, Тарни исказује математички на следећи начин:

$$score(choice_i) = log_2 \frac{p(problem, choice_i)}{p(problem)p(choice_i)}$$

где се тражи максимум ове функције у датом опсегу, како би се нашла она реч $choice_{max}$ која даје највећу вероватноћу заједничког појављивања (енг. co-occurence) уз реч $problem$.

Уколико су $problem$ и $choice_i$ независни, вероватноћа заједничког појављивања је $p(problem, choice_i) \approx p(problem)p(choice_i)$, односно $PMI(problem, choice_i) \approx 0$. У супротном постоји контекстуална семантичка корелација између посматраних $problem$ и $choice_i$.

С обзиром да је log монотоно растућа функција, проблем се може поједноставити разматрањем функције:

$$score(choice_i) = \frac{p(problem, choice_i)}{p(choice_i)}$$

Тарни ову функцију даље модификује увођењем једне врсте претраге веб докумената помоћу машине за претрагу *Altavista*. Ако се са $hits(query)$

означи број докумената који *Altavista* нађе за задати упит *query*, тада се модификовани алгоритам PMI-IR функције може представити као

$$score(choice_i) = \frac{hits(problem\ AND\ choice_i)}{hits(choice_i)}$$

чиме се може проверити колико веб докумената садржи и $problem$ и $choice_i$, а упитом

$$score(choice_i) = \frac{hits(problem\ NEAR\ choice_i)}{hits(choice_i)}$$

колико докумената садржи и $problem$ и $choice_i$ у непосредној близини. Ради евалуације предложене PMI-IR методе, Тарни користи као корпус скупове тестова за проверу знања енглеског језика TOEFL[221] и ESL[222] у процесу проналажења најбољег одговора који представља резултат оцене семантичке сличности задатих речи. Такође, у наставку истог истраживања, аутор показује да PMI-IR на посматраном скупу[223] даје знатно боље резултате од алтернативне методе латентне семантичке анализе (LSA).[224]

Истражујући даље, Тарни и Литман у раду (Turney & Littman, 2003) предлажу методу PMI-IR као оцену поларитета осећања речи тако што се рачунају вероватноће појављивања посматране речи заједно са позитивно поларисаним речима из унапред датог скупа

$$pwords = \{good, nice, excellent, positive, fortunate, correct, superior\}$$

и негативно поларисаним речима из унапред датог скупа

$$nwords = \{bad, nasty, poor, negative, unfortunate, wrong, inferior\}.$$

Тада се семантичка оријентација речи $phrase$ рачуна на основу јачине њених корелација са скупом $pwords$ позитивних и скупом $nwords$ негативних речи:

[221] Test of English as a Foreign Language
[222] English as a Second Language
[223] Треба имати у виду да одабрани скупови за тестирање нису репрезентативни скупови у анализи осећања на енглеском језику.
[224] Опис методе дат је у наставку одељка.

$$SO(phrase) = \sum_{pword \in pwords} PMI(phrase, pword) - \sum_{nword \in nwords} PMI(phrase, nword)$$

Примену методе PMI-IR у процесу класификације докумената Тарни приказује у раду (Turney, 2002) (в. одељак 4.1.1).

Анализирајући Тарнијев и Литманов рад, Мотхарами[225] са сарадницима (Mohtarami, Amiri, Lan, Tran & Tan, 2012; Mohtarami, Lan & Tan, 2013) закључује да постоји проблем са PMI оценом када се користи корпус. Услед ретке појаве или непостојања заједничког појављивања (енг. co-occurence) речи у корпусу са речима из малог почетног скупа сентименталних речи, вредност PMI често може бити занемарљива или једнака нули. Други проблем настаје када реч која учествује у PMI оцењивању садржи семантичко значење у имплицитном облику или, пак, самостално нема поларитет већ једино као део неке фразе. Да би превазишли ове проблеме, аутори предлажу употребу PMI у оцењивању синонима дате речи. Такође, аутори предлажу вишекласификациони PMI модел у коме се интензитет (*Intensity*) речи $word$ за сваку категорију $category$ (аутори су креирали 12 класа које представљају емоционалне категорије) израчунава помоћу:

$$I_k = Intensity(word, category_k) = \sum_{seed_j \in category_k} cooccur\left(word, seed_j\right)$$

а у случају употребе синонима, уместо интензитета дате речи $word$ оцењује се интензитет синсета коме припада:

$$I_k = \sum_{syn_i \in synset(sense(word))} Intensity\left(syn_i, category_k\right)$$

На основу предложеног вишекласификационог PMI модела, аутори предлажу и одговарајућу структуру – „емоционални вектор“. Ако се за сваку реч $word$ израчуна интензитет сваке класе емоција I_k, тада се $word$ представља одговарајућим вектором интензитета $(I_1, I_2, \dots, I_n)$ који се, најзад, у овом истраживању користи у оцени сличности поларитета осећања парова речи

[225] Mitra Mohtarami

израчунавањем коефицијената корелације између емоционалних вектора X и Y двеју речи :

$$corr(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y}$$

где су: n - број класа емоција, $\bar{X}, \bar{Y}$ средње вредности, а S_x , S_y стандардне девијације вектора X и Y.

Векторски и матрични модели са информацијама о поларитету осећања речи могу се наћи и у радовима Маса[226] и његових сарадника (Maas, Daly, Pham, Huang, Ng & Potts, 2011) и Јесеналине и Шарди (Yessenalina & Cardie, 2011). [227]

Када се ради о оцени семантичке сличности унутар хијерархијских структура или текстова који се могу представити у форми таксономије, Ресник[228] уводи меру „информативност садржаја“ (енг. information content - IC) (Resnik, 1995; Resnik, 1999) као оцену којом се може утврдити степен апстракције или специјализације између посматраних концепата[229]. Ако се посматрана таксономија прошири функцијом $p: C \rightarrow [0,1]$ таквом да је за свако $c \in C$, $p(c)$ вероватноћа појављивања инстанце концепта c, тада се „информативност садржаја“ IC дефинише као:

$$IC = -\log(p(c))$$

одакле се може закључити да IC, као степен информативности, опада када је концепт ближе корену стабла, односно када је апстрактнији, с обзиром да важи релација $p(c_1) \leq p(c_2)$ када је c_1 у IS-A релацији са концептом c_2, односно када c_1 припада подстаблу концепта c_2.

Полазећи од претпоставке да сличност два концепта зависи од степена на коме они деле информације, што би се у структури таксономије могло сматрати оним концептом који има највиши ниво, а притом припада

[226] Andrew L. Maas
[227] Ainur Yessenalina and Claire Cardie
[228] Philip Resnik
[229] У таксономији, концепт који је дубље у структури је специфичнији, а онај који који је ближи корену је апстрактнији.

подстаблу и једног и другог концепта, Ресник дефинише семантичку сличност два концепта у таксономији као:

$$sim(c_1, c_2) = \max_{c \in common(c_1, c_2)} [-\log(p(c))]$$

а $common(c_1, c_2)$ је скуп концепата који чине подстабло заједничко за концепте c_1 и c_2.

Када се тражи семантичка сличност међу речима, а не концептима, она се изражава преко семантичке сличности међу концептима свих значења датих речи, помоћу:

$$sim(w_1, w_2) = \max_{c_1, c_2} [sim(c_1, c_2)]$$

Евалуацију свог модела Ресник спроводи над семантичком мрежом PWN, а вероватноће појављивања инстанци концепата, рачуна коришћењем корпуса[230] тако што се за сваку именицу[231] $noun$ у семантичкој мрежи, број њеног појављивања у корпусу рачуна као појављивање сваке класе у хијерархији којој дата именица припада, па је број појављивања једног концепта c (синсета у *WordNet*-у) једнак суми појављивања свих инстанци концепата (синсетова) његовог подстабла $words(c)$:

$$freq(c) = \sum_{noun \in words(c)} count(noun)$$

одакле је вероватноћа појављивања инстанце концепта c дата са:

$$p(c) = \frac{freq(c)}{N}$$

а N је укупан број синсетова именица у *WordNet*-у.

[230] Brown Corpus of American English

[231] У *WordNet*-у се релација *IS-A* остварује над синсетовима именица, што прави хијерархијску структуру.

Менг[232] и сарадници (Meng, Gu & Zhou, 2012) унапређују Ресников рад тиме што сваки концепт посматрају кроз дубину његовог подстабла, број веза хипонимије или подређености које гради и дубину сваке од тих веза.

Још једна мера за оцену семантичке сличности која се користи у алгоритмима за аутоматску изградњу лексикона сентименталних речи и израза и задацима класификације на основу осећања је латентна семантичка анализа. То је метода која претпоставља да се речи (terms) које су семантички блиске (блиске по значењу) налазе у сродним текстовима. Од речи и докумената генерише се матрица речи и докумената (term-document matrix) (в. одељак 1.4) над којом се примењује поступак декомпозиције на сингуларне вредности (Singular Value Decomposition - SVD). Резултат сингуларне декомпозиције (Radunović, 2003) је нови скуп мање димензије сачињен од речи које су међусобно семантички независне, чиме се одстрањују редунданца и вишезначности улазног скупа речи. LSA је примењена у радовима (Strapparava & Mihalcea, 2008), (Ortega Bueno, Fonseca Bruzon, Muniz Cuza, Gutierrez & Montoyo, 2014), (Turney & Littman, 2003), (Turney, 2002), (Mohtarami, Amiri, Lan, Tran & Tan, 2012; Mohtarami, Lan & Tan, 2013), итд.

Осим поларитета, корпусним методама се може оцењивати и јачина осећања. У раду (Johansson & Moschitti, 2013) користе се синтаксно-семантичка правила да би се генерисали предиктори који омогућавају класификацију према јачини осећања, што статистички значајно унапређују постојеће моделе класификације. Према Виби (Hatzivassiloglou & Wiebe, 2000; Wiebe, 2000), градабилност је семантичка особина која једној речи или фрази омогућава учешће у компаративним конструкцијама и сарадњу са модификаторима интензитета. Градабилним придевима изражавају се особине у одређеном (вишем или нижем) интензитету у односу на подразумевану јачину особина израза којег описују (нпр. „грандиозна грађевина“, „претежак задатак“, итд.). Виби је у свом истраживању (Wiebe, 2000) показала да градабилни придеви представљају добар индикатор

[232] Lingling Meng

субјективности текста. Једна од метода класификације придева на градабилне и неградабилне дата је у раду Хаџивасилоглуа и Виби (Hatzivassiloglou & Wiebe, 2000). Аутори су најпре обучавали логистички регресиони модел да класификује придеве у градабилне и неградабилне придеве, а онда су показали да, поред семантичке оријентације, градабиност придева представља поуздан предиктор класификације субјективности текста.

Кад је реч о специјализованим лексиконима сентименталних речи и израза, један од највећих изазова је генерисање свеобухватног лексикона који би помогао боље „разумевање" текстова чији су извори микроблог системи, системи кратких порука и електронске поште. Методе за изградњу лексикона из Твитер порука могу се наћи у радовима Танга са сарадницима (Tang, Wei, Qin, Zhou & Liu, 2014) који предлажу репрезентацију лексикона помоћу фраза до којих се долази алгоритмом машинског учења, Мохамада са сарадницима[233] (Mohammad, Kiritchenko & Zhu, 2013) који креирају NRC *Hashtag Sentiment Lexicon* применом PMI методе између сваке твитер фразе и хештагова (енг. hashtags) познатог поларитета осећања. Броди и Диакопоулос[234] (Brody & Diakopoulos, 2011) се баве утицајем понављања слова у речима као начином изражавања емоција (нпр: „гоооол", „победаааа", и сл.). Ортега и његови сарадници предлажу (Ortega Bueno, Fonseca Bruzon, Muniz Cuza, Gutierrez & Montoyo, 2014) примену методе латентне семантичке анализе у изградњи доменски оријентисаног лексикона над текстовима који представљају оцене артикала са сајта *Ciao*.[235]

Фенг, Бозе и Чои[236] у свом раду (Feng, Bose & Choi, 2011) предлажу изградњу „речника конотација" који се разликује од лексикона сентименталних речи и израза по томе што га чине речи и фразе без подразумеваног поларитета осећања, али које се често налазе уз речи са

[233] Saif M. Mohammad, Svetlana Kiritchenko and Xiaodan Zhu
[234] Samuel Brody, Nicholas Diakopoulos
[235] http://www.ciao.com/
[236] Song Feng, Ritwik Bose and Yejin Choi

израженим негативним поларитетом какве су, на пример, речи: рат, пиштољ, канцер, дијагноза, или позитивним: награда, унапређење, идеја, итд.

Примена корпуса у задацима класификације на основу осећања описана је у раду (Strapparava & Mihalcea, 2008) (в. одељак 4.5.2). Анализа конзистенције осећања се врши унутар реченице помоћу LSA методе – за сваки од наслова утврђује се семантичка сличности текста наслова и посматране класе. Корпус у виду колекције HTML докумената примењен је у анализи осећања на нивоу реченица на јапанском језику у раду (Kaji & Kitsuregawa, 2007). Канајама и Насукава (Kanayama & Nasukawa, 2006), у процесу изградње доменског лексикона, посматрају заједничко појављивање тзв. „поларизованих атома“[237] не само унутар једне већ и у оквиру суседних реченица. Аутори су генерисали низ лингвистичких правила за проналажење „поларизованих атома“ дубљом синтаксном анализом која почива на претпоставци да се исти поларитет осећања обично преноси на следећу реченицу, а да се његова евентуална измена на прелазима између реченица може утврдити помоћу модификатора као што су: *међутим*, *ипак*, *али*, *мада* и др.

4.6 Методе машинског учења у класификацији на основу осећања

Прва примена метода машинског учења у класификацији субјективности реченица дата је у раду (Wiebe, Bruce & O'Hara, 1999). Скуп за учење креиран је из корпуса новинских чланака[238] и садржао је 1004 простих реченица. Реченице су ручно обележила четири оцењивања ознакама класа (субјективно, објективно). Аутори у овом раду предлажу примену статистичког параметра *Cohen's Kappa* у оцени степена сагласности оцењивача како би се евентуална пристрасност могла аутоматски кориговати. Кориговани скуп за учење примењен је у обуци Наивног

[237] Према Канајами и Насукави, поларизовани атом је најмања синтаксна структура која носи поларитет осећања
[238] Wall Street Journal Treebank Corpus

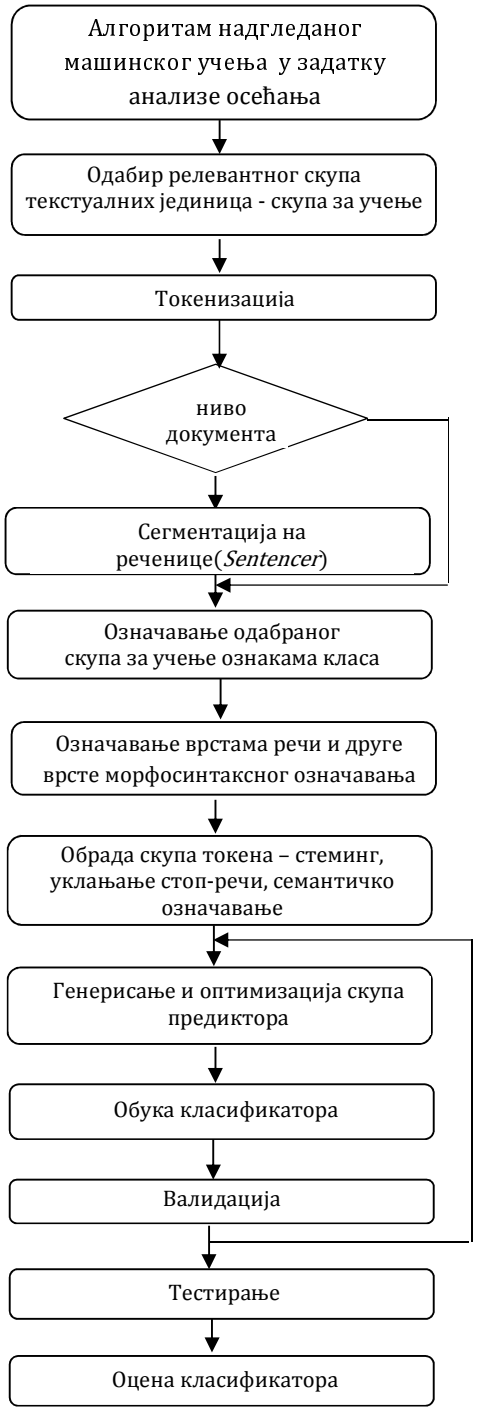
Бајесовог класификатора субјективности, у десетострукој унакрсној валидацији, а коришћени су: бинарни предиктори (индикатори присуства заменица, придева, прилога, бројева и индикатор да ли реченица започиње у новом пасусу) и предиктори који представљају комбинацију класификационе оцене и знака интерпункције. Постигнута је тачност од 72,17% која је додатно побољшана на 81,5% када се употребио подскуп оних реченица скупа за учење који је имао тачније ручно означавање.[239]

Примену метода машинског учења у решавању проблема класификације на основу осећања увели су Пенг и сарадници (Pang, Lee & Vaithyanathan, 2002). Идеја ових аутора да се рецензије филмова велике онлајн базе IMDB[240] употребе као скуп за учење којим ће се обучавати алгоритми надгледаног машинског учења, као и добри почетни резултати класификатора поларитета осећања, мотивисали су истраживаче да и у осталим задацима СА примене низ метода из ове групе и да их интегришу са ресурсима и методама на које су се претходно ослањали. Методе које су Пенг и Ли употребили у свом истраживању су методе машинског учења: Наивни Бајесов класификатор (*NB*), метода максималне ентропије (*ME*) и метода потпорних вектора (*SVM*). Скуп за учење подељен је на две класе: филмове који су позитивно и оне који су негативно оцењени (700 позитивних и 700 негативних рецензија). Класификација на основу осећања рађена је на нивоу докумената, а аутори су експериментисали са различитим врстама предиктора: индикатор присуства униграма, фреквенција појављивања униграма, индикатор присуства биграма, индикатор присуства комбинације униграма и биграма, комбинација униграма и ознаке врсте речи, комбинације униграма и његове релативне позиције у датом тексту и појављивање само придева. Ово почетно истраживање показало је да SVM метода даје нешто боље резултате (82,9%) од ME методе (80,4%) и NB (81%) и то за скуп предиктора који чине униграми дефинисани индикатором присуства.

[239] Од четири оцењивача, двоје су били професионални оцењивачи у области лингвистике.
[240] http://www.imdb.com/

Пенг и Ли су у даљем раду (Pang & Lee, 2004) показали да се надгледано машинско учење може успешно применити у комбинованим задацима класификације. Полазећи од претпоставке да се класификација поларитета на нивоу докумената може побољшати изостављањем реченица које нису субјективне, Пенг и Ли су комбиновали NB и SVM којима се најпре обучава и примењује класификатор субјективности на нивоу реченице, да би класификатор према осећањима користио редукован скуп за учење у коме су остале само субјективне реченице. Тачност класификације према поларитету постигнута оваквом комбинацијом класификатора повећана је на 86,4% NB и 87,15% SVM методом.

Даља истраживања у класификацији према поларитету осећања иду у правцима одабира нових, релевантних скупова предиктора и ка примени ширег скупа техника машинског учења. Ипак, без обзира на усвојену технику, врсту класификационог проблема који се решава и структуру скупа за учење (текстови, твитови, реченице или мање лексичке целине), може се рећи да постоји општи алгоритам класификације према осећањима методама надгледаног машинског учења (слика 4.5) чији је предмет посматрања текст у општем смислу. Упоредне карактеристике и ефикасност алгоритама машинског учења у класификацији текстова према осећањима могу се наћи у свеобухватним прегледним радовима (Yessenov & Misailovic, 2009; Prabowo & Thelwall, 2009; Khairnar & Kinikar, 2013; Habernal, Ptáček & Steinberger, 2013; Medhat, Hassan & Korashy, 2014; Ravi & Ravi, 2015). У наредним одељцима приказаћемо карактеристике и резултате три најзначајније методе надгледаног машинског учења које се примењују у задацима анализе осећања.



Слика 4.5 Општи алгоритам надгледаног машинског учења у задатку анализе осећања

4.6.1 Класификација на основу осећања Бајесовом методом

Наивни Бајесов класификатор (NB) је група вероватносних метода класификације надгледаног машинског учења. У задацима обраде природног језика примењује се у решавању проблема: проналажења информација (Lewis, 1998), класификације текста (McCallum & Nigam, 1998), машинског превођења (Gupta, Nisheeth & Mathur, 2013), разрешавања вишезначности ентитета (Hristea, 2013), оптичког препознавања карактера (Liu & Fujisawa, 2005), анализе осећања (Tan, Cheng, Wang & Xu, 2009; Gamallo & Garcia, 2014) и др. Примена Наивне Бајесове методе у процесу класификације поларитета осећања неког текста уводи две претпоставке:

(1) поларитет текста (класа текста) зависи од поларитета њених посматраних делова;

(2) посматрани делови текста су међусобно независни, односно имају међусобно независан утицај на поларитет текста.

Први услов претпоставља да текст можемо посматрати кроз коначан, унапред одабран, скуп атрибута (предиктора). То су најчешће речи текста (униграми), колокације (биграми) или, у општем случају, вишечлане лексичке јединице (n-грами). Међутим, предиктори могу бити и други квалитативни (ознаке врста речи, знаци интерпункције, и др.) или квантитативни (фреквенције појављивања неког од квалитативних атрибута, релативне позиције квалитативних атрибута у тексту, и др.) предиктори (в. одељак 4.7). Математички посматрано, вероватноћа да текст T који припада колекцији текстова и представљен је скупом атрибута $\{f_1, f_2, \dots f_n\}$ буде класификован у класу $C_j \in \{C_1, C_2, \dots, C_k\}$ представља заједничку вероватноћу два догађаја – да је одабран текст T и да класа текста T има вредност C_j , тј.

$$p(C_j, T) = p(C_j, f_1, f_2, \dots f_n)$$

Ако применимо правило ланаца, важи:

$$p(C_j, T) = p(C_j, f_1, f_2, \dots f_n) = p(C_j)p(f_1, f_2, \dots f_n|C_j)$$
$$= p(C_j)p(f_1|C_j)p(f_2, \dots f_n|C_j, f_1)$$

$$= p(C_j)p(f_1|C_j)p(f_2|C_j, f_1)p(f_3, \dots f_n|C_j, f_1, f_2)$$
$$= p(C_j)p(f_1|C_j)p(f_2|C_j, f_1)p(f_3|C_j, f_1, f_2)p(f_3, \dots f_n|C_j, f_1, f_2) = \cdots$$
$$= p(C_j)p(f_1|C_j) \dots p(f_{n-1}|C_j, f_1, f_2, \dots f_{n-2})p(f_n|C_j, f_1, \dots f_{n-1}) \quad (77)$$

Уз претпоставку о међусобно независном утицају атрибута f_i на класу C_j важиће:

$$p(f_2|C_j, f_1) = p(f_2|C_j)$$
$$p(f_3|C_j, f_1, f_2) = p(f_3|C_j)$$
$$p(f_{n-1}|C_j, f_1, f_2, \dots f_{n-2}) = p(f_{n-1}|C_j)$$
$$p(f_n|C_j, f_1, \dots f_{n-1}) = p(f_n|C_j)$$

па се (77) трансформише у (78):

$$p(C_j, T) = p(C_j, f_1, f_2, \dots f_n) = p(C_j) \prod_{i=1}^{n} p(f_i|C_j) \quad (78)$$

Ако уведемо Бајесову теорему о условној вероватноћи

$p(C_j|T) = \frac{p(C_j)p(T|C_j)}{p(T)}$

и узмемо у обзир да је, на основу условне вероватноће $p(T|C_j) = p(T, C_j)/p(C_j)$,

$p(T|C_j) \propto p(T, C_j)$,

тада Наивни Бајесов модел представља расподелу условне вероватноће случајне променљиве C по свим исходима сваког документа T из посматране колекције, где је вероватноћа да за одабрани текст T класа текста има вредност C_j дата са

$$p(C_j|T) = \frac{p(C_j)p(T|C_j)}{p(T)} \propto \frac{p(C_j)p(T,C_j)}{p(T)} = \frac{1}{Z} p(C_j) \prod_{i=1}^{n} p(f_i|C_j) \quad (79)$$

где $Z = p(T)$.

Процес учења у класификаторима Наивне Бајесове методе састоји се у оцени вероватноће $p(f_i|C_j)$ за сваки предиктор f_i и сваку класу C_j, док процес класификације најпре подразумева да се непознати документ T_x представи помоћу скупа предиктора $\{f_1^{(x)}, f_2^{(x)}, \dots, f_m^{(x)}\} \subseteq \{f_1, f_2, \dots f_n\}$ и одговарајућих условних вероватноћа $p(f_i|C_j)$ добијених моделом (79). Оптимални резултат класификације непознатог документа T_x је она класа $C_{map} \in \{C_1, C_2, \dots, C_k\}$

која има максималну вредност условне вероватноће (*maximum a posteriory probability - MAP*) $p(C_j|T)$, j ∈ {1,2, … , k}

$$C_{map} = \underset{\text{j} \in \{1,2,\dots,k\}}{argmax} \frac{1}{Z} p(C_j) \prod_{i=1}^{m} p\left(f_i^{(x)} \middle| C_j\right)$$

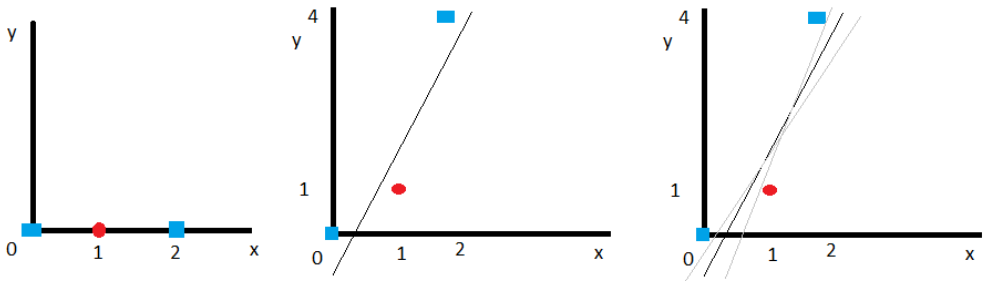
Иако су Пенг и Ли (в. претходни одељак) показали да класификација методом Наивног Бајесовог алгоритма даје нижу тачност од неких других метода, он има своје примене у многим задацима класификације текста. Шта више, изложено је више побољшања ове методе у виду полу-наивног (*Semi-naive Bayesian*) или унапређених верзија наивног класификатора. Међу њима занимљив је приступ дат у раду (Kang, Yoo & Han, 2012) који се бави анализом оцена ресторанских услуга и којим се постиже максимална тачност од 83,6% при класификацији осећања унапређеним NB методама у односу на основну NB методу којом се постиже 73,5% тачности.

4.6.2 Класификација на основу осећања методом потпорних вектора

У досадашњем излагању истакли смо да је метода потпорних вектора бинарна дискриминативна, невероватносна метода надгледаног машинског учења. Налази примену у многим проблемима класификације (в. одељке 4.1.1, 4.4.1, 4.4.2). У многим случајевима, над истим скупом за учење даје боље резултате од већине метода надгледаног ML (в. одељак 4.6). Сваки члан скупа за учење апроксимира се вектором вредности унапред дефинисаног скупа од n атрибута. Геометријски посматрано, сваки члан скупа за учење је тачка у n-димензионалном простору. Процес учења састоји се у проналажењу оне n-димензионалне хипер-равни која на оптималан начин одваја чланове скупа за учење тако да се они чланови који припадају посматраној класи нађу са једне стране, а они који не припадају са друге стране дате хипер-равни. Оптимална хипер-раван је она која максимизује одстојање од чланова скупа за учење, односно којом се постиже да је ширина раздвајања класа максимална. Када скуп за учење није линеарно раздвојив, проблем постаје сложенији и тада се решава у две фазе: напре се трага за тзв. језгреном функцијом $\varphi: R^n \rightarrow R^m$ (енг. kernel function) којом се n-димензионални

векторски простор у коме је скуп за учење линеарно нераздвојив пресликава у m-димензионални векторски простор $(m > n)$ такав да је у њему пресликани скуп за учење линеарно раздвојив, а затим се проналази она m-димензионална хипер-раван која максимизује одстојање од чланова пресликаног скупа за учење.

У следећем, једноставном примеру размотримо улогу језгрене функције у решавању проблема линеарне нераздвојивости скупа за учење. У том смислу, нека је дат једнодимензионални векторски простор и у њему скуп за учење који се састоји од три члана {0,1,2} – од којих два припадају скупу квадрата, а трећи не припада (слика 4.6 а). Скуп квадрата је линеарно нераздвојив у векторском простору R, међутим уколико за језгрену функцију одаберемо $\varphi: R \to R^2$, $\varphi(x) = x^2$, нови скуп $\{(0,0), (1,1), (2,4)\}$ постаје линеарно раздвојив (слика 4.6 б). Из скупа хипотеза, односно раздвајајућих права, бира се она којом се ширина маргине раздвајања класа максимизује (слика 20ц).



Слика 4.6 Примена језгрене функције у решавању проблема линеарне нераздвојивости скупа за учење а) Линеарно неодвојив скуп за учење б) линеарно одвојив скуп за учење - примена језгрене функције ц) права којом се максимизује ширина маргине раздвајања класа

4.6.3 Класификација на основу осећања методом максималне ентропије

Метода *MaxEnt*, приказана у одељку 2.6 примењује се успешно у задацима обраде природног језика. Као метода у задацима анализе осећања

уведена је у раду (Pang, Lee & Vaithyanathan, 2002), а њену примену у овој тези приказаћемо у поглављу 5.

У одељку 2.6 такође смо истакли да се принцип максималне ентропије не бави процесом селекције предиктора. С друге стране, број предиктора у задацима обраде природног језика може нарасти на стотине или хиљаде што даје велику важност методама селекције предиктора.

4.7 Методе селекције предиктора

У поглављу 2 истакли смо да је важан део сваког задатка машинског учења одабир репрезентативног скупа атрибута којим се на оптималан начин апроксимира скуп за учење. Процес селекције релевантних предиктора (Hastie, Tibshirani & Friedman, 2011; Tang, Alelyani & Liu, 2014) састоји се у оцени и одбацивању редундантних, нерелевантних и међусобно зависних атрибута. На пример, уколико је потребно класификовати спортске текстове према теми коју обрађују, тенис би могао бити идентификован на основу појмова из скупа {рекет, сет, гем, бекенд, форхенд,...}, репрезентативни скуп појмова у текстовима који описују кошарку може бити {рекет, тајмаут, обруч, тројка, слободно бацање, лична грешка,...}, скуп атрибута {меч, сусрет, судија, играч, победа, пораз,...} биће заједнички за многе спортске дисциплине, па ће, самим тим, представљати мање релевантан скуп у датом задатку класификације, док ће скуп уобичајених речи у тексту {данас, је, игра, напор, али, људи,...} представљати ирелевантан скуп појмова. Ако претпоставимо да скуп предиктора у задатку машинског учења садржи само n-грамске предикторе, онда можемо очекивати да векторски простор буде велики и да број димензија премаши више стотина. Међутим, у процесу моделирања докумената, врло често се користе и синтаксни, семантички и стилски атрибути (Abbasi, Chen & Salem, 2008) што додатно утиче на увећање броја предиктора. Ипак, неки од атрибута нису релевантни у посматраном процесу класификације и могли би бити елиминисани. Процес селекције атрибута, односно проналажења оног скупа предиктора модела машинског учења који је довољно велики и комплексан

да би побољшао тачност модела и довољно мали да обезбеди ефикасност (израчунљивост у коначном временском интервалу са унапред датим рачунарским ресурсима) и спречи пренаученост модела (overfitting) спада у технике хеуристике.

Дефиниција 8.1 За дати скуп предиктора $F = \{f_1, f_2, \dots, f_n\}$ селекција предиктора $F_s \subset F$ је онај подскуп предиктора h из скупа H свих могућих подскупа предиктора такав да је

$$F_s = argmax_{\ h \in H}\{\Phi(H)\}$$

односно, онај подскуп h којим се максимизује функција Φ оцене алгоритма учења да изврши дати задатак класификације.

Методе којима се смањује димензија векторског простора предиктора (енг. dimension reduction) у процесу машинског учења могу се поделити у две основне групе:

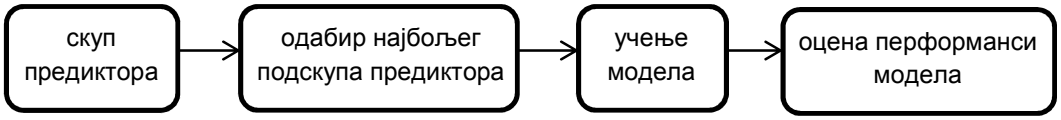
1) методе селекције предиктора (енг. Feature selection - FS methods)
2) методе екстракције предиктора (енг. Feature extraction - FE methods).

Методе селекције предиктора заснивају се на одабиру оптималног подскупа предиктора из скупа предиктора дефинисаних скупом за учење. Циљ ове групе метода је смањење димензије векторског простора предиктора одабиром најоптималнијег подскупа. Методе селекције предиктора могу бити:

1) методе селекције филтрирањем (енг. filters) (Almuallim & Dietterich, 1991)
2) методе селекције помоћу претходног учења (енг. wrappers) (John, Kohavi & Pfleger, 1994)
3) методе селекције уметањем (енг. embedded methods) (Guyon & Elisseeff, 2003)

Методе филтрирања нису зависне од метода учења и обично им претходе у задацима машинског учења (слика 4.7). Филтер користи сопствени алгоритам којим, на основу претпостављене метрике, оцењује допринос или сваког предиктора појединачно, генеришући уређену листу предиктора из

које се бира најоптималнија подлиста, или појединачних подскупова предиктора оцењујући их тако што се над сваким подскупом примени функција метрике и налази онај подскуп који максимизира ту функцију.



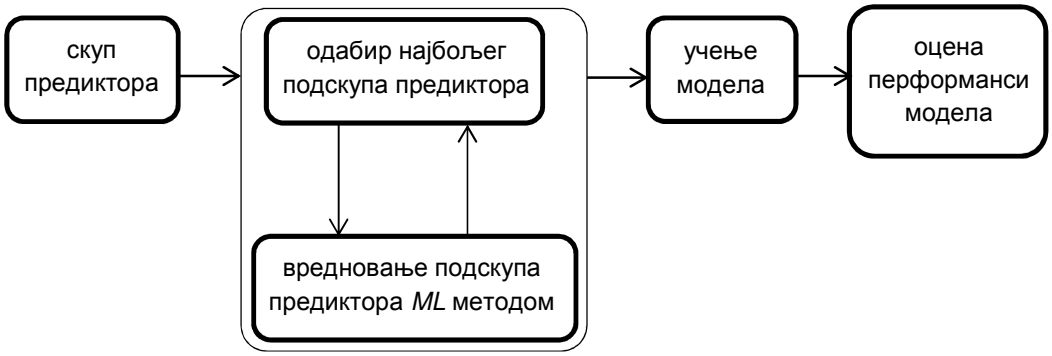
Слика 4.7 Примена селекције предиктора методом филтрирања у задатку машинског учења (претходи фази учења)

Добре стране ове врсте метода селекције су робустност и ефикасност (могу обрадити велике скупове предиктора уз прихватљиве перформансе рачунарске обраде) као и применљивост у различитим системима (због независности од ML метода), а лоше се огледају у ограниченим перформансама коначног модела, с обзиром да филтри не узимају у обзир ефекат одабраног подскупа на процес учења и претпостављају међусобну независност предиктора.

За разлику од филтера, методе селекције уз помоћ тзв. претходног учења узимају у обзир ефекат одабраног подскупа на процес учења, па су веома зависе од алгоритма одабраног у поступку учења модела. Избор подскупа оцењује се помоћу алгоритма машинског учења који садржи технике вредновања предиктора и уско је везан за сам алгоритам учења којим се решава постављени проблем (слика 4.8). Иако има више различитих техника претраге подскупова предиктора као што су: *први најбољи* (енг. best-first), *разгранај па ограничи* (енг. branch-and-bound) и др., најбоље резултате у практичним применама показују похлепне технике које се деле на технике са избором предиктора у сваком кораку (енг. Stepwise Forward Selection) и технике са елиминацијом предиктора у сваком кораку (енг. Stepwise Backward Elimination).

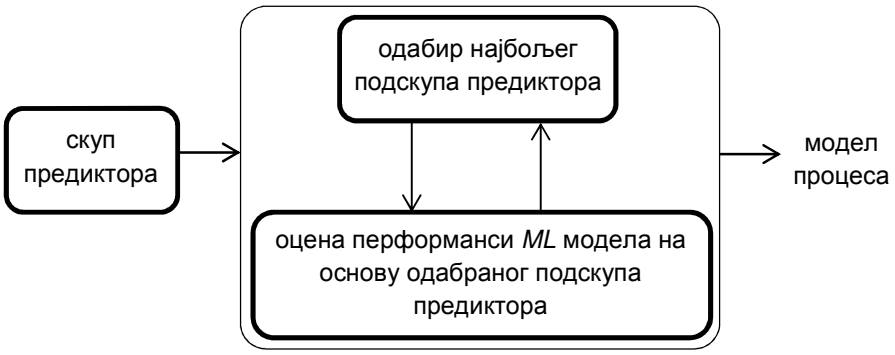
Похлепне технике са избором предиктора у сваком кораку, полазећи од празног скупа, додају по један предиктор у сваком итеративном кораку и то онај који даје највећи допринос подскупу предиктора при оцени

перформанси алгоритма за учење у посматраном кораку селекције. Поступак се прекида када процес додавања предиктора почне да смањује перформансе алгоритма за учење. Друга група аналогних похлепних техника полази од потпуног скупа предиктора, уклања по један у сваком итеративном кораку и то онај који даје највеће смањење перформанси алгортма за учење у том кораку.



Слика 4.8 Примена методе селекције помоћу претходног учења у задатку машинског учења

Добре стране ове врсте метода селекције су што утичу на повећање тачности алгоритма машинског учења, нарочито у задацима обраде доменских знања и могу да узму у обзир и међусобне корелације предиктора. Недостаци се огледају у мањој ефикасности од филтера и немогућности директне сарадње са различитим ML алгоритмима. Методе уметања подразумевају да се процес селекције предиктора врши у току саме обуке модела (слика 4.9). Приликом сваке обуке оцењује се разлика у перформансама модела над тренутно актуелним скупом предиктора у односу на скуп са додатним (или уклоњеним) предиктором. Из тог разлога, ове методе су високо зависне од одабраног алгоритма учења. Ипак, методе уметања могу бити знатно брже од метода селекције помоћу претходног учења, јер не врше обуку сваког подскупа предиктора из почетка.



Слика 4.9 Примена методе селекције предиктора техником уметања у задатку машинског учења

Осим метода селекције, у процесу одабира оптималног скупа предиктора у задацима машинског учења користе се и методе генерисања новог скупа предиктора ниже димензије у односу на полазни које се често називају и методама екстракције. Хибридне методе користе добре стране обе врсте метода. Све методе које се баве неким начином проналажења оптималног скупа предиктора једним именом се зову методе редукције векторског простора. О екстракцији и хибридизацији предиктора биће више бити речи у наредном поглављу.

5. Класификација текстова на основу осећања

5.1 Екстракција векторског простора предиктора

У задацима класификације текста и задацима анализе осећања у тексту, могу се применити различите методе екстракције векторског простора предиктора. Екстракција предиктора је процес генерисања новог скупа из постојећег, са особинама које обезбеђују већу информативност и већи степен дискриминације међу класама. Најчешће коришћене методе екстракције у задацима класификације текста су: груписање предиктора (*clustering*), вишедимензионално скалирање (*Multidimensional scaling* - MDS), анализа главних компоненти (*Principal Component Analysis* - PCA) линеарна дискриминантна анализа (*Linear Discriminant Analysis* - LDA), генетски и еволутивни алгоритми и регресиона анализа. Методе груписања (у литератури се још сређе појам кластеровања) заснивају се на груписању „сличних“ предиктора чија се сличност мери неком од мера за оцену сличности (Еуклидско, косинусно, Махаланобисово растојање, корелација и др.) и заменом такве групе центроидом (просеком свих предиктора у групи) који постаје нови предиктор. Метода вишедимензионог скалирања – MDS коју је предложио Крускал (Kruskal, 1964) је техника која пројектује оригинални скуп за учење у нови скуп у простору ниже димензије, тако да су растојања тачака у оригиналном простору сачувана у највећој могућој мери. Анализа главних компоненти – PCA (Jolliffe, 2002) је метода редукције димензије векторског простора којом се налази она линеарна пројекција тачака полазног високодимензионалног простора у простор ниже димензије којим се максимизује варијанса полазног скупа. Метода линеарне дискриминантне анализе – LDA (Duda, Hart & Stork, 2000) налази ону линеарну трансформацију која максимизује одвојивост класа у векторском простору предиктора са мање димензија, када су познате ознаке класа у оригиналном простору. У задацима анализе осећања у раду (Vinodhini & Chandrasekaran, 2013) приликом екстракције предиктора коришћена је метода анализе главних компоненти, а линеарна дискриминантна анализа у

раду (Wang, Li, Wei & Li, 2009). Методе груписања заснивају се на правилима генерисаним помоћу: семантичких мрежа – као у радовима (Liu, Hu & Cheng, 2005; Carenini, Ng & Zwart, 2005; Zhang & Li 2010), морфолошких правила – као у радовима (Zhang, Xu & Wan, 2012; Wang & Wang, 2008; Gamon, 2004), правила специјализације (Riloff, Patwardhan & Wiebe, 2006) или генерализације чланова хијерархијских структура (Zhao, Li & Wang, 2013). Осим груписања, редукција се може спровести и дефинисањем функције пресликавања одређених група предиктора са сличним семантичким особинама у скуп репрезената тих група (Gaudette & Japkowicz, 2011).

Након генерисања новог скупа предиктора, потребно је извршити његову оцену (редундантност, репрезентативност, релевантности и међусобну зависност). Квалитет предиктора можемо оцењивати статистичким параметрима као што је стандардна девијација (Yousefpour, Ibrahim, Nuzly & Hamed, 2014), вероватносним мерама као што су веродостојност (Gamon, 2004), максимална веродостојност (MLE) (Zhai, Liu, Xu & Jia, 2010) или дистрибуција вероватноће (Zhang, Xu & Wan, 2012), мерама које оцењују семантичку сличност, нпр. PMI (*point-wise mutual information*) (Popescu, Yates & Etzioni, 2004) и др. Свеобухватна анализа метода редукције векторског простора предиктора у задацима анализе осећања може се наћи у радовима (Jiliang, Salem & Huan, 2014) и (Asghar, Khan, Shakeel & Kundi, 2014).

5.2 Хибридне методе екстракције предиктора

Идеја хибридизације метода и скупова предиктора није нова. Хибридизација метода се може постићи комбиновањем FS и FE метода. На пример, Барти и Синг[241] (Bharti & Singh, 2015) предлажу примену FS методе за одабир подскупа најзначајнијих предиктора над којом се затим примењује FE метода, односно генерише се нови скуп са још мањим бројем предиктора. Друга врста хибридизације подразумева интеграцију семантичких ресурса у

[241] Kusum Kumari Bharti and Pramod Kumar Singh

поступцима генерисања нових предиктора. На пример, Алгрин[242] и сарадници (Ahlgren, Malo, Sinha, Korhonen & Wallenius, 2012), ослањајући се на бројна претходна истраживања, користе Википедију као ресурс којим се појмови из текста линковима повезују у одговарајуће концепте. Таква мрежа концепата користи се да се предикторима скупа за учење додели семантичка информативност након чега се може применити метода редукције којом се семантички концепти групишу тако да дају најбољу одвојивост скупа за учење.

У радовима (Liu, Hu & Cheng, 2005; Carenini, Ng & Zwart, 2005; Moravec, Kolovrat & Snásel, 2004) као средство у поступку груписања користи се семантичка мрежа ворднет – праћењем хијерархије релације *hypernym* (Moravec, Kolovrat & Snásel, 2004) или имплицитне релације *synonym* (Liu, Hu & Cheng, 2005; Carenini, Ng & Zwart, 2005).

Најзад, идеја да се различите појаве семантички повезаних предиктора успешно могу заменити јединственим представником у самом скупу за учење применили су Лиу и сарадници (Liu, Hu & Cheng, 2005). Ако се, на пример, као објекат субјективног оцењивања узме *фото-апарат*, а одговарајући скуп за учење садржи следеће реченице:

„Слика је добра, живот батерије дуг".

„Квалитет слике добар, батерија траје кратко".

Ли и сарадници предлажу да се у свим триграмима, унапред познати атрибути фото-апарата замене фиксном речи *[feature]*, након чега би скуп за учење изгледао:

„*[feature]* је добра, живот *[feature]* дуг".

„Квалитет *[feature]* добар, *[feature]* траје кратко"

чиме би се скуп триграмских предиктора редуковао.

[242] Oskar Ahlgren

5.3 Метода редукције векторског простора хибридизацијом предиктора

У овој тези предложена је нова метода редукције векторског простора (Mladenović, Mitrović, Krstev & Vitas, 2015) која представља комбинацију метода груписања и метода које користе функцију пресликавања датог скупа предиктора у нови, редуковани скуп. Груписање се не одвија над комплетним улазним скупом предиктора, већ над једним делом и том приликом се генеришу два подскупа предиктора. Над тим подскуповима се затим примењују функције пресликавања којима се сваки од подскупова замени са два нова предиктора. На тај начин се смањује почетни векторски простор предиктора. С обзиром да они предиктори улазног скупа који нису груписани остају исти и у новом векторском простору, ова метода спада у групу хибридних, јер је резултујући простор сачињен од комбинације старих и нових предиктора.

Метода је заснована на две разнородне врсте претходних истраживања. Једна од њих се односи на методе пресликавања предиктора, а друга на методе екстракције флективних и деривационих форми основног облика речи (леме) из електронских морфолошких речника. Иако инспирацију у нашем истраживању чине рад Лиуа и сарадника (Liu, Hu & Cheng, 2005), као и истраживања утицаја ворднета на процес редукције векторског простора објављена у радовима (Liu, Hu & Cheng, 2005; Carenini, Ng & Zwart, 2005; Moravec, Kolovrat & Snásel, 2004) и описана у претходном одељку, метода пресликавања предиктора коју предлажемо заснована је на резултатима радовима Саифа[243] (Saif, He & Alani, 2012) и Виби[244] (Wiebe et al., 2004) и њихових сарадника.

Саиф је са сарадницима предложио методу којом се креира скуп нових предиктора на основу значења оних речи или фраза у датом тексту које представљају одређене концепте (нпр. *Steve Jobs* представља се предиктором *person*, а *London* предиктором *city,* итд.). Скуп нових предиктора, заснован на

[243] Hassan Saif
[244] Janyce Wiebe

концептуалној репрезентацији ентитета у тексту, анотираних помоћу три алата: *AlchemyAPI*, *Zemanta* и *OpenCalais*, искоришћен је у обуци наивног Бајесовог класификатора осећања у твитовима. Саиф је експериментисао са скупом нових предиктора на три начина:

- у првом случају је извршена замена предиктора у скупу за учење одговарајућим семантичким репрезентацијама, што је довело до редукције векторског простора;
- у другом случају је оригинални векторски простор проширен скупом нових предиктора, додавањем и семантичких репрезентација као нових предиктора;
- у трећем је извршена интерполација основних предиктора са новим предикторима.

Резултати истраживања Саифа и његових сарадника показали су да скуп предиктора, генерисан на основу анотације речи и фраза у тексту семантичким концептима, унапређује основну методу класификације: Најбољи резултат добијен је скупом интерполираних предиктора (тачност 75.95%) у односу на увећани скуп који је садржао основне и нове предикторе (тачност 71.33%) и скуп који је редукован новим предикторима (тачност 68.90%).

За разлику од предложене методе Саифа, у овој тези нови скуп предиктора генеришемо на основу информација о поларитету осећања речи и фраза које улазе у процес генерисања предиктора. Информације екстрахујемо из два семантичка ресурса: лексикона сентименталних речи и израза и Српског ворднета. Даље, за разлику од Саифа који користи само унограме као предикторе, ми предлажемо и користимо и биграме и триграме, узимајући у обзир резултате истраживања Виби и њених сарадника (Wiebe et al., 2004; Wiebe, Wilson & Bell, 2001). Резултујући скуп предиктора користимо у обуци *MaxEnt* класификатора на основу осећања на нивоу документа у текстовима на српском језику. Начин интеграције новог скупа предиктора у скуп за учење идентичан је првом начину (из разлога што смо желели директну редукцију векторског простора) који је у свом раду (Saif, He & Alani, 2012) описао Саиф и односи се на супституцију свих

груписаних предиктора новим, након чега је димензија новог векторског простора V_r смањена и може се представити као:

$$|V_r| = |V| - \sum_{g \in Groups} |V_g| + 4. \quad (80)$$

где је $|V|$ димензија оригиналног векторског простора предиктора, која се умањује за збир груписаних предиктора, а затим увећава за укупан број супституција свих група (максимално четири). Математички модел предложеног модела редукције биће детаљно приказан у наредном одељку.

У раду (Wiebe et al., 2004), Виби са сарадницима уводи тзв „јединствене генерализоване *n*-граме" тј. *ugen-n-gram* предикторе. Истражујући потенцијално субјективне елементе (речи и фразе које текст чине субјективним, односно којима се исказују осећања), аутори су показали да метода примењена у идентификацији ових елемената има већу прецизност када се у алгоритам укључе *ugen-n-gram* предиктори. У том раду се, у скупу за учење, све оне речи које се појављују само једном у неком документу (*hapax legomena*) замењују јединственом речи *UNIQUE* (скраћено *U*) уз додавање ознаке врсте речи која је замењена. На тај начин генеришу се *ugen-1-gram* предиктори који су облика $U - POS$ и који представљају све речи типа *hapax legomena*. Пошто се реч *UNIQUE* сада налази у текстовима скупа за учење, она учествује и у формирању колокација од којих се могу генерисати *unique generalized 2-gram* (*ugen-2-gram*) предиктори облика:

U-adverb U-verb (пример таквог *ugen-2-gram* предиктора је U- *fascinate* генерисан на основу колокације *unceasingly fascinate* – уколико претпоставимо да је прилог (*adverb*) *unceasingly* у посматраном документу скупа за учење употребљен само једном);

U-adj as-prep (извори таквих *ugen-2-gram* предиктора су колокације *drastic as*, *predatory as*, итд.)

На исти начин генерисани су *ugen-3-gram* и *ugen-4-gram* предиктори (предиктори дужине 3 и 4 речи у којима учествују *hapax legomena*).

Полазећи од овог истраживања, претпоставили смо да у задатку класификације није потребно имати све облике речи као предикторе већ само информацију о њиховој субјективности (јесу ли или нису субјективно

информативни) и поларитету осећања (позитиван или негативан поларитет) који носе, па је стога довољно да све *n*-грамске предикторе који носе информацију о субјективности заменимо идентичном ознаком, додајући јој и ознаку поларитета (*POS* или *NEG*). Такође, да бисмо у истраживањима могли да оцењујемо појединачне утицаје семантичких ресурса које користимо, одлучили смо да ознака којом замењујемо препознату субјективну реч или фразу, у тексту скупа за учење, указује на семантички ресурс на основу кога је препозната. То значи да су субјективне речи и фразе препознате Српским ворднетом замењене ознаком *WordNet*, а оне које је препознао лексикон сентименталних речи и израза – ознаком *Lexicon*. Најзад, да бисмо сачували информацију о поларитету, ознаке добијају наставак *POS* или *NEG*, па је укупан скуп нових ознака, а самим тим и нових предиктора у нашем истраживању био

$$\{WordNetPOS, WordNetNEG, LexiconPOS, LexiconNEG\} \quad (81).$$

Друга врста истраживања на којој је заснована метода редукције коју предлажемо односи се на електронске морфолошке речнике srpskog jezika (SrpMD) простих речи (DELAF) и вишечланих лексичких јединица (DELACF) (в. одељак 3.2). У радовима Крстев (Krstev, 2008), Крстев и Витас (Krstev & Vitas, 2009), Витас и Крстев (Vitas & Krstev, 2013) изложене су методе за ефикасан, аутоматски развој електронских морфолошких речника засноване на теорији коначних аутомата (Gross, 1988). Употребом овог ресурса могуће је генерисати све флективне форме простих речи и вишечланих лексичких јединица. У нашем истраживању електронски морфолошки речници су употребљени за проширење лексикона сентименталних речи и израза (в. одељак 5.4.1), лексикона генерисаног из Српског ворднета (в. одељак 5.4.2) и листе „стоп речи“ (в. одељак 5.4.3). О начину имплементације SrpMD у поступку редукције предиктора биће више речи у наредним одељцима.

5.3.1 Математички модел редукције векторског простора хибридизацијом предиктора

Нека имамо скуп за учење од m докумената $D = \{d_1, d_2, \dots, d_m\}$ представљен помоћу скупа F од k предиктора који представљају n-граме речи, где је $n \leq 3$

$$F = \{f_1, f_2, \dots, f_k\},$$

и нека су

$$F' = \{f_1', f_2', \dots, f_q'\}$$

$$F'' = \{f_1'', f_2'', \dots, f_p''\}$$

дисјунктни скупови n-грама речи који представљају елементе скупова лексикона сентименталних речи и израза и лексикона ворднета, респективно. Ако дефинишемо релације μ и η

$$\mu(f) = F \cap F' = \{f \mid f \in F \wedge f \in F'\}$$

$$\eta(f) = F \cap F'' = \{f \mid f \in F \wedge f \in F''\}$$

тада подскупови $F_s' = F \cap F'$ и $F_s'' = F \cap F''$ дефинишу оне предикторе $f \in F$ који се могу груписати и заменити новим предикторима, чиме бисмо за редуковани скуп предиктора добили

$$|F_r| = |F| - (|F_s'| + |F_s''|). \tag{82}$$

Нека је, даље, $H = \{h_1, h_2, \dots, h_c\}$ скуп ознака класа којима је обележен скуп за учење и скуп G нових предиктора дефинисаних помоћу (81). У нашем случају $c = 2$, POS и NEG, али како имамо два дисјунктна скупа F' и F'', то ће бити различито обележене њихове POS и NEG вредности, односно имаћемо и дисјунктне скупове обележја:

$H' = \{h_1', h_2'\} \subset H$ и $H'' = \{h_1'', h_2''\} \subset H$, такве да $H' \cup H'' = H$, одакле је

$$|H| = 4. \tag{83}$$

Нека је $g \subset H \times G$ функционална релација дата са

$g(h) = \{(h_1', LEXICONPOS), (h_2', LEXICONNEG), (h_1'', WORDNETPOS), (h_2'', WORDNETNEG)\}$

Нека су, затим, дате функционалне релације ρ' и ρ'':

$\rho' \subseteq F' \times H'$, $\rho'' \subseteq F'' \times H''$ чији су домени:

$$Dom(\rho') \stackrel{\text{def}}{=} \{f' \in F' \mid (\exists h \in H')(f', h) \in \rho'\}$$

$$Dom(\rho'') \stackrel{\text{def}}{=} \{f'' \in F'' \mid (\exists h \in H'')(f'', h) \in \rho''\}$$

Тада се, композицијом пресликавања $g \circ (\mu(f) \circ \rho'(f)) \subseteq F \times G$ дефинисаном са

$$\mu(f) \circ \rho'(f,h) \circ g(h,g) = \{g \in G' \mid f \in F \wedge f \in F' \wedge (f,h) \in \rho' \wedge (h,g) \in g\}$$

сваки предиктор f оригиналног скупа F, такав да је истовремено и елемент скупа F', пресликава у тачно један елемент $g \in G'$ новог скупа предиктора $G' \subset F_r$.

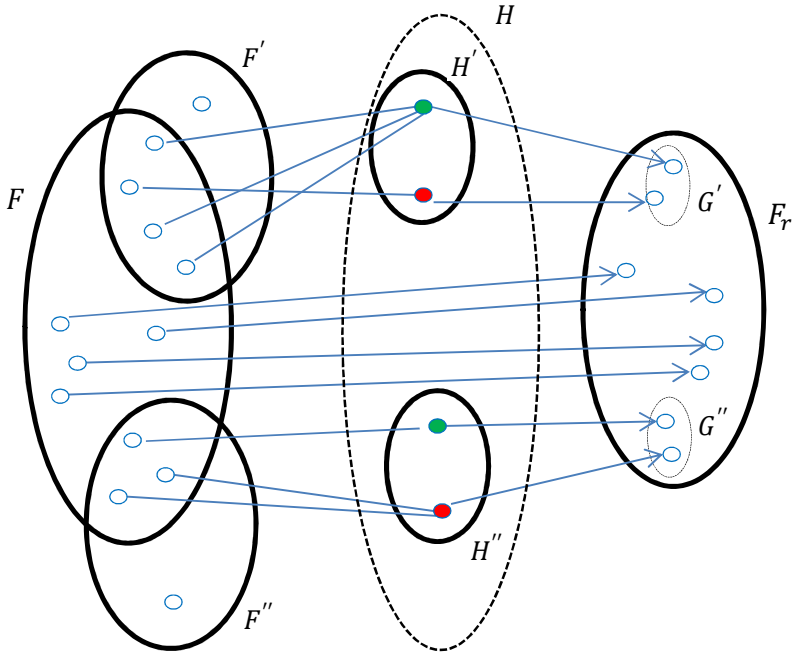
На исти начин је

$$\eta(f) \circ \rho''(f,h) \circ g(h,g) = \{g \in G'' \mid f \in F \wedge f \in F'' \wedge (f,h) \in \rho'' \wedge (h,g) \in g\}$$

На основу једначина (82) и (83), коначни, редуковани скуп предиктора, еквивалентан (80), је димензије:

$$|F_r| = |F| - (|F_s'| + |F_s''|) + |H|.$$

На слици 5.1 представљени су: почетни скуп предиктора F, груписање предиктора на основу истовремене припадности скуповима F и F' или F и F'', резултати пресликавања груписаних предиктора у нове $g \in G'$ и $g \in G''$, на основу релације са скуповима ознака класа H' и H''.



Слика 5.1 Венов дијаграм редукције векторског простора методом хибридизације предиктора

У наредним одељцима описаћемо процес изградње скупова лексикона сентименталних речи и израза F′ и лексикона ворднета F′′ и предложити методу за њихово проширење флективним облицима, како би скупови F ∩ F′и F ∩ F′′ били што већи, па самим тим и редукција броја предиктора боља. Такође, изложићемо алгоритам редукције, пресликавањем у скупове G′ и G′′ оних *n*-грамских предиктора који носе подразумевани поларитет осећања (енг. prior polarity).

5.3.2 Генерисање флективних облика употребом електронских морфолошких речника

У одељку 3.2 истакли смо да морфолошки речник *DELAF* садржи одреднице дате форматом:

inflection,lemma.K_n*[+SynSem]*: [codes_of_morphosyntactic_values]**

где су: *inflection* – један флективни облик речи чији је основни облик дат ознаком *lemma*, K_n – врста речи (изражено великим латиничним словом где је: *N*-именица, *V*-глагол, *A*-придев, *ADV*-прилог, *PRO*-заменица, *NUM*-број, PREP-предлог, *CONJ*-везник и др.) праћена кôдом који указује на коначни трансдуктор којим се једнозначно описују особине флективне класе којој припада дата лема W_{lex}. *SynSem* је низ синтаксно-семантичких ознака које описују лему, наводе се произвољним редоследом у низу, а одвајају знаком „плус“ (+). Њима се описују: подкатегорија речи којој лема припада, значење леме, деривациона информација, синтаксна информација, информација о дијалекту, информација о оригиналном преводу страних речи, информација о негацији речи уколико постоји. Додатком *codes_of_morphosyntactic_values* дефинишу се морфосинтаксне информације о датој флективној форми. Оне се такође наводе произвољним редоследом без међусобног раздвајања. На пример, у случају када се ради о именици (вредност *POS* ознаке је N), облик речи зависи од рода, броја и падежа, па се та информација обележава морфосинтаксном информацијом која садржи: број (*s*-једнина, *m*-множина, *w*-паукал), род (*f*-женски, *m*-мушки, *n*-средњи) и падеж (означен цифром од 1 до

7). Рецимо, ознака N:fs2 односи се на флективни облик именице у генитиву једнине женског рода.

Осим наведених ознака, у *DELAF* речнику се морфосинтаксним кодовима могу описати живи/неживи концепти, степени градације придева, глаголска времена, итд. Следећи примери представљају одреднице речника *DELAF*:

učiteljicu,učiteljica.N+Hum+GM:fs4v флективни облик именице (N) живог бића (v) женског рода (f), једнине (s), у акузативу (4), чија је лема *učiteljica* и која означава концепт људског бића (Hum) и именицу која је изведена из одговарајуће именице мушког рода (GM).

bavio,baviti.V+Imperf+It+Ref:Gsm флективни облик глагола (V) који означава глаголски облик имперфекат (Imperf), непреносни (It), повратни (Ref) у мушком лицу (m) једнине (s).

Слично речнику *DELAF,* дефинише се речних вишечланих лексичких јединица *DELACF* форматом:

inflection,lemma.POS+C|Comp[+SynSem] : [codes_of_morphosyntactic_values]**

који се од *DELAF* формата разликује у додатној ознаци *+C* или *+Comp* чиме се исказује сложена лексичка целина (тзв. *compound*). Пример одреднице речника *DELACF* је

*godišnjem dobu,godišnje doba.N+*Comp*:ns3q*

где је *godišnjem dobu* флективни облик вишечлане лексичке јединице *godišnje doba*, именице средњег рода једнине у дативу. Узимајућу у обзир пример леме *ljubav*, процедура упита над *MEDS* као резултат даје 17 флективних облика, односно 4 различита облика речи:

ljubavima,ljubav.N:fp3:fd6:fp7
ljubavlju,ljubav.N:fs6
ljubavi,ljubav.N:fs2:fs3:fs5:fs6:fs7:fp1:fp2:fp4:fp5
ljubav,ljubav.N:fs1:fs4.

а упит који тражи вишечлане лексичке јединице које садрже лему *ljubav,* између осталих даје и свих 9 различитих флективних облика фразе *ljubavna veza*:

ljubavna veza,ljubavna veza.N:fs1

ljubavne veze,ljubavna veza.N:fs2:fp1:fp4:fp5

ljubavnoj vezi,ljubavna veza.N:fs3:fs7

ljubavnu vezu,ljubavna veza.N:fs4

ljubavna vezo,ljubavna veza.N:fs5

ljubavnom vezom,ljubavna veza.N:fs6

ljubavnih veza,ljubavna veza.N:fp2

ljubavnim vezama,ljubavna veza.N:fp3:fs6:fs7

ljubavnima vezama,ljubavna veza.N:fp3:fs6:fs7.

Лексикон сентименталних речи и израза српског језика има структуру облика:

одредница поларитет_осећања.

Одредница је основни облик просте речи (лема) или вишечлане лексичке јединице. *Поларитет_осећања* означава да ли је осећање које се исказује одредницом позитивно или негативно. О изградњи овог лексикона биће више речи у одељку 5.4.1. Примери јединица овог лексикона су:

ljubavna veza POS

duševna patnja NEG

nahuditi NEG

odvažan POS

osećanje zadovoljstva POS.

Проширење овог лексикона свим флективним облицима одредница у лексикону извршено је помоћу електронских морфолошких речника. Алгоритам проширења речника (алгоритам 5.1) дат је у наставку. Улазну структуру алгоритма чини лексикон основних речи и израза, а излазну проширени лексикон који поред основних речи и израза садржи и све њихове флективне облике. С обзиром да се значење не мења променом флективног облика, сви флективни облици једне одреднице наслеђују исту ознаку *поларитет_осећања*.

Алгоритам 5.1 Проширење лексикона сентименталних речи и израза флективним формама

```
Input: Sentiment_Lexicon_X={x_1, x_2, ..., x_n} // x_i is lemma
Output: Sentiment_Lexicon_E={e_ij}, i∈{1,2,...,m}, j∈{1,2,...,k} // Inflectionally Expanded
// Sentiment Lexicon where e_ij is lemma or MWU or their inflection
// m is the total number of lemmas and MWU units in the input Lexicon,
// k is the total number of inflectional forms of the current lemma or MWU unit
 1. for each lemma in Sentiment_Lexicon_X
 2.    inflections = findInflectionsInDELAF(lemma)
 3.    for each inflection in inflections //iterate simple forms
 4.       addToSentiment_Lexicon_E(inflection, lemma.Polarity);
       // find all MWUs containing substring equal to lemma
 5.    mwus= findAllMWUsInDELACFcontain(lemma)
 6.       for each mwu in mwus //iterate mwus and find all inflections of each mwu
 7.          inflections = findInflectionsInDELACF(mwu)
 8.          for each inflection in inflections // iterate inflections of mwu
 9.             addToSentiment_Lexicon_E(inflection, lemma.Polarity);
10. return Sentiment_Lexicon_E
```

Димензија лексикона (означен са *Sentiment_Lexicon_X*) пре проширења је n и он садржи само основне облике речи (леме). Након проширења, лексикон (означен са *Sentiment_Lexicon_E*) садржи m одредница, где је $m > n$, а лексикон садржи и све флективне облике свих лема (применом процедуре *findInflectionsInDELAF* над сваком појединачном лемом полазног речника), све вишечлане лексичке јединице у којима се појављују све леме (применом процедуре *findAllMWUsInDELACFcontain* над сваком појединачном лемом полазног речника) и све њихове флективне облике (применом процедуре *findInflectionsInDELACF* над сваком вишечланом лексичком јединицом).

Проширење лексикона размотримо на следећим примерима. Нека посматрамо одредницу лексикона која представља лему *ljubav* и која је означена позитивним поларитетом осећања, тј. ознаком POS. Процедура *findInflectionsInDELAF* као резултат даје 15 флективних облика, односно 4 различита облика речи: *ljubav, ljubavi, ljubavlju, ljubavima*. Како флективне форме носе исти поларитет осећања као и лема, речник је проширен са три одреднице:

ljubavi POS

ljubavlju POS

ljubavima POS.

Затим су, применом процедуре *findAllMWUsInDELACFcontain*(lemma), у SrpMD нађене све вишечлане лексичке јединице у којима постоји реч *ljubav*:

ljubavna priča, ljubavne reči, ljubav na prvi pogled и др. Применом процедуре *findInflectionsInDELACF*(mwu), за сваку од вишечланих лексичких јединица пронађени су сви флективни облици којима је проширен лексикон, при чему су све одреднице наследиле исту ознаку *поларитет_осећања* речи *ljubav*.

На исти начин проширење флективним облицима извршено је и код друга два ресурса која користимо у овом истраживању: лексикона генерисаног из Српског ворднета и листе „стоп речи".

Други важан разлог проширења описаних лексикона флективним облицима лежи у природи задатка који решавамо и који је јако осетљив на морфолошке промене над предикторима, што ћемо објаснити у наставку.

У одељку 4.6 представљен је општи алгоритам анализе осећања у тексту. Попут осталих задатака класификације текстова и SA садржи методе које представљају подразумеване кораке као што су: подела текста на реченице, токенизација, стеминг, означавање врстама речи. Међутим, у језицима са развијеном морфологијом (*Morphologically Rich Languages* - MRL), где семантички значајне врсте речи (именице, глаголи, придеви) могу имати велики број флективних облика, уобичајене процедуре кореновања, означавања врстама речи и парсирања морају бити прилагођаване природи таквих језика (Tsarfaty et al., 2010). Стеминг или кореновање, као процес свођења флективног или деривационог облика на основу (корен), представља један од начина за редукцију векторског простора предиктора класификатора. Међутим, употреба стемера у задацима класификације осећања у текстовима на српском језику може произвести два нежељена ефекта. Први се тиче могућности да деривације једног корена које настају додавањем афикса имају различит поларитет осећања у односу на корен. То значи да би се њиховим свођењем на корен изгубила информација о њиховом поларитету. На пример, речи: *мотивисан*, *немотивисан*, *демотивисан*, *безмотиван* имају исти корен *мотив*, али префикси *не-*, *де-*, *без-* дају негативан поларитет осећања речи *мотивисан* чији је поларитет позитиван. Ако желимо да овај проблем превазиђемо дефинисањем правила којим ће деривације добијене поменутим префиксима имати поларитет осећања супротан поларитету корена од кога се граде, можемо се суочити са

бројним одступањима од таквог правила. Рецимо, речи *крајње* и *бескрајно* имају исти корен *крај*, али нису супротних поларитета осећања, упркос постојању префикса *бес-*[245] у речи *бескрајно*.

Други нежељени ефекат који се може произвести применом кореновања односи се, такође, на могућност постојања различитог поларитета осећања речи које представљају парониме. Пароними су речи које се добијају деривацијом из истог корена, имају слично значење, али нису синоними. Напротив, некада могу имати метафоричко значење које им даје супротно семантичко значење од дословног што може значити и промену поларитета осећања. На пример, речи *политичар* и *политикант* су парониме и имају исти корен *политик*. Реч *политичар* нема поларитет осећања, односно није реч са субјективним значењем, док реч *политикант* има изразито негативан поларитет. Такође, речи *чистоћа* и *чистота* имају исти корен *чист*, али реч *чистоћа* нема поларитет, док *чистота* носи позитиван поларитет осећања.

Ови примери показују да коришћење стемера са сврхом редукције векторског простора предиктора може довести до губитка или до погрешног означавања поларитета осећања, што се у морфолошки богатим језицима може мултиплицирати због великог броја флективних и деривационих форми. У циљу превазилажења ове врсте проблема, у овој тези предлажемо другачији приступ редукцији векторског простора скупа за учење уместо кореновања. Идеја је да се не користи стемер, већ да се задрже сви флективни облици виђени у скупу за учење. На тај начин се задржава исправна информација о поларитету. С друге стране, с обзиром да лексикони којима располажемо представљају скупове субјективних речи и израза и свих њихових флективних облика, можемо установити функцију пресликавања којом се свака сентиментална реч скупа за учење, нађена у лексикону, замењује тачно једном фиксном ознаком из скупа датог у (81) која даје информацију о поларитету осећања.

[245] Префикси *бес* и *без* су аломорфи.

5.3.3 Принцип семантичког пресликавања концепата

Алгоритам пресликавања сентименталних речи, којим се скуп за учење мења, дат је у наставку (алгоритам 5.2). Алгоритам се на исти начин примењује и над скупом за тестирање. Идући секвенцијално кроз документ, алгоритам проверама сваки n-грам речи ($n \leq 3$), почев од триграма ка краћим, утврђујући да ли он у неком од лексикона постоји. Ако постоји, утврђује се његов поларитет у лексикону и генерише одговарајућа замена за њега у облику фиксне речи из скупа (81). Када се заврши са испитивањем триграма, иста процедура спроведе се над биграмима и униграмима, а затим се прелази на следећи документ.

Алгоритам 5.2 Пресликавање сентименталних речи као и сентименталних израза чије су дужине 2 или 3 речи

Input: Document_set_D=$\{d_1, d_2, ..., d_n\}$
//Set of textual documents where n is the total number of documents
Output: Document_set_$D_{changed}$=$\{dc_1, dc_2, ..., dc_n\}$
//Set of changed input textual documents in which, all MWUs whose length is 2 or 3
//and after that all sentiment single words,
//detected by lexicons, are mapped to corresponding fixed words.

```
1. for each document in document_set_D
2.    document1= document
3.    for each 3-gram in document1 //first iterate 3-grams
4.       changeContent(3-gram, var document1);
5.    for each 2-gram in document1 //then iterate 2-grams
6.       changeContent(2-gram, var document1);
7.    for each 1-gram in document1 //finally, iterate 1-grams
8.       changeContent(1-gram, var document1);
9. return document1;
```

Procedure: changeContent(n-gram, var document)

```
1. if (sentimentLexicon.find(n-gram).Positive)
2.    document.replace(n-gram, 'LexiconPOS')
3. else if (sentimentLexicon.find(n-gram).Negative)
4.    document.replace(n-gram, 'LexiconNEG')
5. else if (WordNetLexicon.find(n-gram).Positive)
6.    document.replace(n-gram, 'WordNetPOS')
7. else if (WordNetLexicon.find(n-gram). Negative)
8.    document.replace(n-gram, 'WordNetNEG');
```

Принцип измене приказаћемо на следећем примеру. Претпоставимо да је скуп за учење представљен следећим текстом (3 реченице) који ћемо означити са *Модел1*.

Та *љубавна прича* започела је летос. Њиховом *љубавном причом* се сви баве. Та *љубав* траје и данас.

Уз претпоставку да се два флективна облика фразе *љубавна прича* (*љубавна прича, љубавном причом*) и лема *љубав* налазе у лексикону сентименталних речи и израза, применом алгоритма пресликавања (*Algorithm 2) Модел1* трансформише се у *Модел2*.

Та LexiconPOS започела је летос. Њиховом LexiconPOS се сви баве. Та LexiconPOS траје и данас.

На основу примера трансформације модела *Модел1* у *Модел2*, може се уочити да се применом предложене функције пресликавања постиже редукција векторског простора. Пре свега се смањује укупан број униграма који представљају кандидате за подскуп униграмских предиктора. Такође, може се уочити да у првој и трећој реченици модела *Модел1* фигурирају триграм (*ta ljubavna priča*) и биграм (*ta ljubav*), док се у моделу *Модел2* они пресликавањем трансформишу у исти биграм (*ta LexiconPOS*).

После извршене трансформације скупа за учење, генеришу се предиктори *MaxEnt* класификатора. Осим редукованог основног скупа n-грамских предиктора дефинишемо и четири нова предиктора у одговарајућем формату (в. одељак 2.6.1), где је c ознака класе.

$$f_1(c,x) = \begin{cases} 1 & x = LexiconPos, c = Positive \\ 0 & otherwise \end{cases}$$
$$f_2(c,x) = \begin{cases} 1 & x = LexiconNeg, c = Negative \\ 0 & otherwise \end{cases}$$
$$f_3(c,x) = \begin{cases} 1 & x = WordNetPos, c = Positive \\ 0 & otherwise \end{cases}$$
$$f_4(c,x) = \begin{cases} 1 & x = WordNetNeg, c = Negative \\ 0 & otherwise \end{cases}$$

У одељку 5.4.6 биће приказан однос броја предиктора пре и после примене редукције и биће упоређене оцене обе групе метода (оних код којих је примењена предложена редукција векторског простора у односу на оне код којих није).

5.4 САФОС – Радно окружење за класификацију текстова на српском језику на основу осећања

5.4.1 Изградања лексикона сентименталних речи и израза

Како смо истакли у поглављу 4, лексикони сентименталних речи и израза имају значајну улогу у SA, посебно у методама заснованим на лексиконима. У поступку изградње таквог лексикона за српски језик, кренули смо од Плутчикове димензионалне теорије емоција изражених тзв. Плутчковим точком (в. одељак 4.4.2). Скуп од 24 концепта емоција представљао је основу (тзв. *seed words*) изградње нашег лексикона. Речник је имао две фазе развоја. У иницијалном кораку прве фазе, лексикон је проширен синонимима речи које су представљале 24 концепта и у ту сврху је коришћен асоцијативни речник (Jovanović & Atanacković, 1980). Из датог речника аутоматски су екстраховане одреднице које представљају асоцијативне појмове датих речи. Да би се лексикон могао и даље ширити синонимима основних речи, креирана је веб апликација, чија је веб страница за претрагу дата на примеру речи *изненађење* и приказана у прилогу 5.1. Из примера се може уочити да се основна Плутчикова емоција *surprise* (*изненађеност*) може проширити синонимима: *атракција, атрактиван, банути, бити изненађен, зачуђен, изнебуха, изненадни догађај, изненађеност, нагло, напречац, ненадан гост, ненајављен, неочекивани догађај, одједаред* и др. Међутим, лексикон није проширен искључиво синонимима исте врсте речи (нпр. *атракција, изненађеност*) већ и другим врстама речи (нпр. *банути, напречац, ненајављен*) као и фразама (нпр. *ненадан гост, изненадни догађај*). Након проширења основног скупа појмова, наш лексикон је садржао 1053 речи и израза којима се изражавају емоције (осећања) на српском језику.

У другој фази развоја, тих 1053 одредница проширено је флективним облицима који су пронађени уз помоћ морфолошких електронских речника српског језика алгоритмом 5.1 описаним у одељку 5.3.2. Веб апликација која омогућава доградњу и претрагу лексикона приказана је у прилогу 5.2 на примеру претраге проширења речи *зачудити* флективним облицима као што су: *зачудише, зачудивши, зачудио, зачудила, зачудило* и др. Након проширења,

лексикон је садржао 10704 флективних облика речи и фраза полазног скупа одредница.

5.4.2 Генерисање ознака поларитета осећања у синсетовима SWN

Српски ворднет (SWN), описан у одељку 3.3.3, за потребе извршења задатка редукције скупа предиктора, најпре је проширен синсетовима придева који носе субјективно значење. Укупно је додато 472 синсета. У наредном кораку, SWN је проширен етикетама POSITIVE и NEGATIVE. Вредности ових етикета у сваком синсету додељиване су на основу поларитета и интензитета осећања које носи концепт репрезентован синсетом и то је учињено уз помоћ ресурса *SentiWordNet 3.0* (в. одељке 3.5 и 4.5.1). У случајевима када се радило о специфичним концептима балканских језика (синсетови означени BILI ознакама) или специфичним концептима српског језика (синсетови означени SRP ознакама), такви синсетови су аутоматски наслеђивали вредности етикета POSITIVE и NEGATIVE етикета од својих родитеља с којима су повезани релацијом *hypernym*. Уколико родитељи нису имали дефинисане вредности ових етикета оне су уношене ручно уколико их је било могуће дефинисати, а у супротном су остајале недефинисане. На крају овог процеса интеграције ресурса *SentiWordNet 3.0* у SWN, укупно 4044 синсетова је било означено вредностима позитивног и негативног поларитета осећања. Тако проширени SWN употребили смо у изградњи другог лексикона сентименталних речи и израза. Из сваког синсета чији поларитет осећања има вредност већу од 0.75 у позитивном или негативном смислу, екстраховане су све синонимске леме, а из дефиниције (енг. gloss) креирани су биграми и триграми. Контрола биграма и триграма у циљу екстракције фраза са субјективним значењем извршена је ручно. На тај начин је добијен лексикон са 981 флективним обликом. На пример, у случају синсета који представља концепт *ратоборно,* и чију *XML* репрезентацију наводимо у наставку, у лексикон се уносе одреднице добијене екстракцијом из етикета литерала (*ратоборно*, *непријатељски, с непријатељством*), док се из етикете дефиниције формирају биграми (*на ратоборан, ратоборан*

непријатељски, *непријатељски начин*) и триграми (*на ратоборан непријатељски, ратоборан непријатељски начин*).

```
<SYNSET>
 <ID>ENG30-00242478-b</ID>
 <POS>b</POS>
 <SYNONYM>
  <LITERAL>ratoborno<SENSE>1</SENSE><LNOTE /></LITERAL>
  <LITERAL>neprijateljski<SENSE>1</SENSE><LNOTE /></LITERAL>
  <LITERAL>s neprijateljstvom<SENSE>1</SENSE><LNOTE /></LITERAL>
 </SYNONYM>
 <DEF>i na ratoboran neprijateljski način</DEF>
 <ILR>ENG30-01244410-a<TYPE>derived</TYPE></ILR>
 <NL>yes</NL>
 <STAMP>jeca 12/06/2012 00:00:00</STAMP>
 <SUMO>SubjectiveAssessmentAttribute<TYPE>+</TYPE></SUMO>
 <SENTIMENT>
  <POSITIVE>0.00000</POSITIVE>
  <NEGATIVE>0.75000</NEGATIVE>
 </SENTIMENT>
 <DOMAIN>factotum</DOMAIN>
</SYNSET>
```

5.4.3 Изградња и оцена ресурса „Стоп-речи српског језика“

Стоп-речи представљају листу вискофреквентних речи неког језика које су обично граматичке речи које не носе значење, па се у неким задацима обраде природног језика (као што је анализа осећања) могу изоставити. У задацима класификације текста и анализе осећања, њихово уклањање представља један од ефикасних начина редукције векторског простора предиктора. Иако нека истраживања, попут ван Ријсбергенових[246] (van Rijsbergen, 1979), показују да процес уклањања вискофреквентних речи може довести до ступња на коме преостали предиктори нису адекватан репрезентативни скуп датог текста, изградња оптималног скупа стоп-речи (Manco, Masciari, Ruffolo & Tagarelli, 2002) је важан део многих задатака обраде природног језика. Листа стоп-речи може бити листа у општем смислу (тзв. generic stop-words), када се креира на основу неког корпуса опште

[246] Cornelis Joost van Rijsbergen

намене, са сврхом да буде примењена у решавању проблема у различитим семантичким доменима, а може бити и доменски оријентисана, када се креира из корпуса који се односи на специфични домен знања. Према Менингу (Manning, Raghavan & Schütze, 2008) уобичајена стратегија за развој листе стоп-речи подразумева сортирање свих појмова према фреквенцији појављивања у датој колекцији документа и одабир k најфреквентнијих (TF метода). Постоје и друге, често примењиване технике као што су TF.IDF (Manning, Raghavan & Schütze, 2008), delta TF.IDF (Martineau & Finin, 2009), фреквенција суседних речи (Rose, Engel, Cramer & Cowley, 2010), статистичко моделирање (Zou, Wang, Deng, Han & Wang, 2006) и др. Пошто нека истраживања (Surendran, Platt & Renshaw, 2005) указују да доменски оријентисане листе стоп-речи показују већу ефикасност, ми смо у овом истраживању одлучили да упоредимо ефекте примене једне листе стоп-речи опште намене и двеју доменски оријентисаних листи добијених TF.IDF и TF техникама над скупом за учење.

Техника TF.IDF (Manning, Raghavan & Schütze, 2008) најчешће се користи у задацима проналажења информација да покаже степен релевантности посматраног упита у односу на један документ у датој колекцији. Састоји се у корелацији двају параметара: TF – логаритма фреквенције појављивања $tf_{t,d}$ упитне кључне речи t у документу d и IDF – логаритма односа укупног броја докумената N и броја докумената df_t који садрже упитну кључну реч.

$$TF.IDF_{t,d} = \left(1 + \log\left(tf_{t,d}\right)\right) \times \log\left(\frac{N}{df_t}\right)$$

Релевантност (репрезентативност) једне упитне кључне речи за један документ, мерена фактором $TF.IDF_{t,d}$, повећава се уколико се повећава број њених појављивања у документу ($tf_{t,d}$) и смањује број докумената у колекцији у којима се она појављује (df_t). Стоп-речи су речи чија је фреквенција појављивања у једном документу висока, али је и број докумената у колекцији у којима се појављују велики, па имају мали TF.IDF фактор. Можемо речи да су стоп-речи слабо репрезентативне у погледу TF.IDF фактора.

Техника TF оцењује релевантност упитне кључне речи само на основу фреквенције појављивања у датој колекцији докумената. Све три листе су одабране тако да садрже исти број лема (577) али после проширења флективним облицима коначна величина ових листи је била различита. Поступак проширења флективним облицима изведен је применом алгоритма 5.1 који је описан у одељку 5.3.2. Листу стоп-речи опште намене креирали смо применом критеријума одабира 577 најфреквентнијих речи из 122-милионског корпуса савременог српског језика описаног у одељку 3.1. Листа коју смо добили из корпуса садржала је речи у произвољном флективном облику, али смо увели претпоставку да уколико је нека реч у одређеном облику не носи значење, не носе ни други флективни облици те речи. Зато су све речи описане листе најпре преведене у облик леме помоћу софтвера за екстракцију n-грама[247] (Utvić, 2014b), а затим је листа проширена, помоћу електронских морфолошких речника (алгоритам 5.1), флективним облицима. Након проширења листа је садржала 831 лему и све флективне облике лема заменица, бројева, предлога, везника и помоћних глагола. За разлику од листе стоп-речи опште намене која је садржала само семантички нерелевантне врсте речи, друге две доменски оријентисане листе садржале су све врсте речи. Једна од њих је генерисана из скупа за учење описаног у наредном одељку одабиром 577 речи са најмањом вредношћу TF.IDF фактора. Након проширења флективним облицима, ова листа је имала 1428 флективних облика. Друга доменски оријентисана листа, добијена TF техником, након флективног проширења садржала је 1372 облика речи. Оцена доприноса ових листи стоп-речи процесу редукције векторског простора предиктора у задатку класификације осећања извршена је у SAFOS радном окружењу за анализу осећања (в. одељак 5.4.6) над скупом за тестирање „оцене филмова“. Утврђивањем доприноса сваке листе дошли смо до показатеља да је он приближно исти за све три (в. одељак 5.4.7), па смо се одлучили да у наставку истраживања користимо листу стоп-речи опште намене.

[247] http://arhimed.matf.bg.ac.rs/~misko/ngram/ngram.php

5.4.4 Изградња скупова за учење и тестирање

Изградња скупа за учење у задатку машинског учења је процес који врло често захтева експертска знања у процесу анотације чланова скупа за учење, што подразумева селекцију великог броја узорака и њихову анотацију, односно класификацију према унапред датим критеријумима. У случају анализе осећања, потребно је одабрати скуп текстова и анотирати их, према поларитету осећања који одликује дати текст, једном од двеју ознака: позитивно и негативно. У поступку анотације скупа за учење могу се користити аутоматске, полу-аутоматске и ручне технике означавања. У овом раду представићемо једну од аутоматских техника одабира и анотације текстова намењених изградњи скупа за учење у решавању проблема класификације осећања на нивоу докумената.

У том поступку смо најпре одабрали веб портале, односно њихове сегменте који се баве објављивањем тзв. „црних хроника“ - текстова који описују и исказују негативна осећања изазвана исто таквим догађајима (убиства, пљачке, проневере, итд.). Затим смо одабрали портале који се баве промоцијом друштвено одговорног понашања и добрих навика у свакодневном животу и раду. У првом случају смо одабрали један сајт[248], а у другом три[249] како бисмо обезбедили баланс у укупном броју негативно и позитивно субјективно оријентисаних текстова. Овај први део процеса који се односи на одабир сегмената описаних портала је једини изведен ручно, али је и унапред дефинисао природу докумената који су селектовани. У наставку процеса користили смо софтвер за аутоматско преузимање дела садржаја веб сајта и креирали софтвер за екстракцију текста из HTML формата. Затим смо уједначили кодне стране текстова и преименовали их тако да садрже ознаку POS или NEG у зависности од садржаја. Као резултат, добили смо скуп за учење од 2000 (1000 позитивних и 1000 негативних) докумената са 48.328 реченица и 1.006.822 облика речи.

[248] http://www.kurir-info.rs/crna-hronika
[249] http://www.dobrevesti.rs; http://www.svrljig.info/vesti/dobre-vesti; http://www.ilovezrenjanin.com/category/ drustvena-odgovornost-2/

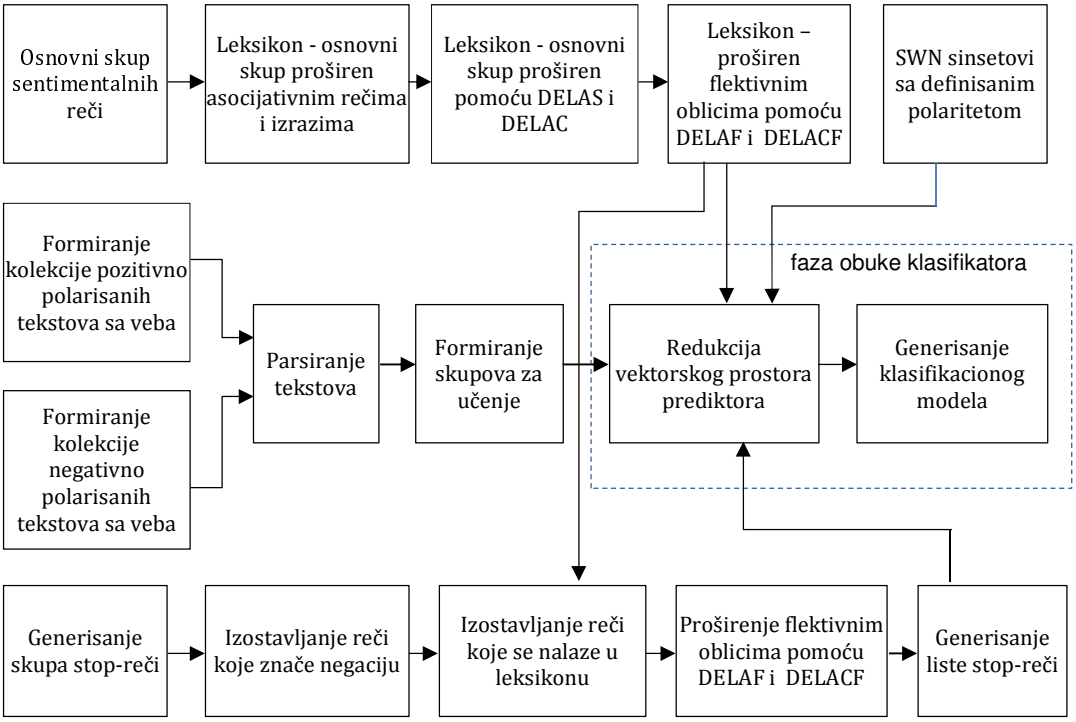
За потребе процеса оцене система SAFOS (в. одељак 5.4.5) генерисали смо по истој, претходно описаној, логици два скупа за тестирање. Један од њих, „оцене вести“, креиран је на основу негативно поларисаних текстова објављених на сајту[250] који се бави појавама корупције и неправилности у јавном раду и позитивно поларисаних текстова објављених на сајту[251] дневих вести. Парсирано је и означено 408 позитивних текстова и 371 негативан текст, па је коначан скуп за тестирање „оцене вести“ бројао 779 докумената са 16.304 реченица и 438.025 облика речи.

Други скуп за тестирање односио се на оцене филмова, а генерисан је на основу садржаја сајта[252] за оцену и рангирање филмова, музике, ТВ серија и рачунарских игара на српском језику. Ми смо одабрали само сегмент који се односио на оцене филмова. Сваки филм у овом сегменту оцењивао је модератор сајта оценом израженом децималним бројем из опсега [0,00-10,00] и описивао пропратним текстом чија је дужина у просеку 175 речи. Филмове чија је оцена била већа или једнака 5,00 означили смо ознаком POS, а остале ознаком NEG. Овај скуп за тестирање „оцене филмова“ садржао је после парсирања, уједначавања кодне стране и доделе ознака класа 2237 докумената са 16.534 реченица и 392.218 облика речи. Скуп је, за разлику од претходног, небалансиран у погледу броја чланова обеју класа јер је садржао 1890 позитивно и 347 негативно означених текстова, што је последица начина оцењивања модератора сајта.

5.4.5 САФОС - класификација текстова на српском језику на основу осећања

Радно окружење за анализу осећања у текстовима на српском језику (Sentiment analysis framework for Serbian - SAFOS) је скуп софтверских алата које смо креирали и повезали тако да користе *MaxEnt* ML методу за класификацију текстова према осећањима на нивоу докумената. Структура система SAFOS представљена је на слици 5.2.

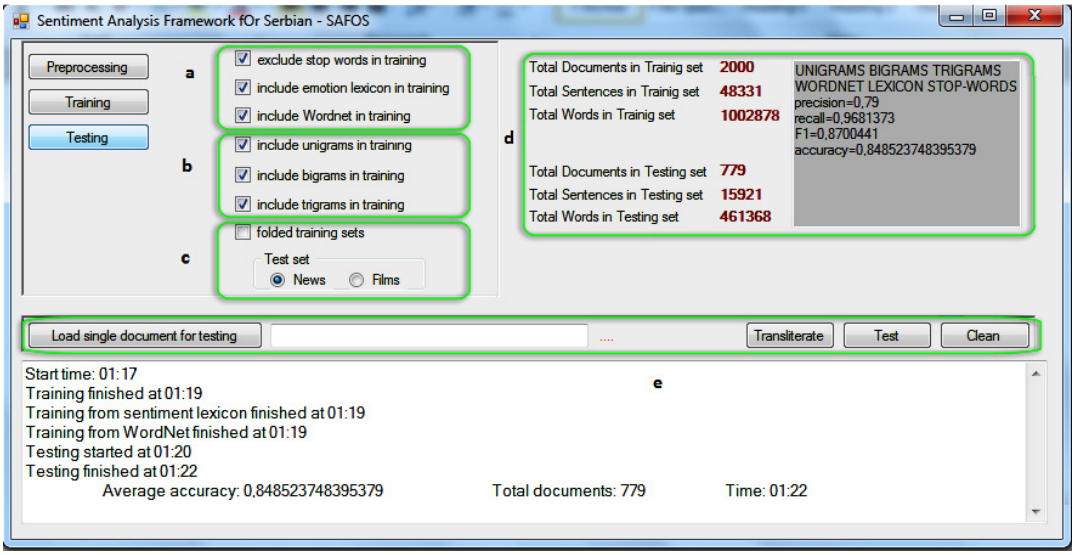
[250] http://pistaljka.rs/
[251] http://www.vesti.rs/Dobre-vesti
[252] http://2kokice.com/



Слика 5.2 SAFOS – модуларни систем за дефинисање скупа предиктора у фази пре учења и обуку модела за класификацију текстова на основу осећања

SAFOS интегрише све ресурсе и поступке њиховог генерисања које смо претходно описали у овом поглављу: скуп за учење, листу стоп-речи, лексикон сентименталних речи и израза и лексикон добијен на основу поларитета осећања синсетова Српског ворднета. Сви ови ресурси могу учествовати у поступку редукције векторског простора предиктора описаног у одељку 5.3.3 и реализованог алгоритмом 5.2. У корисничком интерфејсу SAFOS-а приказаног на слици 5.3 може се извршити избор оних ресурса (стоп-речи, лексикон, ворднет лексикон) који учествују у процесу редукције као и структура предиктора (униграми, биграми, триграми) који се укључују у процес генерисања класификационог модела (дугме *Preprocessing*). Такође, може се након обуке (дугме *Training*) одабрати скуп за тестирање („оцене вести“, „оцене филмова“) за оцену параметара система (дугме *Testing*). Параметри којима се оцењује систем описани у одељку 2.7 (прецизност, одзив, F1-мера и тачност) као и остали параметри система који доводе до

конкретних вредности ових оцена приказују се на форми главне стране SAFOS-а. Након изградње модела, могуће је анотирати произвољан непознат текст његовим учитавањем и тестирањем (дугмад *Load single document for testing* и *Test*).



Слика 5.3 Кориснички интерфејс система SAFOS а) избор параметара у фази претпроцесирања б) избор параметара у фази обуке-тренирања ц) избор скупа за тестирање д) приказ резултата евалуације система, параметара скупова за учење и предиктора који учествују у евалуацији е) примена модела - анотирање појединачних докумената

5.4.6 Оцена модела у систему САФОС

SAFOS обезбеђује два начина евалуације модела: 10-струку унакрсну валидацију (CV) и примену скупа за тестирање (TS). У случају примене скупа за тестирање могу се користити „оцене вести“ или „оцене филмова“. Евалуацијом оцењујемо структуру предиктора која даје најбоље вредности параметара за оцену. У 10-струкој унакрсној валидацији, у свакој од 10 итерација, скуп за учење садржи 1800 докумената, а преосталих 200 чине скуп за тестирање. Вредности параметара евалуације израчунавају се као просек свих итерација. Евалуацију вршимо тако што меримо најпре допринос само униграмскх предиктора (U), затим униграмских и биграмских (U+B) и

најзад униграмских, биграмских и триграмских (U+B+T). Сваку од три комбинације предиктора анализирамо на 4 различита начина:

1. без редукције скупа предиктора
2. са редукцијом предиктора само применом листе стоп-речи (примену ове врсте редукције означићемо са *SWL*)
3. са редукцијом предиктора само применом оба лексикона (примену ове врсте редукције означићемо са *mapping*)
4. са редукцијом предиктора применом листе стоп-речи и оба лексикона (примену ове врсте редукције означићемо са *SWL + mapping*)

Експеримент смо извели 12 пута са различитим скуповима предиктора и добили 12 различитих модела класификације чије смо резултате унакрсне валидације упоредили. Резултати су приказани у табели 5.1.[253] Ознаком Cf_1 означени су резултати оних модела у којима нема примене лексикона, па самим тим ни нашег предложеног модела редукције предиктора. Ознаком Cf_2 означени су резултати оних модела у којима се примењују лексикони и уводи предложени модел редукције. Анализом табеле се може уочити да редукција коју предлажемо у нашем истраживању унапређује модел класификације по сва 4 параметра ($Cf_2 > Cf_1$), а у одељку 5.4.7 анализираћемо који од модела показују статистички значајна побољшања.

Други експеримент који смо спровели користи скуп за тестирање „оцене вести“ чија је структура описана у одељку 5.4.4. Спроведени начин евалуације исти је као и у случају методе унакрсне валидације. Резултати су приказани у табели 5.2, чијом се анализом може уочити да редукција коју предлажемо унапређује модел класификације по сва 4 параметра.

Трећи експеримент који смо спровели користи скуп за тестирање „оцене филмова“, на исти начин као код методе унакрсне валидације. Резултати су приказани у табели 5.3, а анализа показује да модел са редукцијом коју предлажемо показује боље вредности сва 4 параметра класификације.

[253] Табеле 5.1 до 5.5 преузете из рада (Mladenović, Mitrović, Krstev & Vitas, 2015)

Табела 5.1 Оцена модела класификације 10-струком унакрсном валидацијом

Feature set	precision	recall	F1	accuracy
$Cf_1(U)$	0.925	0.944	0.934	0.950
$Cf_2(U + mapping)$	0.937	0.950	0.943	**0.956**
$Cf_1(U + SWL)$	0.922	0.943	0.932	0.948
$Cf_2(U + SWL + mapping)$	0.934	0.946	0.940	0.954
$Cf_1(U + B)$	0.889	0.914	0.910	0.909
$Cf_2(U + B + mapping)$	0.904	0.919	0.912	0.914
$Cf_1(U + B + SWL)$	0.919	0.943	0.931	0.946
$Cf_2(U + B + SWL + mapping)$	0.929	0.951	0.940	0.950
$Cf_1(U + B + T)$	0.779	0.825	0.800	0.787
$Cf_2(U + B + T + mapping)$	0.783	0.820	0.800	0.788
$Cf_1(U + B + T + SWL)$	0.899	0.921	0.910	0.920
$Cf_2(U + B + T + SWL + mapping)$	0.908	0.924	0.916	0.923

Табела 5.2 Оцена модела класификације применом скупа за тестирање „оцене вести“

Feature set	precision	recall	F1	accuracy
$Cf_1(U)$	0.708	0.821	0.760	0.729
$Cf_2(U + mapping)$	0.727	0.824	0.772	0.746
$Cf_1(U + SWL)$	0.710	0.824	0.763	0.732
$Cf_2(U + SWL + mapping)$	0.730	0.821	0.773	0.747
$Cf_1(U + B)$	0.727	0.880	0.796	0.764
$Cf_2(U + B + mapping)$	0.752	0.887	0.814	0.788
$Cf_1(U + B + SWL)$	0.736	0.855	0.791	0.764
$Cf_2(U + B + SWL + mapping)$	0.767	0.865	0.813	**0.792**
$Cf_1(U + B + T)$	0.720	0.814	0.764	0.737
$Cf_2(U + B + T + mapping)$	0.745	0.824	0.782	0.760
$Cf_1(U + B + T + SWL)$	0.736	0.841	0.785	0.759
$Cf_2(U + B + T + SWL + mapping)$	0.768	0.850	0.807	0.787

Табела 5.3 Оцена модела класификације применом скупа за тестирање „оцене филмова“

Feature set	precision	recall	F1	accuracy
$Cf_1(U)$	0.831	0.736	0.781	0.651
$Cf_2(U + mapping)$	0.877	0.800	0.837	0.736
$Cf_1(U + SWL)$	0.828	0.728	0.775	0.642
$Cf_2(U + SWL + mapping)$	0.877	0.802	0.838	0.738
$Cf_1(U + B)$	0.839	0.882	0.860	0.758
$Cf_2(U + B + mapping)$	0.855	0.895	0.874	**0.783**
$Cf_1(U + B + SWL)$	0.839	0.818	0.828	0.714
$Cf_2(U + B + SWL + mapping)$	0.869	0.851	0.860	0.766
$Cf_1(U + B + T)$	0.839	0.879	0.859	0.755
$Cf_2(U + B + T + mapping)$	0.853	0.889	0.871	0.777
$Cf_1(U + B + T + SWL)$	0.836	0.848	0.842	0.731
$Cf_2(U + B + T + SWL + mapping)$	0.858	0.875	0.867	0.772

Предложена метода редукције примењена у систему SAFOS смањује на следећи начин укупан број предиктора:

- модел U+SWL користи 92.155, а редуковани U+SWL+mapping 91.039 предиктора
- модел U+B+SWL користи 404.911, а редуковани U+B+SWL+mapping 401.907 предиктора
- модел U+B+T+SWL користи 476.765, а редуковани U+B+T+SWL+mapping 476.471 предиктор.

Резултати свих експеримената показују да класификација текстова према осећањима која користи методу пресликавања предиктора сентименталних израза да би се редуковао векторски простор предиктора постиже боље резултате у односу на исту класификацију која предложену методу не користи. У случајевима експеримената са скуповима за тестирање, највећа тачност класификације скупа за тестирање „оцене филмова“ од 78,3% постиже се комбинацијом униграмских и биграмских предиктора редукованих предложеном методом пресликавања. У случају скупа за тестирање „оцене вести“ најбољи модел класификације са тачношћу од 79,2% даје исти скуп предиктора, редукован стоп-речима и методом пресликавања. У методи 10-струке унакрсне валидације највећа тачност од 95,6% постиже се скупом униграма који су редуковани пресликавањем. Свеукупна боља тачност класификације методом унакрсне валидације у односу на примену два независна скупа за тестирање објашњава се чињеницом да су и скуп за учење и скуп за тестирање из истог доменског простора, док се код независних скупова за тестирање ради о подацима који моделу нису познати (тзв. unseen data).

5.4.7 Статистичка значајност модела у систему САФОС

У циљу оцене статистичке значајности резултата представљених у претходном одељку, применили смо *t*-тест за зависне узорке. У случају 10-струке CV методе, постављањем нулте хипотезе желели смо да оценимо перформансе система у погледу тачности класификације, па она гласи:

H_0: Разлика тачности два класификациона модела није статистички значајна на нивоу поверења од 95%.

При оцењивању статистичке значајности тачности класификације у случајевима примене скупова за тестирање применили смо z-тест јер је број елемената који су директно поређени био већи од 30. Дефиниција нулте хипотезе иста је као у случају методе унакрсне валидације. Резултати статистичких тестова, дати у табели 5.4, показују да код методе унакрсне валидације употреба униграма редукованих методом пресликавања групе предиктора сентименталних израза, постиже статистички боље резултате у односу на исту комбинацију предиктора без редукције ($p < 0{,}05$).

У случају скупа за тестирање „оцене филмова“ сви тестирани скупови предиктора дају статистички боље резултате класификације када се изврши редукција методом пресликавања групе предиктора сентименталних израза у односу на исту комбинацију предиктора без редукције. Чак све комбинације предиктора редуковане и стоп-речима исказују боље резултате са нивоом поверења од 99% ($p < 0{,}01$).[254] У случају скупа за тестирање „оцене вести“ два модела дају статистички боље резултате класификације ($p < 0{,}05$): модел комбинације униграма и биграма уз редукцију методом пресликавања и исти модел уз додатну редукцију помоћу стоп-речи.

У табели 5.5 приказали смо резултате експеримента којим смо оценили доприносе трију различитих листи стоп-речи описаних у одељку 5.4.3. Оцена је извршена применом скупа за тестирање „оцене филмова“, а одабран је скуп предиктора U+B+T+SWL+mapping. Анализом резултата датих у табели може се закључити да листа стоп-речи опште намене и TF.IDF листа дају бољу тачност класификације, али приближно исту под истим условима, па смо се одлучили да у SAFOS-у, у коначној инстанци, користимо листу стоп-речи опште намене. Стога су сви резултати које смо претходно описали, а односе се на редукцију предиктора применом листе стоп-речи, користили ону опште намене.

254 Додатна хипотеза H_0: Разлика тачности два класификациона модела није статистички значајна на нивоу поверења од 99%.

Најзад, анализирајући утицај листе стоп-речи (SWL) у процесу редукције предиктора, можемо истаћи њен променљив утицај на тачност класификације. У 10-струкој унакрсној валидацији, примена ресурса SWL статистички значајно унапређује тачност класификације у случају свих скупова предиктора, осим код скупова U и U+mapping, где незначајно смањује тачност. Код скупа „оцене вести“ SWL не смањује тачност у случају скупа U+B, а незначајно побољшава у случају свих осталих скупова предиктора. Код скупа „оцене филмова“ SWL утиче на смањење тачности класификације, сем у случају U+mapping када долази до незначајног побољшања. Иако овај елеменат редукције исказује променљив утицај на статистику класификације, с обзиром да је највећу номиналну тачност исказао скуп U+B+SWL+mapping, одлучили смо да SWL остане саставни део SAFOS система.

Табела 5.4. Резултати тестирања статистичке значајности модела класификације, зависно од скупа предиктора и метода редукције
($* p < 0{,}05$; $** p < 0.01$)

Dataset	Feature sets	p-value
10-fold CV	$Cf_2(U + mapping) > Cf_1(U)$	0.032*
	$Cf_2(U + SWL + mapping) > Cf_1(U + SWL)$	0.048*
	$Cf_2(U + B + mapping) > Cf_1(U + B)$	0.114
	$Cf_2(U + B + SWL + mapping) > Cf_1(U + B + SWL)$	0.054
	$Cf_2(U + B + T + mapping) > Cf_1(U + B + T)$	0.448
	$Cf_2(U + B + T + SWL + mapping) > Cf_1(U + B + T + SWL)$	0.190
news testing set	$Cf_2(U + mapping) > Cf_1(U)$	0.22
	$Cf_2(U + SWL + mapping) > Cf_1(U + SWL)$	0.23
	$Cf_2(U + B + mapping) > Cf_1(U + B)$	0.03*
	$Cf_2(U + B + SWL + mapping) > Cf_1(U + B + SWL)$	0.04*
	$Cf_2(U + B + T + mapping) > Cf_1(U + B + T)$	0.10
	$Cf_2(U + B + T + SWL + mapping) > Cf_1(U + B + T + SWL)$	0.07
films reviews testing set	$Cf_2(U + mapping) > Cf_1(U)$	<0.0001**
	$Cf_2(U + SWL + mapping) > Cf_1(U + SWL)$	<0.0001**
	$Cf_2(U + B + mapping) > Cf_1(U + B)$	0.03*
	$Cf_2(U + B + SWL + mapping) > Cf_1(U + B + SWL)$	<0.0001**
	$Cf_2(U + B + T + mapping) > Cf_1(U + B + T)$	0.03*
	$Cf_2(U + B + T + SWL + mapping) > Cf_1(U + B + T + SWL)$	0.002**

Табела 5.5 Поређење резултата класификације применом три различите листе стоп-речи над скупом за тестирање „оцене филмова“

Stop word list	Generic stop words list	Stop words list from training set by TF.IDF method	Stop words list from training set by TF method
precision	0.858	0.864	0.868
recall	0.875	0.859	0.828
F1 score	0.867	0.862	0.847
accuracy	**0.772**	0.767	0.748

5.4.8 Поређење са постојећим моделима анализе осећања

Када је реч о српском језику, према нашим сазнањима, два досад објављена резултата анализе осећања односе се на модел идентификације поларитета осећања на нивоу реченица (твитова) методама надгледаног машинског учења. У раду (Jolić, 2014) коришћене су методе NB, ME и SVM са две врсте предиктора. У случају примене униграма као предиктора постигнута је тачност ME модела 80,5%, а у случају примене униграма и биграма тачност ME модела је 82,7%. У раду није наведена метода оцене модела (тестирање независним скупом или унакрсна валидација). У другом раду, (Milošević, 2012), дат је приказ изградње модела идентификације поларитета осећања на нивоу реченица (твитова) NB методом надгледаног машинског учења. Резултати у погледу тачности класификације нису објављени. Када је реч о другим језицима и примени ME метода, можемо посматрати две врсте модела: моделе тестиране унакрсном валидацијом, приказане у табели 5.6, и моделе тестиране независним тест подацима приказане у табели 5.7.

Табела 5.6. Резултати класификације осећања ME методом са унакрсном валидацијом, на различитим језицима и скуповима за учење

Аутор	**Скуп за обуку**	**Језик**	**Резултати класификације осећања ME методом са унакрсном валидацијом**
(Pang, Lee & Vaithyanathan, 2002)	movie dataset	енглески	Average three-fold cross-validation: Unigrams Acc=0,804 Uni+Big Acc=0,808 Bigrams Acc=0,774
(Zhang, Huang & Wu, 2008)	movie dataset	енглески	Average 10-fold cross-validation: best ME Acc=0,889
(Blinov et al., 2013)	1.book reviews 2. camera reviews 3. movie reviews	руски	1. Acc=0,884 2. Acc=0,937 3. Acc=0,831
(Habernal, Ptàček & Steinberger, 2014)	Czech CSFD movie dataset	чешки	Average 10-fold cross-validation: ME baseline Acc=0,775 sspace Acc=0,787 Dir Acc=0,805 sspace + Dir Acc=0,815
(Altrabsheh, Cocea & Fallahkhair, 2014)	collection of feedback in lectures	енглески	Average 10-fold cross-validation: best ME Acc=0,57
(Wang et al. 2014)	product reviews	енглески	camera Acc=0,807 laptop Acc=0,926 radio Acc=0,828
(Øye, 2015)	twitter	норвешки	Acc=0,72 (PA скуп предиктора) Acc=0,75 (PB скуп предиктора) Acc=0,71 (PC скуп предиктора)
(Mladenović, Mitrović, Krstev & Vitas, 2015)	**новинске вести**	**српски**	**Average 10-fold cross-validation:** **Un + mapping Acc=0,956** **Un+Bi+mapping Acc=0,914** **Un+Bi+Tr+mapping Acc=0,923**

Табела 5.7 Резултати класификације осећања ME методом над независним скупом за тестирање, на различитим језицима и скуповима за учење

Аутор	**Скуп за обуку**	**Скуп за тестирање**	**Језик**	**Резултати класификације осећања ME методом над независним скупом за тестирање**
(Mehra, Khandelwa & Patel, 2002)	imdb movie	imdb movie	енглески	Acc=0,59 (резултат означен са Ashu)
(Lee & Renganathan, 2011)	product reviews	product reviews	кинески	1.без фазе претходног процесирања Acc=0,837 2.са претходним процесирањем Acc=0,870
(Johnson, Shukla & Shukla, 2012)	twitter	1.twitter 2.Gallup poll's data	енглески	Acc=0,60
(Saif, He & Alani, 2012)	1. twitter (STS) 2. Health CareReform (HCR)	1. twitter (STS) 2. Health CareReform (HCR)	енглески	1. са стоп речима Acc=0,807 без стоп речи Acc=0,775 2. са стоп речима Acc=0,711 без стоп речи Acc=0,737
(Evert, Proisl, Greiner & Kabashi, 2014)	twitter data + SMS data	twitter + live journal entries	енглески	Acc=0,725 (Message polarity classification)
(Mladenović, Mitrović, Krstev & Vitas, 2015)	**1.новинске вести (скуп А)**	**1.новинске вести (скуп А)** **2. оцене филмова**	**српски**	**1. Un+Bi+SWL+mapping Acc=0,792** **2.Un+Bi+mapping Acc=0,783**

У овој тези евалуација модела извршена је на оба начина: унакрсном валидацијом и независним скупом. Резултати које смо добили унакрсном валидацијом бољи су или у рангу са моделима генерисаним у неколико последњих година (Blinov et al., 2013; Wang et al., 2014). При примени независних скупова, резултати које смо добили слични су резултатима који се добијају овом методом на другим језицима. Треба истаћи да је метода ME осетљива на промену домена класификације у односу на домен над којим се врши обука, што се види из свих резултата ове групе у односу на унакрсну валидацију која користи исти домен у обуци и класификацији и што доводи

до нижих вредности свих статистичких показатеља у случају примене независних скупова за тестирање.

У досадашњем излагању разматрали смо текстове који садрже дословно значење. Систем за класификацију SAFOS обучаван је скупом за учење који не препознаје фигуративни језик. С друге стране, истраживања која су објавили Рејес и Роко (Reyes & Rosso, 2012b) показала су да се тачност класификације у задатку СА може знатно побољшати (са 54% на мах. 89.05%) када се укључе предиктори којима се препознаје фигуративни говор, у односу на скуп предиктора који текст третирају у дословном значењу. Рентоуми и њене колеге (Rentoumi et al., 2010) унапредили су СА методу машинског учења интегришући је са методом заснованом на правилима којом се, са већом тачношћу него основном методом, препознаје фигуративна употреба језика. У раду (Glasgow, Fink & Boyd-Graber, 2014) предложен је метод проналажења лексичких (тзв. мртвих) метафора ради унапређења процеса класификације твитова тако да је F1 мера побољшана са 0,61 на 0,86.

У следећем поглављу предлажемо структуре и методе развоја ресурса којима се може открити фигуративни говор у текстовима на српском језику и којима се може унапредити процес класификације према осећањима.

6. Препознавање фигуративног говора

Анализа текста у одређеним доменима, а нарочито при примени одређених стилова, захтева примену техника препознавања фигуративног говора. Фигуративни говор, односно употреба реторичких фигура, доводи до промене уобичајене или подразумеване употребе природног језика. Постоје бар четири разлога употребе фигура у свакодневног говору и писању:

- да се аудиторијум убеди у тврђења која се излажу
- да се изразе лична осећања аутора
- да се истакну другачија виђења неких чињеница
- да се аудиторијум забави и заинтересује за дату тему.

Применом реторичких фигура мења се дословно значење текста тако да оно може бити делимично или у потпуности промењено. Поступак аутоматске или полуаутоматске идентификације реторичких фигура у тексту може побољшати процесе класификације текста, анализе осећања, машинског превођења, ауторства, итд. На пример, традиционални систем класификације текста на основу осећања (Pang, Lee & Vaithyanathan, 2002) би на исти начин интерпретирао реченице „Он је брз као муња“ и „Он је брз као пуж“, док би укључењем процеса препознавања фигуративног говора друга реченица била посматрана као појава реторичке фигуре ироније, чиме би њен поларитет осећања био у потпуности промењен у односу на дословни.

Да бисмо изградили ефикасан систем аутоматског препознавања фигуративног говора, потребно је, најпре, формално дефинисати и описати реторичке фигуре, а да бисмо постојеће системе анализе осећања унапредили могућношћу препознавања измењеног значења текста, потребно је утврдити скуп фигура чијом се применом мења интензитет или поларитет осећања речи, фраза, реченица или већих текстуалних целина. У наставку овог поглавља описаћемо поступак изградње прве формалне доменске онтологије реторичких фигура у српском језику. Затим ћемо предложити поступак аутоматског генерисања онтологије задатака (в. одељак 3.4) која настаје интеграцијом доменске онтологије реторичких фигура и онтологије генерисане из семантичке мреже Српски ворднет. Описаћемо поступак

доградње Српског ворднета новом семантичком релацијом која омогућава проналажење једне групе реторичких фигура којима се постиже промена поларитета осећања у тексту, а на крају поглавља особине система који препознаје фигуративни говор.

6.1 Модели класификације реторичких фигура

Реторичке или стилске фигуре (енг. rhetorical devices, stylistic figures, figures of speech) примењиване су и истраживане још у античким временима. Према Аристотелу „Реторичар је онај који има способност убеђивања". Прву значајну класификацију реторичких фигура, познату као „*quadripartita ratio*", дали су Римљани. Она описује четири основне реторичке операције којима се граде све познате фигуре: додавање (*addition*), изостављање (*omission*), пермутација (*permutation*) и премештање (*transposition*). Антички реторичари су тврдили да у тексту све реторичке фигуре могу бити генерисане комбинацијом четири основне реторичке операције на различитим језичким нивоима: речи, фраза, реченица, пасуса и текстова већег обима од пасуса. Овакав поглед на природу и класификацију реторичких фигура задржан је и у средњем веку, када је Пичам[255] (Peacham, 1593) употребио реторичке операције: понављање (*repetition*), изостављање (*omission*), поделу (*separation*) и повезивање (*conjuction*).

Теорија реторичких фигура доживела је снажан напредак средином двадесетог века захваљујући групама истраживача Оулипо и Група μ (Bagić, 2012), а касније и радовима Лејкофа и Џонсона (Lakoff & Johnson, 1980), Ортонија (Ortony, Clore & Collins, 1990), Глуксберга (Glucksberg, 2001), Тарнера (Turner, 2002), Алм-Арвиус (Alm-Arvius, 2003), Гибса (Gibbs & Colston, 2012) и других, на основу којих је извршена свеобухватна анализа и класификација реторичких фигура. Са појавом система за развој формалних онтологија и унапређењем рачунарских алата у области рачунарске лингвистике, генеришу се различити системи формалне класификације

[255] Henry Peacham the Elder (1546–1634)

(Kelly, Abbott, Harris, Di Marco & Cheriton, 2010), базе података (Lönneker-Rodman, 2008), базе знања реторичких фигура (Harris & Di Marco, 2009) и софтверски алати за њихово препознавање и анотацију (Gawryjolek, 2009). Системи формалне класификације реторичких фигура користе разноврсне структуре података и различите критеријуме груписања. Тако је, на пример, Дуранд (Durand, 1987) увео релације између лингвистичких елемената: истоветност (*identity*), сличност (*similarity*), разлика (*difference*), супротност (*opposition*) и лажна хомологија (*false homologies*). Морис (Sloane, 2001) је креирао (семио-)синтактичку дводимензионалну класификациону табелу састављену од: лингвистичких операција и лингвистичких нивоа. Харис[256] (Harris, 2011) класификује све фигуре у три групе: „оних које се односе на наглашавање; оних које се односе на физичку организацију, транзицију и распоред; и оних које се односе на декорацију.“ Шатклиф[257] при класификацији користи шест група фигура: граматичких (*grammar*), семантичких (*meaning*), поредбених (*comparison*), раздвајајућих (*parenthesis*), понављајућих (*repetition*) и реторичких у ужем смислу (rhetoric). Општију класификацију даје Шварц[258] (Schwartz, 2009) чијом се поделом формирају три групе: фигуре говора (*figures of speech*), фигуре звука (*sounds*) и остале реторичке фигуре (*other rhetorical devices*). Једно од најсвеобухватнијих истраживања у новије време дала је Инкпот[259] група у оквиру *Rhetfig* пројекта. Према Кели[260] и њеним сарадницима (Kelly, Abbott, Harris, Di Marco & Cheriton, 2010), члановима Инкпот групе, постоје концептуално различити приступи реторичким фигурама, због чега они дефинишу три различита начина њихове класификације. Реторички приступ врши поделу према реторичким карактеристикама на тропе (*tropes*), схеме (*schemes*) и хрома (*chroma*) фигуре. Лингвистички приступ врши поделу фигура према лингвистичким карактеристикама у шест група концепата (фонолошких, морфолошких, синтаксних, лексичких, семантичких и ортографских).

[256] http://www.virtualsalt.com/rhetoric.htm
[257] Sutcliffe - http://opundo.com/figures.php
[258] Schwartz - http://cla.calpoly.edu/~dschwart/teaching.html
[259] http://create.uwaterloo.ca/matt/inkpot/projects/rhetorical_about.html
[260] Ashley R. Kelly

Когнитивно-концептуални приступ најпре дефинише лингвистичке технике које се користе, а затим врши поделу фигура према примењеним техникама које могу бити: понављање (*repetition*), изостављање (*omission*), понављање неке особине у низу (*series*), идентичност (*identity*), сличност (*similarity*), симетрија (*symmetry*) и супротност (*opposition*).

У теорији српске књижевности такође је дато више приступа у разврставању реторичких фигура. Према Тартаљи (Tartalja, 2003) оне се деле у три групе и то на: фигуре изговора (дикције), фигуре исказа и фигуре казивања. Фигуре исказа се деле на фигуре речи или тропе и фигуре везане за реченицу (фигуре конструкције). Друга значајна класификација може се наћи у теорији књижевности Солара (Solar, 1987). Он, пре свега, истиче комплексност процеса разврставања реторичких фигура која проистиче из неслагања реторичара и лингвиста у погледу броја фигура и интерпретације значења и функција појединих фигура (Solar, 1987, pp. 66). Ипак, он предлаже класификацију за коју каже да је најближа савременом схватању језика, према којој се све фигуре могу поделити у четири групе: фигуре дикције, фигуре речи или тропе, фигуре конструкције и фигуре мисли. Фигуре дикције, које се још називају и гласовне фигуре или звучне фигуре, заснивају се на дејству одређених гласова, односно звукова у говору. Понављање одређених гласова или група гласова, њихово изостављање или уметање на неочекиваним местима, опонашање одређених звукова и шумова из природе утиче на појачавање или смањење значаја лингвистичких структура над којима се изводи. Ова врста фигура, према аутору, нема утицаја на значење структура над којима се гради и не мења их, већ само наглашава њихово основно значење. Фигуре конструкције настају мењањем распореда речи у реченици или у некој другој већој целини (одломку, стиху и сл.) у односу на уобичајени тј. подразумевани, па се могу посматрати као фигуре поретка или распореда. Насупрот овим групама, фигуре речи или тропи имају задатак да промене основно значење неке речи. Фигурама мисли такође се мења основно значење лингвистичке структуре која је комплекснија од речи.

Очигледно, велики број начина класификације фигура из перспективе како реторике, тако и лингвистике указују на потребу да се, при изградњи

формалне онтологије, у обзир узму особености фигура посматраних са свих аспеката.

6.2 Изградња дескриптивне онтологије реторичких фигура у српском језику

Дескриптивна онтологија се, према Полију[261] (Poli, 2003) и Обрсту[262] (Obrst, 2010), бави колекцијама *prima facie* информација у одређеној области или шире, добијених на основу општеприхваћених друштвених појава, људске спознаје и интерпретације на природном језику. У наставку овог поглавља приказаћемо структуру и примену онтологије *RetFiguresOnto.* Реч је о формалној репрезентацији особина, начина и подручја деловања као и примера употребе реторичких фигура које се користе у српском језику. *RetFiguresOnto* је дескриптивна онтологија чија је сврха формална спецификација знања из домена фигуративне употребе језика (Mladenović & Mitrović, 2013).

Имајући у виду сложени и модуларни поступак изградње једне онтологије који укључује: прикупљање и организацију доменског знања (знања експерата из неког домена), дефинисање примене, опсега важења и грануларности онтологије, изградњу таксономије, дефинисање веза, ограничења и правила над ентитетима онтологије (Devedžić, 2002), рад на изградњи формалне онтологије реторичких фигура српског језика поделили смо у две фазе. У првој фази смо дефинисали скуп фигура и сакупили колекцију њихових инстанци у виду примера употребе у језику, како бисмо креирали доменско знање као основу за изградњу онтологије. У другој фази је сама онтологија развијена и припремљена за даље коришћење.

[261] Roberto Poli
[262] Leo Obrst

6.2.1 Формирање колекције реторичких фигура и примера њихове употребе у српском језику

У процесу изградње доменског знања, први корак се односио на креирање структуре базе података реторичких фигура *RetFig* чији су мета-подаци: *назив фигуре на српском*, одговарајући *назив на енглеском* (паралелизација фигура извршена је са аналогним ресурсом[263] енглеског језика Инкпот групе), *дефиниција или опис фигуре*, *етимологија имена*, додатна *напомена* и *примери употребе*. Свакој од фигура додељена су три класификациона податка: *реторички тип*, *лингвистички тип* и *врста лингвистичке операције*. У циљу прикупљања примера употребе и евентуалне корекције дефинисаних релација, креирали смо веб апликацију[264] којом се база података *RetFig* може допуњавати (прилог 6.1). У процесу допуне, као изворе примера фигура користили смо текстове белетристике, поезије и стручне текстове из области лингвистике (Bagić, 2012; Tartalja, 2003; Solar, 1987). На крају овог процеса дефинисано је 98 фигура[265] класификованих, према Солару, у четири реторичке групе: фигуре наглашавања, фигуре замене значења – тропи, фигуре конструкције и фигуре ширења и сужавања мисли. Типични представници групе фигура наглашавања у овој бази су: афереза, апокопа, алитерација, ономатопеја и др. У фигуре конструкције спадају: анаграм, анафора, палиндром, плеоназам и др. Репрезентативни примери тропи фигура су: метафора, метонимија, иронија, синегдоха, оксиморон, компарација и др. Представници групе фигура којима се исказане мисли могу проширити или сузити су: словенска антитеза, амплификација, лаконизам, парадокс и др. Све фигуре такође су подељене и у пет лингвистичких категорија. Ако је лингвистички елемент који учествује у изградњи фигуре слово, група слова или слог, реч је о групи фонолошких фигура. Уколико се фигура гради употребом различитих флективних или деривационих облика речи ради се о групи морфолошких фигура. Када се

[263] http://rhetfig.appspot.com/
[264] http://sm.jerteh.rs/MemberZone/RetFigS.aspx
[265] Листа свих фигура са примерима употребе и класификационим информацијама приказана је на веб страници http://sm.jerteh.rs/RetFig.aspx

фигуром мења уобичајени (подразумевани) поредак у реченици, мења врста речи којој припада нека реч, додају или уклањају делови реченице, онда таква фигура припада групи синтаксних фигура. Ако се фигуром мења дословно значење реченице или неког њеног дела, онда она припада групи семантичких фигура. Уколико се промена значења односи на лингвистичку целину већу од реченице, група је прагматска. Најзад, све фигуре настају применом неке лингвистичке операције над лингвистичким елементима. У овој тези предложили смо и применили осам операција: додавање, изостављање, понављање, премештање, спајање, раздвајање, замену и симетрију.

Тако дефинисана *RetFig* база може бити серијализована у XML формат (прилог 6.1) и може се преузети са странице веб апликације.[266]

6.2.2 Изградња онтологије *RetFiguresOnto*

Доменска онтологија (в. одељак 3.4) реторичких фигура српског језика *RetFiguresOnto.owl* изграђена је са циљем да:

- представља формално знање којим се једнозначно описују и дефинишу реторичке фигуре које се користе у српском језику
- да буде дељена и повезана са другим лингвистичким ресурсима и онтологијама као што су *SWNonto*, DOLCE и SUMO (в. одељке 3.3 и 3.4).
- да представља основ за изградњу онтологије задатака (в. одељак 3.4) која ће се користити у процесу онтолошке анотације реторичких фигура у српском језику.

Техника моделовања која је примењена у овом раду је техника. „одозго-наниже“ (*top-down*). Онтологија *RetFiguresOnto* креирана је помоћу алата *Protege*[267] 4.2, слободног софтвера отвореног кода за моделовање знања и изградњу онтологија. Коришћен је OWL *2 Web Ontology Language*. Како је реч о дескриптивној онтологији, њен раст није интензиван и карактерише га углавном пораст броја инстанци тј. појединачних фигура.

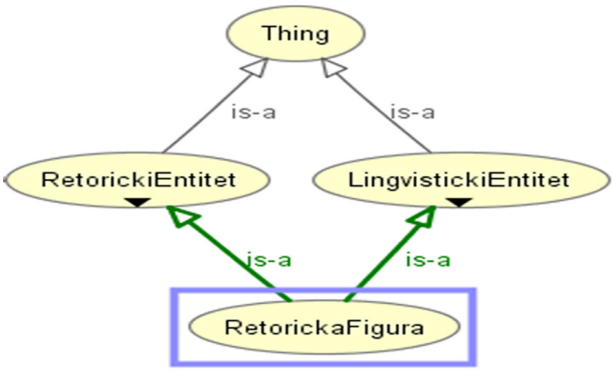
[266] http://sm.jerteh.rs/memberzone/RetFig.aspx
[267] http://protege.stanford.edu/

6.2.3 Таксономија онтологије *RetFiguresOnto*

Таксономија онтологије *RetFiguresOnto* садржи две главне класе:

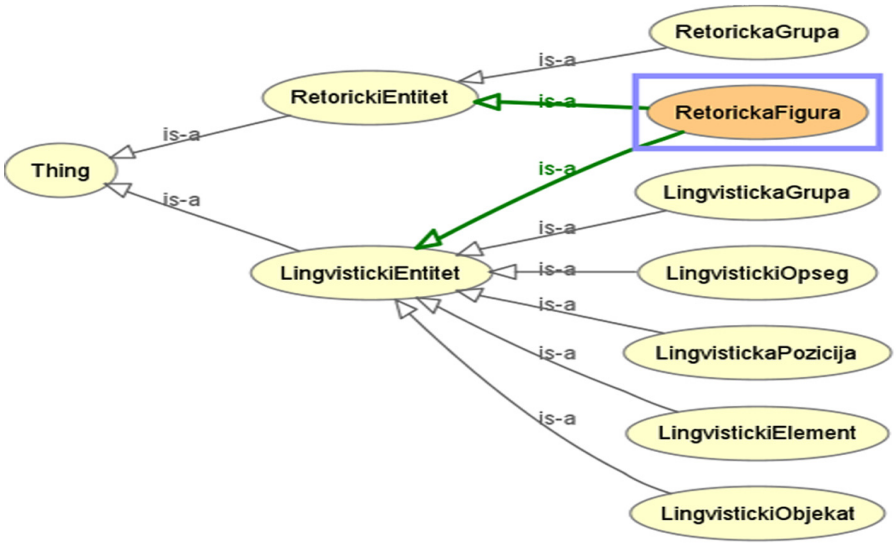
- RetorickiEntitet (*RhetoricalEntity*)
- LingvistickiEntitet (*LinguisticEntity*).

Класа *RetorickaFigura* (*RhetoricalFigure*) дефинисана је, узимајући у обзир сва претходна истраживања о природи реторичких фигура, истовремено и као лингвистички и као реторички концепт (слика 6.1).



Слика 6.1 Главне класе онтологије реторичких фигура

На нижем нивоу, *RetorickiEntitet* представљен је класама: *RetorickaGrupa* (*RhetoricalGroup*) и *RetorickaFigura* (*RhetoricalFigure*). Класа *RetorickaGrupa* основ је пресликавања на онтологије чији је предмет изучавања реторика. Чланови ове класе су: *фигуре наглашавања, фигуре конструкције, фигуре ширења и сужавања мисли, фигуре замене значења - тропи*. Класа *LingvistickiEntitet* представљена је концептима: *LingvistickaGrupa* (*LinguisticGroup*), *LingvistickiOpseg* (*LinguisticRange*), *LingvistickiObjekat* (*LinguisticObject*), *LingvistickaPozicija* (*LinguisticPosition*), *LingvistickiElement* (*LinguisticElement*) и *RetorickaFigura* (*RhetoricalFigure*). Класа *LingvistickaGrupa* основ је пресликавања на лингвистичке онтологије високог нивоа (*Upper level linguistic ontologies*) као што су: GOLD и DOLCE (в. одељак 3.4.2). Чланови ове класе означавају области науке о језику чији предмети проучавања су објекти трансформације садржани у реторичким фигурама. То су области лингвистике: *Фонологија, Морфологија, Синтакса, Семантика и Прагматика*. Таксономија онтологије приказана је на слици 6.2.



Слика 6.2 Таксономија доменске онтологије реторичких фигура

Класа *RetorickaFigura* је централна класа онтологије. Инстанце ове класе су у релацијама (*object properties*)[268] са инстанцама свих осталих класа. Такође, ова је класа декларисана и атрибутима (*class atributes*), описима (*descriptions*) и исказима (*assertions*). Декларација класе *RetorickaFigura* дата је у прилогу 6.2. Чланови ове класе су појединачне фигуре дефинисане и описане у бази *RetFig*. Да бисмо описали улогу осталих класа које су изведене из класе *LingvistickiEntitet* потребно је да објаснимо улогу појмова: лингвистички опсег, објекат, елемент, позиција и операција. У одељку 6.1 истакли смо да су још антички реторичари тврдили да се све реторичке фигуре могу генерисати комбинацијом четири основне реторичке операције на различитим језичким нивоима: речима, фразама, реченицама, пасусима и деловима текста већег обима од пасуса. Узимајући у обзир овај став и новија истраживања (Harris & Di Marco, 2009) и (Kelly, Abbott, Harris, Di Marco & Cheriton, 2010), сваку реторичку фигуру у тексту можемо дефинисати помоћу скупа ентитета (опсега, објекта, елемента као дела објекта, позиције елемента унутар објекта) и релације (лингвистичке операције). Лингвистички опсег или оквир (*linguistic scope*) текста (контекста) у коме

[268] Релација типа *object properties* представља бинарну релацију која се успоставља између две класе или две инстанце класа.

фигура делује може бити, почев од најмање лингвистичке целине речи, и шира целина: фраза, реченица, стих, строфа или пасус. Унутар такве целине, дефинишемо лингвистички објекат (*linguistic object*) чија трансформација помоћу лингвистичке операције (*linguistic operation*) представља одређену реторичку фигуру. Лингвистички објекат може бити реч, фраза, стих или реченица. Међутим, трансформација може бити извршена или над целим објектом или над неким његовим делом (деловима), па стога дефинишемо и лингвистички елемент (*linguistic element*) као део објекта који ће бити под дејством операције. У случају када се трансформација врши над целим објектом, објекат и елемент су идентични. Лингвистичка позиција (*linguistic position*) дефинише положај елемента унутар објекта. Однос описаних ентитета приказаћемо на примеру фигуре *апхереза* (слика 6.3). Пример употребе *аферезе* генерисан[269] је упитом над Корпусом савременог српског језика (в. одељак 3.1) у облику конкорданце и представља, по дефиницији, појаву изостављања првог слова у некој речи:

… vršio, ali njegov otac je sumnjičavo vrteo glavom i često mu govorio: "Od crtanja se, sine, ti ***leba*** *najesti nećeš". Ali, jedan od prvih poslova koje je Živa dobio bio je da uradi veliki mural …*

Слика 6.3 Однос лингвистичког опсега, објекта, елемента и позиције, повезаних лингвистичком операцијом - на примеру фигуре *апхереза*

Анализом процеса настанка реторичких фигура, утврдили смо да међусобни однос описаних лингвистичких ентитета дефинише појаву

[269] Извор конкорданце: *Politikin magazin (2002). Beograd: Politika novine i magazini*.

конкретне фигуре. На пример, када је лингвистички објекат *reč*, лингвистички елеменат *slovo*, а операција *seIzostavlja*, тада се могу генерисати реторичке фигуре:

- афереза – ако се изостави прво слово (лингвистичка позиција је *početak*)
- апокопа – ако се изостави последње слово (лингвистичка позиција је *kraj*)
- синкопа – ако се изостави слово које није ни прво ни последње (лингвистичка позиција је *sredina*).

У општем случају, лингвистичка позиција може описивати позицију када се:

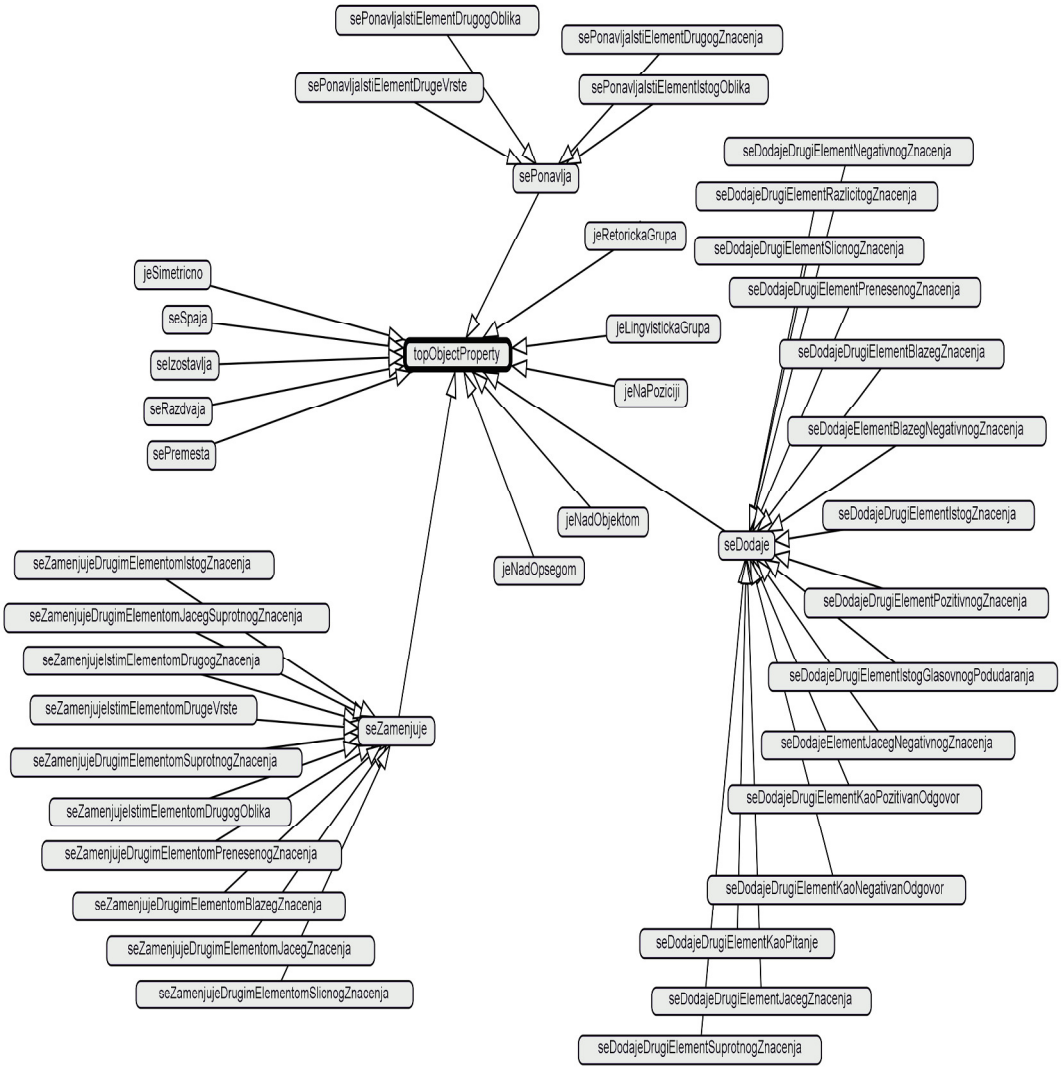
- лингвистички елемент налази на почетку лингвистичког објекта
- лингвистички елемент налази у средини лингвистичког објекта
- лингвистички елемент налази на крају лингвистичког објекта
- лингвистички елемент налази на почетку и на крају лингвистичког објекта
- лингвистички елемент налази на произвољној позицији унутар лингвистичког објекта
- лингвистички елемент налази на произвољној позицији унутар лингвистичког објекта, али се операција трансформације протеже на низ сукцесивних елемената у датом објекту
- лингвистички елемент налази на произвољној позицији унутар лингвистичког објекта, али се операција трансформације протеже на низ сукцесивних елемената у датом објекту, тако да се постиже симетричност.

6.2.4 Дефинисање релација, ограничења и правила

Лингвистичке операције у онтологији представљамо релацијама које повезују инстанце класе *RetorickaFigura* као домена релације (*Domain*) са инстанцама класе *LingvistickiElement* као опсега вредности (*Range*). Основна подела лингвистичких операција, па самим тим и релација у онтологији *RetFiguresOnto* је на релације: *seDodaje*, *seIzostavlja*, *sePonavlja*, *sePremesta*, *seSpaja*, *seRazdvaja*, *seZamenjuje* и *jeSimetricno*. За све релације дефинисали смо одговарајуће карактеристике релација (*Object Properties Characteristics*) – функционалност, рефлексивност, симетричност и транзитивност. Дефинисали смо ограничења (*Disjoin Properties*)[270] као и одговарајуће даље поделе релација *sePonavlja, seDodaje* и *seZamenjuje* на подрелације (*subProperties*) јер представљају везе комплексне структуре. Програмски додатак алата *Protege* за графичку репрезентацију, *Graf PG ETI Sova*[271], употребили смо да би представили комплетну структуру релација у онтологији *RetFiguresOnto* на слици 6.4. Са приказаног графа се може уочити да је, поред лингвистичких операција, уведено и неколико других које повезују инстанце класе *RetorickaFigura* као домена релације (*Domain*) са инстанцама осталих класа: релација *jeLingvistickaGrupa* са инстанцама класе *LingvistickaGrupa*, релација *jeRetorickaGrupa* са инстанцама класе *RetorickaGrupa*, релација *jeNaPoziciji* са инстанцама класе *LingvistickaPozicija*, релација *jeNadObjektom* са инстанцама класе *LingvistickiObjekat* и релација *jeNadOpsegom* са инстанцама класе *LingvistickiOpseg.* Пример декларације релације *jeLingvistickaGrupa* дат је у наставку:

Declaration(ObjectProperty(ont:jeLingvistickaGrupa))
SubObjectPropertyOf(ont:jeLingvistickaGrupa owl:topObjectProperty)
AsymmetricObjectProperty(ont:jeLingvistickaGrupa)
IrreflexiveObjectProperty(ont:jeLingvistickaGrupa)
ObjectPropertyDomain(ont:jeLingvistickaGrupa ont:RetorickaFigura)
ObjectPropertyRange(ont:jeLingvistickaGrupa ont:LingvistickaGrupa)

[270] *Disjoin Properties* представља ограничење којим се онемогућава да било које две инстанце класа могу бити истовремено повезане обема релацијама за које важи ово ограничење - чиме онемогућавамо да једна реторичка фигура буде, на пример, и у релацији *seDodaje* и у релацији *seIzostavlja* са истим инстанцом неке класе (нпр. инстанцом која представља реч).
[271] http://protegewiki.stanford.edu/wiki/SOVA

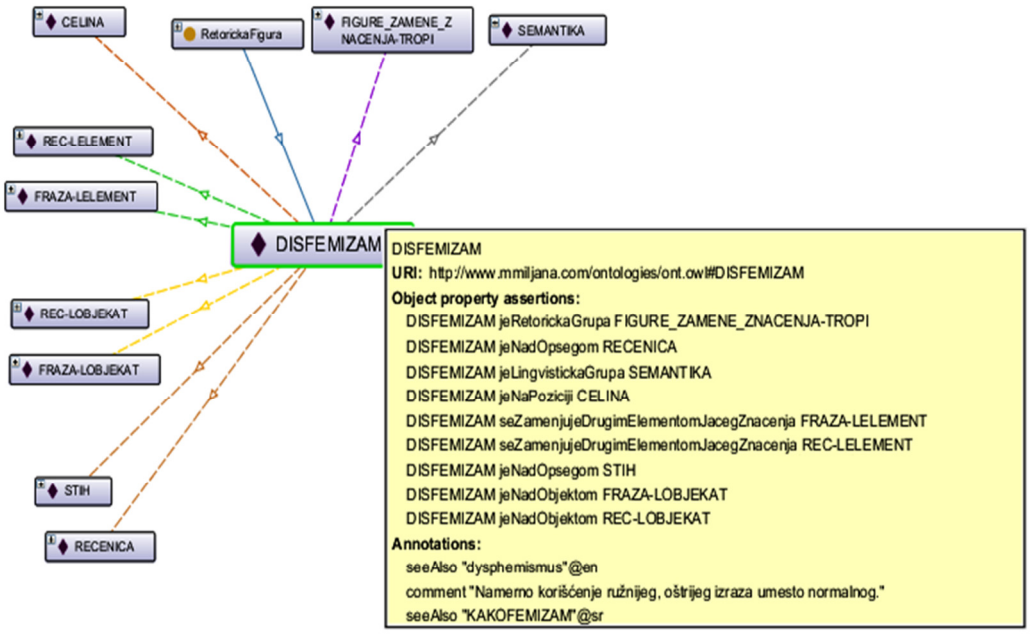


Слика 6.4 Хијерархија релација у онтологији *RetFiguresOnto*

6.2.5 Креирање атрибута и унос инстанци класа

Након дефинисања релација (*Object Properties*), дефинисали смо и инстанце (*Individuals*), тј. чланове свих класа. Онтологија *RetFiguresOnto* описује 98 реторичких фигура у облику инстанци класе *RetorickaFigura*. За сваку од њих дефинисано је којој реторичкој и којој лингвистичкој групи припада. Дефинисани су: лингвистички опсег, објекат, елемент и позиција елемента који учествују у креирању дате фигуре. Свака фигура добила је одговарајуће анотације: кратку дефиницију (*comment*), информацију о називу

реторичке фигуре на енглеском језику која би помогла у процесу поравнања са аналогном онтологијом (Harris & Di Marco, 2009) енглеског језика (*owl:sameAs*), информацију о алтернативном називу исте реторичке фигуре на српском. На слици 6.5 дат је, као пример, *OntoGraf*[272] приказ реторичке фигуре *дисфемизам* са свим дефинисаним везама у *RetFiguresOnto*, одакле се може добити потпуна информација о томе шта ова фигура представља. Из анотација се сазнаје да је реч о намерном коришћењу ружнијег, оштријег израза уместо уобичајеног. Сазнаје се да је назив одговарајуће фигуре у енглеском језику *dysphemismus*, а да на српском постоји и алтернативни назив *какофемизам*. *Дисфемизам* је реторичка фигура из групе тропи, њоме се бави област лингвистике – семантика. Може се наћи унутар реченице или стиха (лингвистички опсег) и добија се тако што се постојећа фраза или реч (лингвистички објекат/елемент) у реченици, у целини (лингвистичка позиција) замени другом фразом или речју јачег значења (лингвистичка операција – „*seZamenjujeDrugimElementomJacegZnacenja*“).



Слика 6.5 Пример декларације реторичке фигуре дисфемизам *RetFiguresOnto*

[272] *Protege* додатак за визуелну репрезентацију онтологија. https://github.com/protegeproject/ontograf

Декларација инстанце *дисфемизам* дата је у наставку:

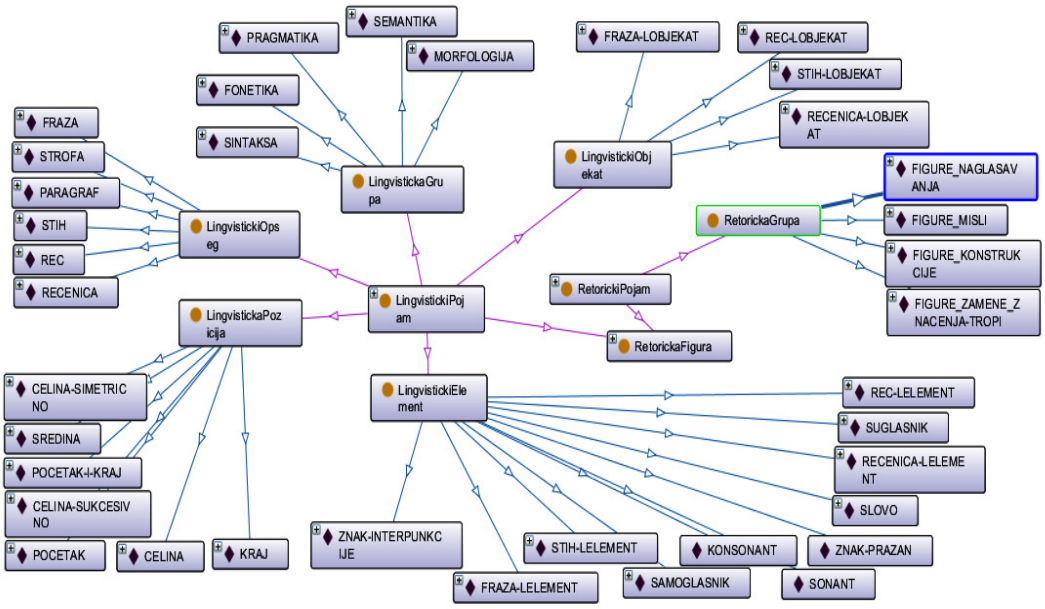
```
<owl:NamedIndividual rdf:about="&ont;DISFEMIZAM">
<rdf:type rdf:resource="&ont;RetorickaFigura"/>
   <ont:naziv rdf:datatype="&xsd;string">DISFEMIZAM</ont:naziv>
   <rdfs:comment>Namerno korišćenje ružnijeg, oštrijeg izraza umesto
normalnog.</rdfs:comment>
   <rdfs:seeAlso xml:lang="en">dysphemismus</rdfs:seeAlso>
   <rdfs:seeAlso xml:lang="sr">KAKOFEMIZAM</rdfs:seeAlso>
   <ont:jeNaPoziciji rdf:resource="&ont;CELINA"/>
   <ont:jeRetorickaGrupa rdf:resource="&ont;FIGURE_ZAMENE_ZNACENJA-TROPI"/>
   <ont:seZamenjujeDrugimElementomJacegZnacenja rdf:resource="&ont;FRAZA-LELEMENT"/>
   <ont:jeNadObjektom rdf:resource="&ont;FRAZA-LOBJEKAT"/>
   <ont:seZamenjujeDrugimElementomJacegZnacenja rdf:resource="&ont;REC-LELEMENT"/>
   <ont:jeNadObjektom rdf:resource="&ont;REC-LOBJEKAT"/>
   <ont:jeNadOpsegom rdf:resource="&ont;RECENICA"/>
   <ont:jeLingvistickaGrupa rdf:resource="&ont;SEMANTIKA"/>
   <ont:jeNadOpsegom rdf:resource="&ont;STIH"/>
</owl:NamedIndividual>
```

Ако се, за тренутак, вратимо серијализацији базе података *RetFig*, односно XML документу (прилог 6.1), можемо размотрити неке од примера употребе фигуре дисфемизам у свакодневном говору, како је приказано у наредној структури.

```
<RETFIG>
    <figure>
        <id>4963449e-182f-44ad-a4ac-0c2d8f3efb12</id>
        <name> DISFEMIZAM (KAKOFEMIZAM)</name>
        ...
        <examples>
            <example>On je već hrana crvima</example>
            <example>Automatski  foto aparat - idiot</example>
            <example> Ta džukela laje po ceo dan. (džukela - pas)</example>
            <example> On ne prestaje da loče iako mu je doktor zabranio. (loče -
            pije, konzumira alkohol </example>
        </examples>
    </figure>
...
</RETFIG>
```

Дисфемизам се на основу онтолошке дефиниције лако препознаје. У првом примеру, реч „умро“ замењује се фразом јачег, оштријег значења „храна црвима“. Исто понашање се може показати и у осталим примерима, где се именица *пас* замењује именицом *џукела*, глагол *пије* глаголом *лоче*, а фраза „аутоматски фото-апарат“ именицом *идиот*.

На слици 6.6 дат је одговарајући *OntoGraf* приказ класа и индивидуа које смо досад описали, а чине таксономију онтологије реторичких фигура.



Слика 6.6 Таксономија класа и чланова класа онтологије *RetFiguresOnto*

Графичка репрезентација чланова класе *RetorickaFigura*, која је због великог броја чланова изостављена из приказа на слици 6.6 дата је у прилозима 6.3 и 6.4.

6.2.6 Примена онтологије *RetFiguresOnto*

Онтологија *RetFiguresOnto* може дати двојаке врсте одговора. Пре било какве анализе текста, за унапред дати лингвистички опсег или лингвистички објект, она може указати на то које се реторичке фигуре могу очекивати у дефинисаном опсегу и над којим деловима текста. На пример, ако бисмо анализирали појединачне речи текста, што би било еквивалентно лингвистичком опсегу "REC" (в. слику 6.6), онда нема смисла очекивати појаву реторичке фигуре нпр. дисфемизам, али има смисла очекивати аферезу, апокопу, дијарезу, протезу, синкопу, итд. (в. одељак 6.2.3). SPARQL упит који даје одговор на основу датог примера представљен је на слици 6.7.

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ont: <http://www.mmiljana.com/ontologies/ont.owl#>
SELECT DISTINCT ?subject
        WHERE { ?subject  ont:jeNadOpsegom  ?opseg.
                ?opseg ont:naziv ?nazivOpsega.
                FILTER (?nazivOpsega = "REC") }
```

SIMBOL
PROTEZA
PERIFRAZA
APOKOPA
AFEREZA
ANAGRAM
EPENTEZA
DIJAREZA
METATEZA
SINKOPA

Слика 6.7 Реторичке фигуре које се граде над речима

С друге стране, анализирајући неки текст уз помоћ онтологије можемо утврдити које се лингвистичке операције појављују у њему, па самим тим и о којим је фигурама реч. Примера ради, ако смо утврдили да у неком тексту долази до губитака слова у речима, пресликавањем на онтолошку релацију *seIzostavlja* долазимо до скупа реторичких фигура које се граде на овај начин: аферезе, синкопе, апокопе и елизије. Резултати тог упита дати су на слици 6.8.

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ont: <http://www.mmiljana.com/ontologies/ont.owl#>
SELECT DISTINCT ?figura
        WHERE { ?figura  ont:selzostavlja  ?element.
                ?element ont:naziv ?naziv.
                FILTER (?naziv = "SLOVO") }
```

figura
SINKOPA
ELIZIJA
AFEREZA
APOKOPA

Слика 6.8 SPARQL упит којим се идентификују реторичке фигуре које се формирају изостављањем слова у речима

6.3 Аутоматско генерисање онтологије задатака реторичких фигура у српском језику

Дескриптивну онтологију *RetFiguresOnto* можемо употребити да дефинишемо формална правила којима се врши одабир подскупа фигура за изградњу онтологије задатака *SemRetFig*. У наставку овог одељка показаћемо како се онтологија задатака *SemRetFig* може аутоматски креирати применом правила у онтологијама *RetFiguresOnto* и *SWNOnto*. Онтологију *SemRetFig* креирамо да бисмо је применили у задацима:

- онтолошке анотације реторичких фигура у српском језику
- препознавања фигуративног говора
- побољшања система класификације осећања текстова на српском језику.

Аутоматско препознавање реторичких фигура у тексту и њихова анотација није ново поље интересовања у обради природног језика. Модел препознавања једне групе реторичких фигура, као и одговарајући алат за њихову аутоматску анотацију (*Java Annotation Tool Of Rhetoric* - JANTOR) предложени су у раду (Gawryjolek, 2009). Алат користи комбинацију неколико различитих стратегија: лексички парсер заснован на вероватносној контекстно-независној граматици, хијерархијски парсер ради израчунавања степена различитости између слично структуираних фраза, Портеров стемер, типизиране зависне граматике и семантичку мрежу ворднет, како би генерисао синтаксне обрасце (*syntactic patterns*) који карактеришу сваку од посматраних фигура. На пример, анафора је фигура коју карактерише понављање једне или више речи на почетку сваке у низу од две или више узастопних реченица (или стихова). У JANTOR-у се дефинише синтаксни образац облика

$$< [W_a] \ldots >< [W_a] \ldots >$$

којим се проналази појава ове фигуре, где су <...> границе реченице или стиха, а W_a једна или више речи на њиховом почетку. Резултати примене JANTOR-а показују да се реторичке фигуре из класа фигура конструкције и наглашавања успешно проналазе на овај начин, јер је могуће дефинисати синтаксне обрасце или регуларне изразе који их једнозначно одређују. Обе

ове класе фигура засноване су на синтаксним или морфолошким операцијама у склопу граматичких правила природног језика. Међутим структура реторичких фигура у класама фигура замене значења и проширења мисли је таква да оне немају једнозначан образац примене на нивоу синтаксе. У овој тези истражујемо методе идентификације фигура из групе фигура замене значења (тропи). Реторичке фигуре из ове класе своју улогу заснивају на:

- замени значења посматране речи или фразе значењем неке друге речи или фразе
- додели вишеструког значења посматраној речи или фрази
- генерисању новог значења посматране речи или фразе.

У таквим случајевима није могуће дефинисати морфолошко-синтаксне шаблоне, али је могуће утврдити какве семантичке релације постоје између учесника који граде одређену фигуру и на основу тога дефинисати правила закључивања која се могу применити над семантичким мрежама каква је ворднет или над онтологијама проистеклим из одговарајућих семантичких мрежа. У овој тези приказаћемо поступак аутоматске изградње онтологије једне групе реторичких фигура из класе тропи (онтологија *SemRetFig*) коју чине оне фигуре чија се улога заснива на замени основног значења речи или фразе унутар једне лингвистичке целине која може бити: реч, фраза, стих или реченица. Метода аутоматске изградње ове онтологије, која је по својој природи и намени онтологија задатака, садржи две фазе:

- изградњу основне таксономије класа
- проширење онтологије инстанцама класа.

Таксономију класа генеришемо упитом над доменском онтологијом *RetFiguresOnto*, а инстанце тих класа формирају се из *SWNonto* онтологије употребом правила за њихово генерисање (слика 6.9).

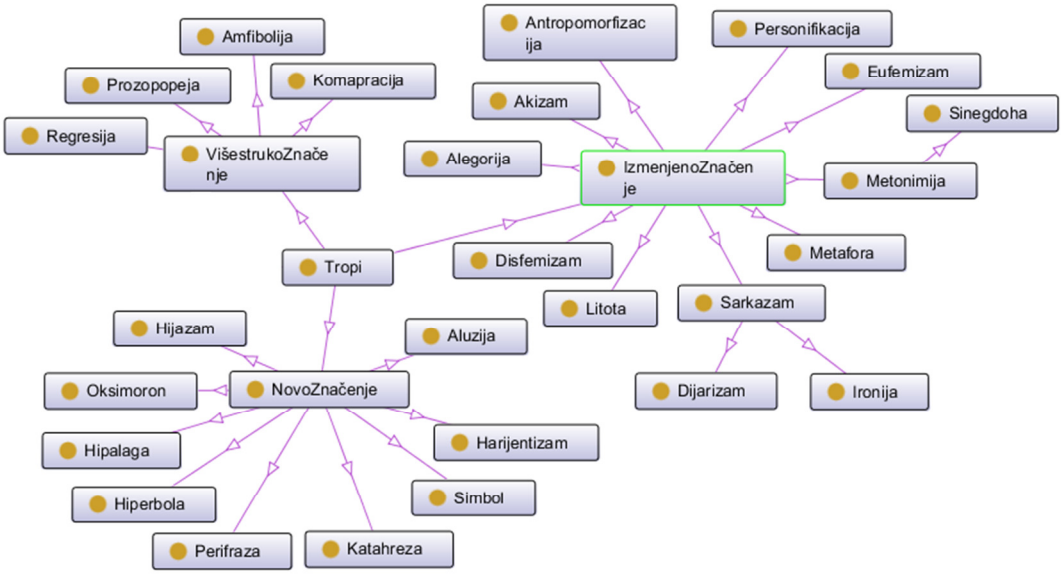
Слика 6.9 Аутоматско генерисање онтологије *SemRetFig* из онтологија *RetFiguresOnto* и *SWNonto*

Фигуре које чине основни скуп класа онтологије *SemRetFig* добили смо SPARQL упитом над онтологијом *RetFiguresOnto*.

```
select distinct ?figura
where { ?figura ont:jeNadObjektom ?objekt.
            ?figura ont:jeRetorickaGrupa ?retGrupa.
            ?objekt ont:naziv ?nazivObjekta.
            ?retGrupa ont:naziv ?nazivRetGrupe.
FILTER (?nazivObjekta ="REC"  &&
            ?nazivRetGrupe="FIGURE_ZAMENE_ZNACENJA-TROPI")}
order by ?figura
```

Тако смо дошли до скупа од 26 реторичких фигура које имају заједничку особину да припадају групи фигура које карактерише замена (?nazivRetGrupe="FIGURE_ZAMENE_ZNACENJA-TROPI") уобичајеног значења једне речи (?nazivObjekta ="REC") значењем друге речи или фразе. Али, тој групи припадају, такође, и оне фигуре које карактерише појава да се једној речи уз њено основно значење (чиме се оно појачава или смањује) придодаје и значење неке друге речи или фразе на основу неке њихове међусобне семантичке релације, као и оне фигуре које једној речи дају потпуно ново значење. Идеја је да се свака од тако селектованих фигура дефинише као класа нове онтологије задатака *SemRetFig* чије ће се инстанце (енг. individuals) генерисати аутоматски, применом правила (изражених помоћу *Semantic Web Rule Language* - SWRL)[273] над онтологијом *SWNonto*. На слици 6.10 приказана је таксономија онтологије *SemRetFig*.

[273] Semantic Web Rule Language - SWRL је језик који се користи у области семантичког веба за представљање формалних логичких израза добијен комбиновањем особина језика OWL DL (Web Ontology Language for Description Logic) и језика RML (Rule Markup Language).



Слика 6.10 Таксономија онтологије *SemRetFig*

На примеру фигуре *ирониja* објаснићемо начин изградње правила над онтологијом *SWNonto* помоћу којих се могу генерисати инстанце класе *Ironija* у онтологији *SemRetFig*. Ирониja је појава речи или израза којим се једно каже, а супротно мисли и разуме. Право значење речи прикривено је и супротно дословном значењу израза. Примери ирониje заснивају се на неколико врста семантичких релација (Bagić, 2012), а често коришћени облици у српском језику заснивају се на употреби придева коме се даје значење њему подразумеваног супротног (антонимског) придева као у примерима:

„*Он је баш генијалан*!“ (Скривено значење је тврдња да је неко заправо глуп.)

„*Види се да је жгољав!*“ (Скривена тврдња да је неко дебео.)

Друга форма се *састоји* од везе именице и придева, али се придеву додељује скривено значење оног придева који се уобичајено везује уз наведену именицу.

„*Брз је као корњача.*“

„*Храбар је као зец!*“

Вејл и Хао (Veale & Hao, 2009) користе још и назив „*иронична компарација*“ како би описали специфичну врсту ирониje изведену из поређења. У

наведеним примерима придев *брз* добија скривено значење придева који представља природну карактеристику именице *корњача*, а то је придев *спор*. Придев *храбар* добија скривено значење придева који је природна карактеристика именице *зец*, а то је придев *плашљив*. Семантичке релације које постоје у онтологији *SWNonto* и које користимо при генерисању кандидата за инстанцирање класе *Ironija* засноване на структури *„иронична компарација“* су пар инверзних релација *specificOf/specifiedBy* [355] којима се повезују инстанце класе именичких синсетова са инстанцама синсетова њима специфичних придева и релација *near_antonym* којом се међусобно повезују инстанце класа синсетова придева. Стога се правило за генерисање ироничних компарација као инстанци класе *Ironija,* примењено над онтологијом *SWNonto* може изразити као OWL правило

$$\{?n \ : specifiedBy \ ?p1. \ ?p1 \ : near\_antonym \ ?p2\} => \{?n \ : irony \ ?p2\}$$

или у форми *SWRL* правила

$$specifiedBy(?n, ?p1), near\_antonym(?p1, ?p2) \ \rightarrow \ Irony(?n, ?p2) \qquad (84)$$

Инстанце класе *Ironija* могу се добити и SPARQL упитом (слика 6.11) над онтологијом *SWNonto*. Резултат упита даје кандидате за инстанцирање класе *Ironija* у форми пара (adjective2, noun), нпр. (храбар, зец), али даје и инстанце класе *Poredjenje* које чине парови (adjective1, noun), нпр. (плашљив, зец).

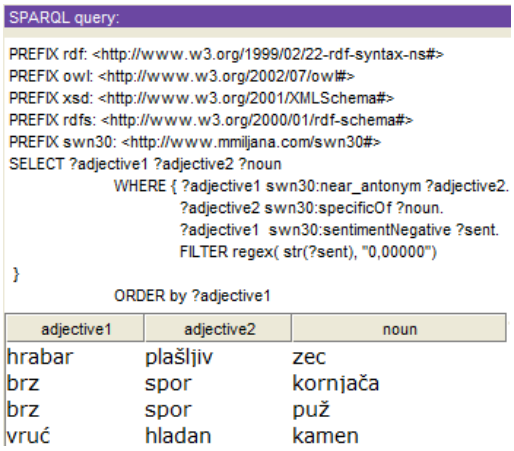
SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swn30: <http://www.mmiljana.com/swn30#>
SELECT ?adjective1 ?adjective2 ?noun
        WHERE { ?adjective1 swn30:near_antonym ?adjective2.
                ?adjective2 swn30:specificOf ?noun.
}
order by ?adjective1
```

adjective1	adjective2	noun
hrabar	plašljiv	zec
brz	spor	kornjača
brz	spor	puž
spor	brz	strela
spor	brz	ideja
lak	težak	slon
vruć	hladan	kamen

Слика 6.11 Кандидати за инстанцирање класа *Ironija* и *Poredjenje*

Резултат упита аналоган је резултату закључивања на основу правила (84), али се са слике 6.11 може уочити да у природном језику није уобичајено коришћење неких од кандидата за инстанцирање чланова класе *Ironija*. На основу интуиције коју у овом тренутку не доказујемо, може се рећи да су природни кандидати класе *Ironija* они код којих је поларитет осећања субјекта (?adjective1) у RDF тројци[274] *?adjective1 swn30:near_antonym ?adjective2* – позитиван. На пример, природнији примери ироније исказани су тврдњама: „*Брз је као корњача*“ или „*Лак је као слон*“ него „*Спор је као стрела*“, или „*Тежак је као перце*“. У том смислу, претрага (слика 6.11) се може унапредити додатним условом који не узима у обзир оне RDF тројке у којима је вредност негативног поларитета осећања субјекта већа од нуле. У циљу идентификације поларитета осећања субјекта, упит се мора проширити контролом својства податка (data property) *swn30:sentimentNegative*, а на основу вредности тог својства бирају се (филтрирају) само они елементи код којих је вредност тог својства једнака нули (слика 6.12).



SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swn30: <http://www.mmiljana.com/swn30#>
SELECT ?adjective1 ?adjective2 ?noun
        WHERE { ?adjective1 swn30:near_antonym ?adjective2.
                ?adjective2 swn30:specificOf ?noun.
                ?adjective1  swn30:sentimentNegative ?sent.
                FILTER regex( str(?sent), "0,00000")
}
        ORDER by ?adjective1
```

adjective1	adjective2	noun
hrabar	plašljiv	zec
brz	spor	kornjača
brz	spor	puž
vruć	hladan	kamen

Слика 6.12 Побољшање претраге кандидата за инстанце класе *Иронија*

Размотримо сада и структуру реторичке фигуре *оксиморон* која, по дефиницији, представља спајање појмова супротних значења у нови појам и

[274] RDF је модел података који користи форму тројке или триплета „субјекат-предикат-објекат“ (subject–predicate–object) да би описао ресурсе семантичког веба, обезбедио складиштење података у графовским базама података и представљање знања у онтолошким моделима.

чији су примери у *RetFig* бази: *виртуелна стварност*, *гласна тишина*, *луда памет*, *живи фосил*, *ватрени лед,* и др. Правила за генерисање индивидуа класе *Oksimoron*, примењена над онтологијом *SWNonto* могу се изразити као:

$$\{?p1 : near\_antonym\ ?p2.?p2 : derived-pos\ ?n\} => \{?p1 : oxymoron\ ?n\}$$
$$\{?p1 : near\_antonym\ ?p2.?p2 : be\_in\_state\ ?n\} => \{?p1 : oxymoron\ ?n\}$$

или у форми SWRL правила:

$$near\_antonym(?p1,?p2), derived-pos(?p2,?n) \rightarrow oxymoron(?p1,?n)$$
$$near\_antonym(?p1,?p2), be\_in\_state(?p2,?n) \rightarrow oxymoron(?p1,?n)$$

Инстанце класе *Oksimoron* могу се добити и SPARQL упитом (слика 6.13) над онтологијом *SWNonto*. Резултат упита даје кандидате за инстанцирање класе *Oksimoron* у форми пара (adjective1, noun), нпр. (гласан, тишина). Са слике 6.13 може уочити да RDF тројке које садрже антонимске придеве постају кандидати за инстанцирање ове класе, када задовољавају услов постојања релације *derived-pos* или *be_in_state*.

SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swn30: <http://www.mmiljana.com/swn30#>
SELECT ?adjective1 ?adjective2  ?noun
        WHERE {
        ?adjective1 swn30:near_antonym ?adjective2.
        {
            { ?adjective2 swn30:derived-pos ?noun.      }
             UNION
            { ?adjective2 swn30:be_in_state ?noun.      }
        }
        }
```

adjective1	adjective2	noun
ružan	lep	lepota
virtualan	stvaran	stvarnost
spor	brz	brzina
vruć	hladan	temperatura
glasan	tih	tišina

Слика 6.13 Кандидати за инстанцирање класе *Oksimoron*

Поред лексичко-семантичких релација које могу бити индикатор постојања одређених тропи фигура, могу се користити и имплицитне релације које постоје у ворднету. На пример, релација синонимије је имплицитна релација којом се могу проналазити кандидати за инстанцирање класе *Perifraza*, фигуре којом се један појам описује или замењује са више речи коришћењем неких битних особина тог појма (слика

6.14). Тако се реч *Pariz* може заменити синонимима: „*Grad svetlosti*“, „*Prestonica Francuske*“, „*Glavni grad Francuske*“, па парови попут (Pariz, Grad svetlosti) постају кандидати за инстанце ове класе.

SPARQL query:

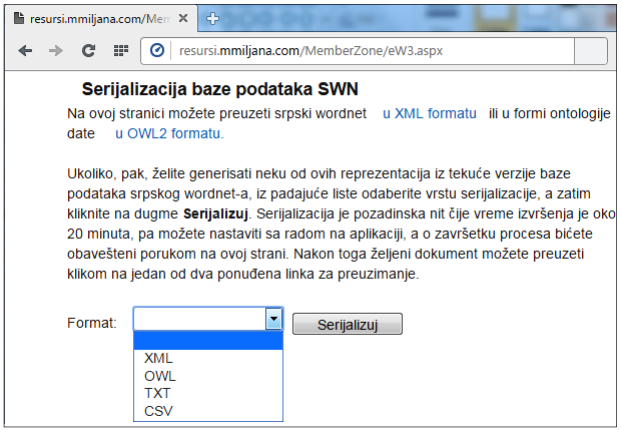
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swn30: <http://www.mmiljana.com/swn30#>
SELECT ?noun ?synonymNoun
        WHERE { ?noun swn30:hasNounWord  ?synonym.
        ?noun rdfs:label  ?os.
        ?synonym swn30:literal ?synonymNoun
        FILTER (CONTAINS ( UCASE(str(?os)), "PARIZ") ) }
```

noun	synonymNoun
Pariz	"prestonica Francuske"
Pariz	"glavni grad Francuske"
Pariz	"Grad svetlosti"
Pariz	"Pariz"

Слика 6.14 Кандидати за инстанцирање класе *Perifraza*

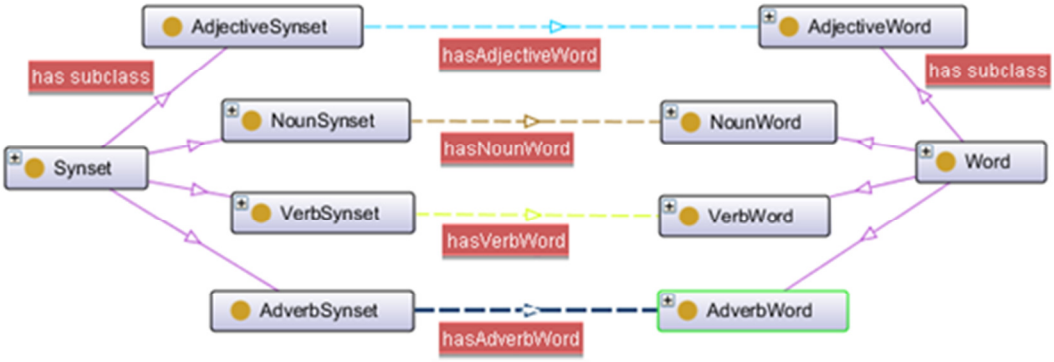
6.4 **Структура онтологије *SWNonto***

Онтологија *SWNonto* се генерише аутоматски из семантичке мреже Српски ворднет (SWN). У одељку 3.3.3 описали смо структуру, а у одељку 3.5 софтверски алат SWNE помоћу кога се Српски ворднет развија. Овај алат, зависно од задатака, равноправно користи три формата (OWL, XML и TXT) за репрезентацију SWN, па је могуће серијализовати садржај ворднета како је то приказано на слици 6.15.



Слика 6.15 Формати серијализације семантичке мреже Српски ворднет

Серијализацијом у OWL формат генерише се онтологија *SWNonto* чија је таксономија класа заснована на ван Асемовом моделу[275] (van Assem, Gangemi & Schreiber, 2006). На врху хијерархијске структура налазе се две класе – класа синсетова (*Synset class*) и класа речи (*Word class*). Обе класе садрже скуп подкласа које се формирају на основу врсте речи и међусобно су повезане релацијама изведених из релације *hasWord*. Таксономија класа онтологије *SWNonto* приказана је на слици 6.16.



Слика 6.16 Таксономија онтологије *SWNonto*

У погледу релација *SWNonto* садржи три врсте од којих две представљају релације између самих објеката (*object properties*), тј. бинарне релације између инстанци двеју класа, а трећа врста је релација доделе типа података (*datatype property*) и успоставља се између инстанце класе и типа података. У онтологији *SWNonto* бинарне релације се успостављају између инстанци класе *Synset* и између инстанци класе *Word*. Дефинисано је укупно 18 релација ове врсте и оне су приказане у табели 6.1. Релације доделе типа података успостављају се између инстанци класа *Synset* или *Word* и типова података дефинисаних у оквиру спецификације језика за опис XML схеме.[276] Дефинисано је укупно 9 релација овог типа и оне су дате у табели 6.2.

[275] в. одељак 3.4.1
[276] http://www.w3.org/TR/xmlschema11-2/

Табела 6.1 Бинарне релације у онтологији *SWNonto*

Relation	Type	Domain	Range	Tran-sitive	Sym-metric	Refle-xive	Inverse operation
hypernym	S → S	N, V, Aj, Av	N, V, Aj, Av	⊤	⊥	⊥	hyponym
near_antonym	S ↔ S	N, V, Aj, Av	N, V, Aj ,Av	⊥	⊤	⊥	near_antonym
holo_part	S → S	N	N	⊤	⊥	⊥	part_meronym
holo_portion	S → S	N	N	⊤	⊥	⊥	portion_meronym
holo_member	S → S	N	N	⊤	⊥	⊥	member_meronym
holo_substance	S → S	N	N	⊤	⊥	⊥	substance_meronym
subevent	S → S	V	V	⊥	⊥	⊥	⊥
cause	S → S	V	V	⊥	⊥	⊥	⊥
verbGroup	S ↔ S	V	V	⊥	⊤	⊥	⊥
participle	S ↔ S	Aj, V	V, Aj	⊥	⊥	⊥	⊥
similar_to	S ↔ S	Aj	Aj	⊥	⊤	⊤	similar_to
also_see	S ↔ S	V, Aj	V, Aj	⊥	⊤	⊥	also_see
synonym	S ↔ S	N, V, Aj, Av	N, V, Aj, Av	⊥	⊥	⊤	synonym
be_in_state	S ↔ S	N, Aj	Aj, N	⊥	⊥	⊥	be_in_state
categoryDomain	S → S	N, V, Aj, Av	N	⊥	⊥	⊥	⊥
usageDomain	S → S	N	N	⊥	⊥	⊥	⊥
regionDomain	S → S	N, Aj	N	⊥	⊥	⊥	⊥
derived	S ↔ S	Av	Aj	⊥	⊥	⊥	derived

Табела 6.2 Релације додељивања типа података у онтологији *SWNonto*

Property	Domain	XML Schema Datatype
hasLiteral	Word	xsd:string
hasDefinition	Synset	xsd:string
hasSynsetID	Synset	xsd:string
hasPos	Synset	xsd:string
hasBcs	Synset	xsd: string
hasSentimentPositive	Synset	xsd:decimal
hasSentimentNegative	Synset	xsd:decimal
hasSumo	Synset	xsd:string
hasDomain	Synset	xsd:string

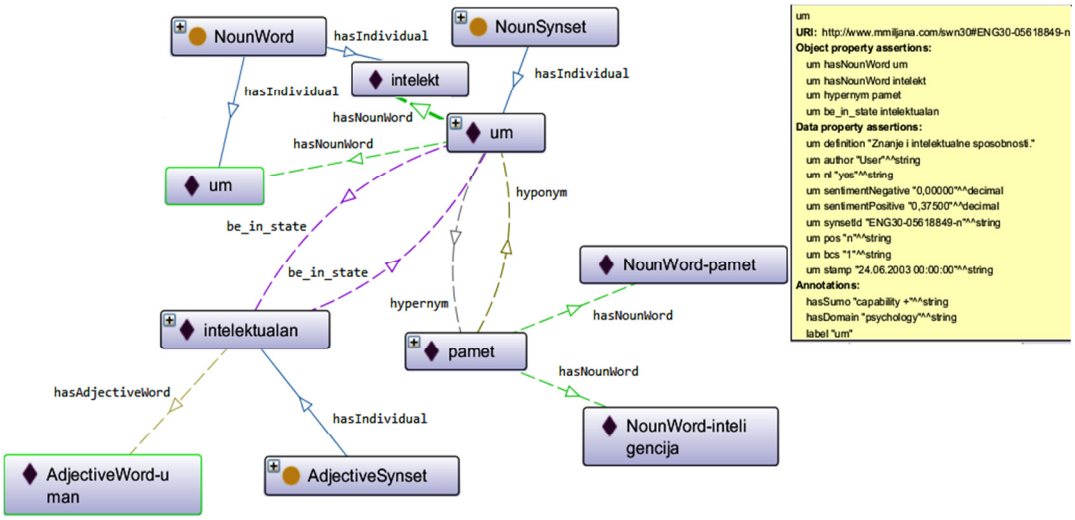
Инстанце класе *Word* описују се релацијама доделе типа података и аналогне су литералима у семантичкој мрежи ворднет. Пример декларације инстанце која је аналогна литералу *um* (интелект) дат је у наставку.

```
<owl:NamedIndividual rdf:about="&swn30;NounWord-um" rdfs:label="um">
   <rdf:type rdf:resource="&swn30;NounWord"/>
   <swn30:literal rdf:datatype="&xsd;string">um</swn30:literal>
</owl:NamedIndividual>
```

Инстанце класе *Synset* описују се бинарним релацијама и релацијама доделе типа података и аналогне су синсетовима у семантичкој мрежи ворднет. Пример декларације инстанце која је аналогна синсету *um* (интелект) дат је у наставку.

```
<owl:NamedIndividual rdf:about="&swn30;ENG30-05618849-n" rdfs:label="um">
  <rdf:type rdf:resource="&swn30;NounSynset"/>
  <swn30:synsetId rdf:datatype="&xsd;string">ENG30-05618849-n</swn30:synsetId>
  <swn30:hasNounWord rdf:resource="&swn30;NounWord-um"/>
  <swn30:hasNounWord rdf:resource="&swn30;NounWord-intelekt"/>
  <swn30:definition>Znanje i intelektualne sposobnosti.</swn30:definition>
  <swn30:pos rdf:datatype="&xsd;string">n</swn30:pos>
  <swn30:bcs rdf:datatype="&xsd;string">1</swn30:bcs>
  <swn30:nl rdf:datatype="&xsd;string">yes</swn30:nl>
  <swn30:stamp rdf:datatype="&xsd;string">24.06.2003 00:00:00</swn30:stamp>
  <swn30:author rdf:datatype="&xsd;string">User</swn30:author>
  <swn30:sentimentPositive rdf:datatype="&xsd;decimal">0,37500
  </swn30:sentimentPositive>
  <swn30:sentimentNegative rdf:datatype="&xsd;decimal">0,00000
  </swn30:sentimentNegative>
  <swn30:hasDomain rdf:datatype="&xsd;string">psychology</swn30:hasDomain>
  <swn30:hasSumo rdf:datatype="&xsd;string">capability +</swn30:hasSumo>
  <swn30:hypernym rdf:resource="&swn30;ENG30-05617606-n"/>
  <swn30:be_in_state rdf:resource="&swn30;ENG30-01332386-a"/>
</owl:NamedIndividual>
```

Визуелна репрезентација наведене декларације инстанце класе *Synset* којом се репрезентује синсет *um* (интелект) дата је на слици 6.17.



Слика 6.17 Инстанца класе *Synset* којом се репрезентује синсет *um* (интелект) и релације које гради са инстанцама других класа

6.5 Примена онтологије задатака реторичких фигура у препознавању фигуративног говора

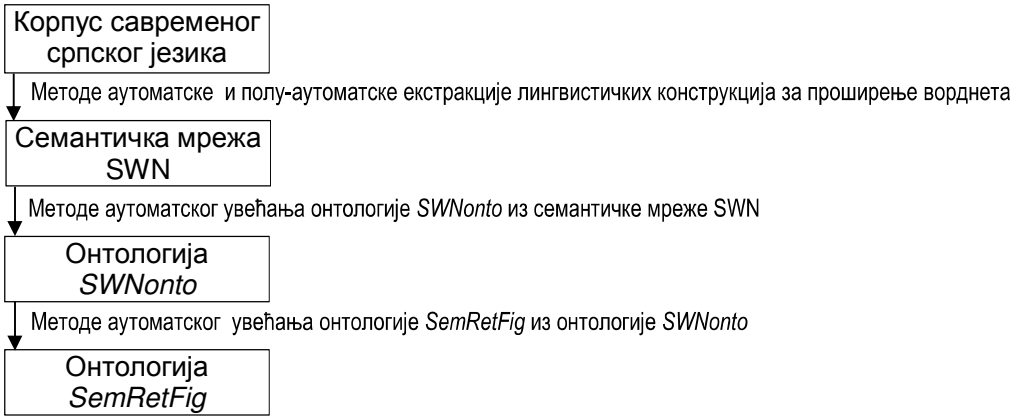
Аутоматско препознавање фигуративног изражавања је нова област интересовања на пољу рачунарске лингвистике. У 2015. је први пут, на глобалном плану, постављен један семантички задатак ове врсте - задатак семантичке анализе фигуративног говора на Твитеру (*International Workshop on Semantic Evaluation* - SemEval-2015).[277] Истраживања у овој области крећу се у два правца:

- ка побољшању метода класификације поларитета осећања (Reyes & Rosso, 2012b; Rentoumi et al., 2010);
- бољем разумевању лингвистичких структура на различитим нивоима (Hao & Veale, 2010; Veale & Hao, 2008).

У многима од њих важну улогу имају семантичке мреже (Marrafa et al., 2006; Mendes, 2006; Barbieri, Ronzano & Saggion, 2015), онтологије (Kontopoulos, Berberidis, Dergiades & Bassiliades, 2013) и лексички ресурси као што су

[277] http://alt.qcri.org/semeval2015/index.php?id=tasks

корпуси (Hao & Veale, 2010), специјализовани речници (Reyes & Rosso, 2012a; Barbieri, Ronzano & Saggion, 2015) и лексикони (Barbieri, Ronzano & Saggion, 2015). У претходним одељцима описали смо структуре онтологија *SWNonto* и *SemRetFig* као и правила, дефинисаних у онтологији *SWNonto*, којима се аутоматски генеришу инстанце класа онтологије *SemRetFig*. Међутим, главни проблем у инжењерингу једне онтологије према (Devedžić, 2006, pp. 43-44) је њено оспособљавање да се мења истовремено са увећањем и изменом доменских знања. Способност онтологије да непрестано „учи“ (енг. ontology learning), тако што се аутоматским и полу-аутоматским поступцима екстрахују и анотирају нова знања из посматраног домена и интегришу са постојећим знањима у онтологији, у многоме зависи од алата који се у ту сврху примењују. Када је реч о онтологији *SemRetFig,* како је њен задатак да препознаје различите реторичке фигуре, врло је важно да она „учи“ различите облике фигуративног говора из примера на природном језику. С обзиром да се процес тог учења заснива на препознавању учесника у свим семантичким релацијама онтологије *SWNonto*, а посебно у релацијама као што су: *specificOf*, *specifiedBy*, *near_antonym*, *derived-pos, be_in_state* и *synonym*, уколико обезбедимо аутоматско или интензивно ручно и континуирано проширење онтологије *SWNonto* овим релацијама и ентитетима који се повезују овим релацијама, утолико ћемо имати онтологију која је потпунија и ефикаснија у постављеним задацима (в. одељак 6.3). У циљу аутоматизације проширења *SWNonto* онтологије наведеним релацијама, у наставку предлажемо методу аутоматског проширења семантичке мреже SWN (самим тим и онтологије *SWNonto*) релацијама *specificOf* и *specifiedBy* као и веб алат којим ће интернет корисницима мреже Српски ворднет бити омогућено да успостављају и остале, одговарајуће, семантичке релације између синсетова. Дијаграм учења ових онтологија приказан је на слици 6.18.



Слика 6.18 Учење онтологија *SWNonto* и *SemRetFig* из корпуса савременог српског језика

6.5.1 Једна метода полу-аутоматског проширења семантичке мреже SWN

Метода коју предлажемо ослања се, са једне стране, на истраживања (Marrafa et al., 2006) и (Mendes, 2006) која се баве односима именица и придева, али за разлику од њих, ми трагамо за специфичним везама између именица и придева, односно разматрамо оне описне придеве који су специфични за мали или врло мали скуп или за само једну именицу. Извор генерисања семантичке релације коју предлажемо лежи у реторичкој фигури компарације, таквој да има значајну фреквентност у текстовима на природном језику. Друга врста истраживања, на коју се ослања наш рад, приказана је у радовима (Veale & Hao, 2008; Hao & Veale, 2010), а тиче се развоја аутоматских метода екстракције семантичког знања из примера примене фигуре компарације. У том смислу, предлажемо (Mladenović, Mitrović & Krstev, 2016) екстракцију лингвистичких конструкција облика „PRIDEV kao IMENICA“ из анотираног дела Корпуса савременог српског језика (в. одељак 3.1). Из анотираног корпуса можемо екстраховати лингвистичке конструкције које се примењују у различитим доменима употребе језика. Предност коришћења анотираног корпуса не састоји се само у скраћивању поступка екстракције у односу на поступак који користи Гугл машину (што је примењено у радовима Вејла и Хао), већ и у генерисању ужег скупа

кандидата (Mladenović, Mitrović & Krstev, 2016) у којима се на крају дате конструкције компарације налазе искључиво именице. Друга значајна предност употребе анотираног корпуса над Гугл машином је могућност поређења добијених резултата по фреквенцији појављивања, што нам допушта да одаберемо подскуп најфреквентнијих, односно најчешће коришћених у природном говору. Метода проширења релацијом *specificOf/specifiedBy* коју предлажемо, биће приказана на примеру проширења SWN, али је применљива на ворднет структуру у општем случају и састоји се из следећих корака:

1) Из анотираног корпуса датог природног језика K_l екстраховати лингвистичке конструкције *sims* облика „PRIDEV kao IMENICA“ и формирати скуп $Sims$

$$Sims = \{\text{„PRIDEV kao IMENICA“}\}, \quad \text{sims} \in \text{Sims} \subset K_l$$

У нашем примеру, из анотираног Корпуса генерисано је 5952 конкорданци „PRIDEV kao IMENICA“ као у следећим примерима:

650810: ri više . - Kakva je ? - <Bela kao mleko> . Ona traži isto kao i

18206219: dan od zatvorenika ; lica <žuta kao limun> , radosno polete prema

26045112: i zaturenoj na potiljak , <crven kao cvekla> , Platon Rjapčikov i jo

2) Из скупа $Sims$ елиминисати све елементе код којих придеви нису дескриптивни

$$SimsRedycedByADJ = \{sims \in Sims \mid \text{PRIDEV 'is descriptive'}\}$$

као у следећим примерима где су у питању присвојни придеви:

251511: za taj dan . Jer reč je <ljudska kao glad> . Nema uvek istu snagu .

137584415: Drugog? Ljubav <majčinska kao vernost> ljubav muško-

У нашем случају добијено је $|SimsRedycedByADJ| = 2030$ елемената.

3) Из скупа SimsRedycedByADJ елиминисати све елементе код којих су именице личне или су замењене скраћеницом (3. пример).

$$SimsRedycedByNOUN = \{\text{sims} \in SimsRedycedByADJ \mid \text{IMENICA 'is a common N'}\}$$

као у следећим примерима:

132719070: Pljevlja bi bila bogata i <bleštava kao Las> Vegas - kaže jedan od

40699798: da bude slavna i <bogata kao Monika> Seleš . Kako se koja

68456010: zatvoru u Beogradu , <opštepoznatom kao CZ> , naći u poziciji onih

У нашем случају је добијено $|SimsRedycedByNOUN| = 1059$ елемената.

4) Из скупа $SimsRedycedByNOUN$ генерисати подскуп најфреквентнијих елемената

$$SimsMostFrequented = \{sims \in SimsRedycedByNOUN \mid \text{freq}(sims) \geq k\},$$

где је k минимална фреквенција појављивања структуре „PRIDEV kao IMENICA“ у посматраном корпусу K_l. У нашем случају, за вредност $k = 1$, укупан број парова PRIDEV-IMENICA, кандидата за проширење ворднета је $|SimsMostFrequented| = 1059$.

5) Од датог скупа $SimsMostFrequented$ креирати текстуалну датотеку *Adjective_As_Noun* парова PRIDEV-IMENICA над којом ће бити примењен алгоритам (алгоритам 6.1 дат у псеудокоду) аутоматског проширења ворднета.

Алгоритам 6.1 Аутоматско проширење ворднета семантичком релацијом између синсетова именица и придева са једним литералом чије значење је означено као прво.

```
Input: Adjective_As_Noun text file
Output: a pair of WordNet mutually inverse
semantic relations (specificOf/specifiedBy)
for each input adjective-noun pair
-----------------------------------------------
foreach adjective-noun pair in adjective-noun pairs
if ((adjective exists in Wordnet.adjective.literals)
    and (noun exists in Wordnet.noun.literals)) {
      if ((Wordnet.senses(adjective).Count==1)
         and (Wordnet.senses(noun).Count==1)
         and (Wordnet.sense(adjective).FirstSense)
         and (Wordnet.sense(noun).FirstSense) ) {
          Create_Relation(specificOf,adjective,noun);
          Create_Relation(specifiedBy,noun,adjective);
      }
      else
       foreach (sense in Wordnet.senses(adjective)) {
         add_to_adjective_senses(adjective,sense,synsetId)}
       foreach (sense in Wordnet.senses(noun)) {
         add_to_noun_senses(noun,sense,synsetId)}  }  }
```

Представљеним алгоритмом секвенцијално обрађујемо улазне кандидате PRIDEV-IMENICA и за сваки пар проверавамо да ли у датом ворднету постоје синсетови придева и именице који су лексикализовани литералима тренутно обрађиваног придева PRIDEV и именице IMENICA. Затим се примењује процедура аутоматског креирања пара семантичких релација *specificOf/specifiedBy* између синсетова придева и именице уколико су лексикализовани само по једним литералом чије је значење прво. Прво значење једне речи је оно значење дате речи у природном језику које је, корпусом или релевантним речником, дефинисано као најчешће коришћено у датом језику. Интуиција на којој се заснива ограничење које уводимо односи се на минимизацију грешке упаривања која постоји у случају када не постоје синоними у посматраном синсету и значење литерала је прво. У том случају, вероватноћа грешке постоји само уколико бар један од синсетова није коректно допуњен литералима и нема исправно додељено значење или жељено значење није прво и оно не постоји као литерал у SWN. У том случају, пошто нам је извор могућих грешака познат и ограничен, могуће је извршити проверу структуре SWN (генерисати тестове) пре саме примене алгоритма. С друге стране, уколико бар један од синсетова има више од једног синонима или има један чије значење није прво, такав се пар PRIDEV-IMENICA одваја у две независне датотеке: датотеку придева (*adjective_senses*) и свих његових значења пронађених у SWN и њој аналогну датотеку именица (*noun_senses*). Ове датотеке се даље користе у веб алату[278] (прилог 6.5) апликације описане у радовима (Mladenović, Mitrović & Krstev 2014) и (Mladenović & Mitrović, 2014) којим корисници ручно упарују придеве и именице, бирајући их према одговарајућем значењу и повезујући их паром релација *specificOf/specifiedBy*. Парови PRIDEV-IMENICA за које се одмах на почетку процеса испитивања утврди да не постоје у облику литерала у датом ворднету, учествују у генерисању кандидата за допуну ворднета на регуларан начин — уносом синсетова.

[278] Веб алат је креиран тако да се могу упаривати и литерали других врста речи помоћу неке од релација које постоје у SWN

Пре примене датог алгоритма, извршили смо испитивање SWN да утврдимо каква је његова структура у смислу описаних ограничења. Тако, SWN, који садржи преко 22 хиљаде синсетова, садржи 1660 синсетова придева са једним литералом, од чега је код 1452 значење тог литерала означено као прво, док је именичних синсетова са једним литералом чије је значење означено као прво 15.035. Применом предложеног алгоритма 6.1, од укупно 1059 парова PRIDEV-IMENICA, са особином „укупно парова чија оба члана имају по једно значење и то значење је прво“ нађено је 69 парова. Парова PRIDEV-IMENICA којих има у SWN, али са више значења или са једним које није прво има укупно 302. Преосталих 688 парова односе се на оне случајеве када бар један члан пара PRIDEV-IMENICA не постоји као литерал у SWN. Дакле, предложеном методом SWN се може одмах, без претходне допуне, проширити са 372 пара релације типа *specificOf/specifiedBy*.

6.5.2 Оцена методе аутоматског проширења онтологије *SWNonto*

Да бисмо оценили да ли је фреквенција појављивања прихватљив параметар генерисања релација *specificOf/specifiedBy* паровима PRIDEV-IMENICA у односу на заступљеност у природном језику, користили смо онлајн анкету и спровели је помоћу *Google Forms*[279]. Поређењем листе (означићемо је са *Листа1*) генерисане аутоматски - помоћу корпуса, филтриране помоћу корака 1-4 описаних у одељку 6.5.1 и уређене у опадајућем низу по фреквентности парова, са листом добијеном анонимним анкетирањем (означићемо је са *Листа2*), желели смо да оценимо која вредност прага фреквенције k обухвата резултате добијене анкетом. Димензија листе *Листа1* одређена је доњим прагом фреквенције појављивања $k = 1$ и износила је 1059 елемената. Чланове листе *Листа2* из листе *Листа1* одабрао је лингвистички експерт ручним путем тако да то буду они парови који су често заступљени у обичном говору без проучавања ширег контекста. С обзиром да је *Листа1* генерисана аутоматски из корпуса, појава шума у

[279] https://www.google.com/forms/about/

подацима била је неизбежна. На пример, поред структура: *čist kao suza*, *lak kao pero*, *hladan kao led*, *veran kao pas*, појављују се и структуре: *dobar kao pisac*, *poznat kao vođa*, *dobar kao oblik* и др. Ручно је одабрано 154 структура и помоћу њих је направљено 4 формулара *Google Forms*.

Анкета је рађена у временском периоду од 5 дана, тако што су формулари објављивани сукцесивно на друштвеној мрежи Фејсбук. Анонимни корисници су на свако питање постављено у облику „Да ли се у обичном говору може рећи за некога/нешто да је PRIDEV као IMENICA?“ обавезно одговарали са или *ДА* или *НЕ*. У табели 6.3 приказана је расподела питања према формуларима као и број испитаника који су дали одговоре.

Табела 6.3 Расподела питања и испитаника по једној анкети (*Google Form*)

Анкета	Број питања по анкети	Број учесника анкете
1	30	46
2	42	138
3	41	150
4	41	100
Укупно	154	434

Најпре смо измерили допринос испитаника како бисмо утврдили скуп оних чије одговоре можемо сматрати релевантним. Релевантност је мерена аритметичком средином одговора, па су дати скуп чинили они испитаници код којих није било значајне разлике између аритметичких средина њихових одговора. Да би се измерио појединачни допринос сваког учесника, четири Гугл формулара подељено је на 7 подскупова, како је приказано у табели 6.4 (сваки Гугл формулар, осим првог, подељен је на два дела). Сваки подскуп имао је 30 или мање питања и конвертован је у матрицу у којој је сваки ред представљао одговоре једног испитаника, а свака колона представљала једно питање у облику *<придев> као <именица>.* Садржај у свакој ћелији матрице имао је вредност 1 ако је испитаник одговорио на питање са *ДА* и вредност 0 уколико је тај одговор био *НЕ*. Редови матрице су међусобно упоређени *t*-тестом за зависне узорке како би се утврдило да ли има битне разлике између аритметичких средина одговора учесника. Из сваког подскупа

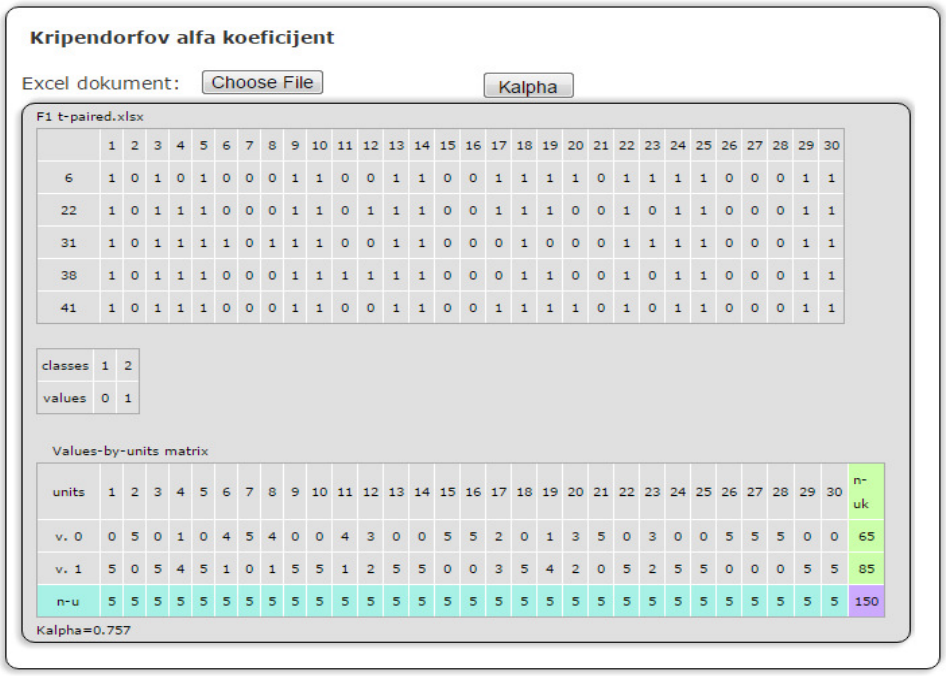
изабрали смо по пет испитаника чија су међусобне разлике аритметичких средина свих одговора, на основу резултата *t*-теста[280] за зависне узорке биле најмање са интервалом поверења 95%.

Након тога је извршено додатно испитивање сагласности тако одабраних испитаника Крипендорфовим α–тестом (Kalpha) као рестриктивнијом методом за утврђивање степена сагласности оцена испитаника. Крипендорфов α коефицијент (Hayes & Krippendorff, 2007) оцењује степен сагласности у оценама посматрача у општем случају када постоје: више од 2 посматрача, већи број питања, већи број различитих одговора, различите врсте метрике (номинална, редна, интервална, пропорционална), питања на која посматрачи нису дали одговоре.[281] Вредност *Kalpha* коефицијента може бити у интервалу [0,1] где *Kalpha=1* представља степен потпуне сагласности, а *Kalpha=0* степен потпуне несагласности. *Kalpha* може узети и негативну вредност из интервала [-1,0) услед грешака узорковања или системског неслагања. Истраживања (Lombard, Snyder-Duchand & Campanella Bracken, 2002; Hayes & Krippendorff, 2007; Maggetti, 2013) о вредностима *Kalpha* коефицијента које означавају нивое поузданости испитаника показују да се сагласност изражена *Kalpha* коефицијентом $\alpha \geq 0{,}667$ може сматрати поузданом, а она код које је $\alpha \geq 0{,}8$ може се сматрати веома поузданом. Резултати које смо добили употребом *Kalpha* теста над скупом од 5 испитаника у свакој од 7 матрица приказани су у табели 6.4. На слици 6.19 приказан је начин обраде прве матрице[282], где је у горњем делу приказана матрица бр. 1 као улазни скуп података анкете бр. 1 са 30 питања и одговорима (0 или 1) 5 испитаника. У доњем делу су приказани резултати *Kalpha* теста у облику матрице *values-by-units* и вредност *Kalpha* коефицијента $\alpha = 0{,}757$.

[280] За ово израчунавање написан је макро у MS Excel 2010 и коришћен је Excel статистички алат *t-Test: Paired Two Sample for Means*

[281] Крипендорфов α–тест реализовали смо у облику алата у веб апликацији SWNE (прилог 6.6) тако да може бити коришћен у разним анкетама и са различитим врстама питања и одговора. Улазни документ се прослеђује у формату xlsx, а излаз се исписује на веб страни алата.

[282] На исти начин обрађено је свих 7 матрица.



Слика 6.19 Резултати *Kalpha* теста над првом анкетом (степен сагласности 5 испитаника означених бројевима 6,22,31,38 и 41)

С обзиром да су прва два формулара и део трећег (1, 2a, 2b,3a) *Kalpha* коефицијентом оцењени као поуздани, узета су у даље разматрање она питања на која је одговор најмање три испитаника био *ДА*, Укупан број тих питања је 53, а њихова дистрибуција у односу на формуларе дата је у последњој колони табеле 6.4.

Табела 6.4. Степен сагласности испитаника у анкетама спроведеним формуларима *Google Forms*, број одговора који припадају формуларима који су *Kalpha* тестом оцењени поузданим, као и број одговора на која су испитаници већински одговорили са ДА

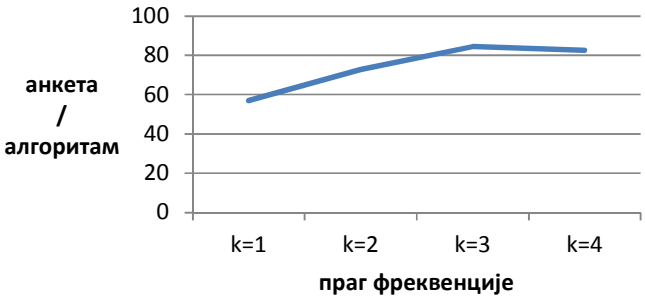
Подскуп питања	Број испитаника	Број питања	Kalpha вредност	Број питања већински означених са *ДА*
1	5	30	$\alpha = 0.757$*	16
2a	5	21	$\alpha = 0.713$*	17
2b	5	21	$\alpha = 0.698$*	15
3a	5	21	$\alpha = 0.688$*	5
3b	5	20	$\alpha = 0.484$	
4a	5	21	$\alpha = 0.434$	
4b	5	19	$\alpha = 0.375$	
Укупно		154		53

На крају, желели смо да утврдимо утицај промене фреквентног прага на релевантност аутоматски одабраних парова PRIDEV-IMENICA, мерено резултатима добијених анкетом. Листа *Листа1* умањена је тако да садржи само формуларе са релевантним одговорима (1, 2а, 2б и 3а), што је чинило 93 пара. Од њих је креирана листа *Листа1'*. Осим тога, креирана је и листа *Листа2'* која је садржавала само оне PRIDEV-IMENICA парове листе *Листа1'* које су испитаници оценили позитивно. Најпре смо поставили фреквентни праг на $k = 4$, што значи да смо алгоритмом обухватили само оне парове чија је фреквенција појављивања у корпусу $k \geq 4$. Таквих парова у листи *Листа1'* је било 23. Од тих 23 у листи *Листа2'* налазило се 19 парова, што значи да испитаници нису позитивно оценили 4 пара која је препознао алгоритам. Целокупна статистика која показује који проценат парова добијених алгоритмом је обухваћен и људском проценом дат је у табели 6.5.

Табела 6.5 Однос броја ручно и аутоматски одабраних парова у зависности од прага фреквенције

праг фреквенције	алгоритам	анкета	анкета / алгоритам
$\boldsymbol{k = 1}$	93	53	57%
$\boldsymbol{k = 2}$	44	32	73%
$\boldsymbol{k = 3}$	32	27	84%
$\boldsymbol{k = 4}$	23	19	83%

Промена односа анкетног и аутоматског одабира, када се мења праг фреквенције, дата је на слици 6.20.



Слика 6.20 Однос броја парова одабраних анкетом у односу на број одабраних алгоритмом, у зависности од промене прага фреквенције

Слика показује начин на који се, на узорку од 93 пара који су садржани у листи *Листа1'* (*Kalpha* сагласни), мења проценат учешћа ручно одабраних парова у подскупу парова добијених избором само оних парова из исте листе чија је фреквенција појављивања у корпусу једнака или већа од датог прага, када се тај праг мења. Остварени резултат од 84% даје нам ручно измерену тачност алгоритма за аутоматско проширење ворднета када је праг фреквенције $k = 3$. На основу добијених резултата можемо закључити да ако у методи проширења SWN описаној у претходном одељку за праг фреквенције узмемо $k = 3$, можемо очекивати тачност одабира парова у износу од 84%.

6.5.3 Ка препознавању фигуративног говора у текстовима на српском језику

Аутоматским уносом 96 парова и ручним повезивањем 302 пара, SWN је, без претходне допуне синсетовима, проширен са 372 пара релације типа *specificOf/specifiedBy*. Од преосталих 688 потенцијалних парова, укупно 564 парова генерисано је уносом и повезивањем нових синсетова у SWN. Број парова релације *specificOf/specifiedBy* са којим смо ушли у даљи процес обраде фигуративног говора био је 936. Овако дефинисаном релацијом у SWN, добили смо исти број инстанци класе *Поређење* у онтологији *SemRetFig*. Поред тога, када концепти придева за које је дефинисана релација *specificOf* имају и свој антонимски пар као у примерима (безобзиран, пажљив), (танак дебео), (мек, тврд), тада постоје и парови и кандидати за инстанцирање компаративне ироније. У примерима фигуре *поређење*: „безобзиран као грубијан“, „танак као сламка“, „мек као свила“ одговарајући кандидати за инстанцирање фигуре *ирониja* су: „пажљив као грубијан“, „танак као буре“, „мек као камен“ итд. Међутим, укупан број придева у SWN повезаних релацијом *near_antomym* износи 792, па је и број парова који могу учествовати у генерисању инстанци класе *Ironija* помоћу правила (84) знатно мањи и износи 274.

Да бисмо оценили степен препознавања реторичких фигура *поређење* и *ирониja*, у текстовима на српском језику, спровели смо тест који се састоји из четири корака:

1. креиран је алат којим се, уз помоћ регуларних израза, улазни текст парсира тако да се екстрактују изрази у форми индивидуа класа *Ironija* и *Poređenje* (прилог 6.7); из прилога се може уочити да су, осим структура „реч *као* реч“, одговарајућим регуларним изразом екстрактоване и структуре облика „реч *попут* реч“;
2. дефинисан је скуп различитих улазних текстова који се састоји од 10 дигитализованих дела (наведених у легенди на слици 6.17): збирки дечијих песама (Антић, Ршумовић, Данојлић и Станчевић), дечијег романа (Орлови рано лете - Ћопић), збирке бајки[283], комедије (Нушић), романа епске фантастике (Мартин), романа (Орвел) и дела студије феноменолошких појава „Метафоре и алегорије“ (Михаило Петровић Алас); овом скупу придружени су и текстови који су у одељку 5.4.4 представљени као скупови за тестирање састављени од оцена филмова и вести са новинских портала (ступци означени лабелама *vesti* и *filmovi* на слици 6.17);
3. из екстрахованих структура, ручно су уклоњене све оне које нису облика „ПРИДЕВ *као* ИМЕНИЦА“ или „ПРИДЕВ *попут* ИМЕНИЦА“;
4. креиран је алат који, на основу идентификације индивидуе, у онтологији *SemRetFig* закључује којој класи фигура припада и на основу тога јој додељује етикету фигуре.

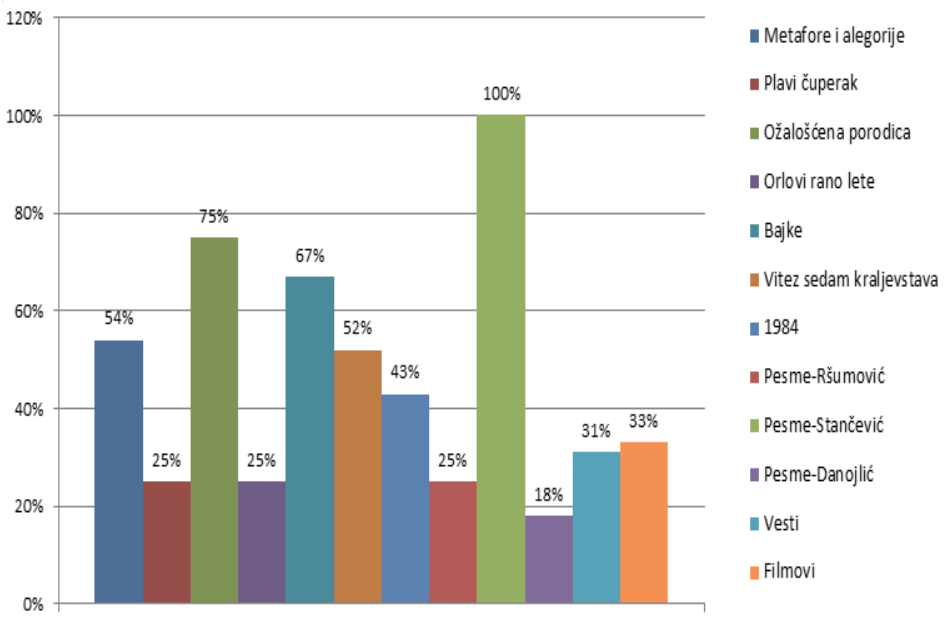
Резултати препознавања фигуре *поређење* приказани су на слици 6.21, а детаљни приказ свих пронађених структура у једном тексту дат је на примеру романа „Витез седам краљевстава“ Џорџа Мартина у прилогу 6.8. Лингвистичке структуре које су добијене овим тестом указују на то да се у поетским текстовима могу наћи поређења неуобичајена у свакодневном говору. На пример, у песмама Данојлића, постоје структуре „добар као слон“,

[283] http://www.zadecu.com/Bajke/

„спокојан као крава“, „устрептала као стрела“ и сл. У песмама Ршумовића могу се наћи: „плаво као ћилим“, „зелено као тепих“, „ружан као четка“, итд.

Препознавање фигуре *ирониja* овом методом није дало задовољавајуће резултате. Иако има ироничних тврђења попут: „*Baš si krasan prijatelj*“, „*Ma baš je pametna*“, „*Sve si uspeo uništiti, prekrasno*“, „*E baš si me obradovao*“, „*Samo ti nastavi da se svađaš*“ у студији феноменолошких појава „Метафоре и алегорије“ аутора Михаила Петровића Аласа (Petrović, 1967), аутоматском методом која садржи правило (84) у датим текстовима нису пронађене структуре компаративне иронијe.

На крају овог теста, анализирали смо структуру скупова за тестирање *filmovi* и *vesti* којима је тестиран SAFOS – систем за анализу осећања текстова на српском језику, како бисмо оценили могућност за његову даљу доградњу у смислу препознавања фигура. Резултати су показали ниску али уједначену тачност препознавања од око једне трећине укупног броја фигура *поређења*. На примерима структура које су екстраховане из ових скупова као што су: „*okaraterisan kao triler*“, „*poznat kao zemlja*“, „*urađen kao nastavak*“ и сл. може се видети да у текстовима постоје и лажно позитивни кандидати за инстанцирање фигуре *поређења*. Међутим, како је препознавање фигура вршено помоћу скупа већ познатих и класификованих фигура онтологије *SemRetFig*, постигнута је прецизност у износу од 100%, чиме су лажно позитивни кандидати одбачени. С друге стране, структуре облика: „*oštrim poput brijača*“, „*hrabrim као SUPERMEN*“, „*puknuti kao vidik*“, „*razmnožavati se kao vinska mušica*“ указују на низак ниво одзива класификације и потребу испитивања и другачијих лингвистичких структура од описаних, узимањем у обзир и других врста сем заједничких именица (личних имена, фраза, скраћеница и др.) са једне као и глагола са друге стране лингвистичке структуре.



Слика 6.21. Тачност препознавања фигуре *поређења* у текстовима на српском језику онтологијом *SemRetFig*

6.5.4 Ка даљем препознавању фигуративног говора

У претходном одељку је истакнуто да примењеним тестом није пронађена компаративна ирониja. Проблеми могу бити двојаки: мали скуп инстанци дате класе и неодговарајући скуп за тестирање. Први од проблема се може решавати унапређењем семантичке мреже Српски ворднет. Унапређење у том погледу подразумева унос нових концепата који су у PWN дефинисани у лексикографским датотекама: *noun.cognition*, *noun.feeling*, *noun.phenomenon*, *verb.cognition*, *verb.creation*, *verb.emotion*, *verb.perception* и *verb.stative* и успостављање и унапређење свих врста лексичко-семантичких релација. Што се тиче другог проблема, он је анализиран у радовима који проучавају фигуративни говор. На пример, у раду (Wallace, 2015) указује се на проблем препознавања ироничних тврђења као посебног задатка класификације текста за који, за разлику од општег класификационог поступка, није довољно посматрати *n*-граме и њихове фреквенције појављивања. Говорећи о новој области класификације текста на ироничне и неироничне – аутоматском препознавању ироније (енг. computational irony) –

аутор истиче неопходност генерисања корпуса лингвистичких јединица различитог нивоа (текстова, реченица, фраза) које носе иронично значење ради моделовања предиктора ироничног понашања у тексту и њихове примене у процесима бинарне класификације текста на ироничне и неироничне лингвистичке јединице.

Анализом постигнутих резултата у претходном одељку као и достигнућа у радовима (Reyes & Rosso, 2012b; Filatova, 2012; Gianti et al., 2012) утврдили смо да је за потпунију методологију препознавања фигуративног говора у текстовима на српском језику неопходно:

- употребити лингвистичке алате - лематизер и алат за препознавање и обележавање врстама речи, чиме би корак 3. у претходном одељку био спроведен аутоматски;
- генерисати, означити и оценити скупове за обуку појединачних фигура, на шта указују бројна истраживања која се тичу фигура: *ирониja*, *сарказам*, *оксиморон*, *метафора,* како је то представљено у наведеним радовима;
- развијати семантичке ресурсе као што су речници синонима, антонимски-парови, ланци градабилних придева и сл. како би се уз њихову помоћ лакше и ефикасније повећавао број лексичко-семантичких релација у Српском ворднету;
- развијати друге нестандардне врсте речника као што су: емотикони и жаргонски речници који се примењују у интернет комуникацији, тзв. AOL[284] и комуникацији кратким порукама;
- развијати различите методе препознавања синтаксних и семантичких правила под којима се граде појединачне фигуре – на пример, за потребе ирониje неопходно је утврдити везе са постојањем нестандардних знакова интерпункције као и са уметањем симболичких ознака и променом фреквенције појављивања одређених речи и израза у непосредној близини фигуре.

[284] http://www.internetslang.com/AOL-meaning-definition.asp

7. Закључак

У овој тези смо пројектовали, креирали, имплементирали и оценили SAFOS - први систем за анализу осећања текстова на нивоу докумената, на српском језику, методом максималне ентропије која припада групи метода надгледаног машинског учења. Предложили смо нову методу редукције скупа предиктора хибридизацијом основних предиктора и нових који су добијени поступцима груписања и пресликавања. Показали смо да ова метода статистички значајно унапређује класификацију на основу осећања у текстовима на српском језику. Такође, показали смо да у задацима обраде природног језика који се тичу анализе осећања, важну улогу играју лингвистички ресурси: речници, лексикони, семантичке мреже и онтологије. Резултати истраживања у овој тези показују да се метода надгледаног машинског учења максималне ентропије, која се у овој тези користи у класификацији текста на основу осећања, може унапредити укључивањем језичких ресурса у процес оптимизације скупа предиктора којим се репрезентује класификациони модел. Класификациони модели су генерисани и испитивани уз помоћ софтверског алата SAFOS пројектованог тако да пружа могућност комбиновања различитих врста n-грамских предиктора и различите начине њихове редукције. Оцењујући резултате добијене класификацијом текстова у SAFOS-у закључили смо да се основни скупови предиктора успешно редукују применом језичких ресурса.

У првом делу тезе бавили смо се анализом дословног значења текста, а у другом изградњом ресурса и формалних поступака препознавања употребе фигуративног говора са циљем даљег унапређења процеса класификације текста на основу осећања. У том смислу, генерисали смо прву доменску онтологију реторичких фигура на српском језику *RetFiguresOnto* и онтологију задатака *SemRetFig* како бисмо креирали систем за препознавање фигуративног говора. У ту сврху је семантичка мрежа Српски ворднет проширена паром нових семантичких релација *specificOf/specifiedBy* и предложена је метода за полуаутоматско проширење SWN мреже овом релацијом. Предложена је метода за интеграцију онтологија *RetFiguresOnto* и

SWNonto чиме је аутоматски генерисана *SemRetFig* онтологија која се у даљем раду може ширити аутоматски, увођењем нових правила за генерисање инстанци реторичких фигура. На крају рада пројектован је, имплементиран и оцењен систем за препознавање фигуративног говора текстова на српском језику.

У даљем раду бавићемо се унапређењем SAFOS система, имплементацијом нових врста предиктора и проширењем скупова за учење. Међутим, треба истаћи да проширење скупова подразумева и повећање способности система да „разуме“ различите дискурсе (хумор, дијалог, фигуративни говор и др.) као и различите системе знакова (емотиконе, скраћени говор, Твитер хештагове и линкове и др.).

У одељку 6.3 истакли смо да су циљеви примене онтологије *SemRetFig*, поред препознавања фигуративног говора у тексту, аутоматска анотација реторичких фигура и побољшање методе класификације осећања. У даљем раду бавићемо се и овим проблемима. Аутоматска анотација реторичких фигура подразумева: изградњу радног окружења као у примеру JANTOR-а у раду (Gawryjolek, Di Marco & Harris, 2009) са могућношћу одабира што већег скупа фигура за означавање, флексибилност у раду са различитим врстама, форматима и величинама текстова и способност генерисања излазних резултата у различитим форматима како би могли бити примењени у даљим истраживањима. Када је реч о унапређењу класификатора анализе осећања, треба истаћи да речју или фразом изражена јачина осећања може бити умањена, увећана или у потпуности промењена у датом контексту употребом стилских фигура тропи. Анализирајући употребу фигуративног говора у облику ироничних компарација, Хао и Вејл (Hao & Veale, 2010) су уочили да оне у тексту делују попут модификатора значења (valence shifters) „not“ , „never“ и „avoid“ (Kennedy & Inkpen, 2006) јер доводе до промене поларитета осећања речи или фраза. У општем случају, модификатори смањују, повећавају или мењају поларитет осећања речи или фраза. На сличан начин делују и тропи. По дефиницији, ирониja и сарказам мењају поларитет, дисфемизам и хипербола појачавају постојећи ниво сентименталне изражајности, док га литота и еуфемизам умањују. Метафора, метонимија,

оксиморон и компарација имају сложеније механизме деловања у оба правца промене и јачине и поларитета осећања. У даљем раду посебно ћемо се бавити методама аутоматског препознавања ових фигура, како бисмо унапредили систем класификације текстова на основу осећања.

Задаци обраде природног језика у којима се решавају проблеми семантичког значења некада се не могу ослонити само на статистичке алате. Када атрибути неког процеса нису нумерички и при томе њихово значење није једнозначно и зависи од контекста, семантичке мреже, лексикони, доменске онтологије и други лексички ресурси представљају важна средства унапређења статистички заснованих система у рачунарској лингвистици.

Методе анализе осећања и истраживања мишљења добијају све већи значај на подручју маркетинга и продаје производа. Према Гримсу,[285] једном од водећих аналитичара у области пословне интелигенције, унапређивање метода анализе осећања и истраживања мишљења повећава утицај циљаног маркетинга, помаже брже откривање повољних прилика и претњи у области пословања, заштиту угледа пословних брендова, а крајњи циљ је повећање профита.[286] Када је реч о подручју науке, недавно је објављено[287] да IBM *Watson*, технолошка платформа са особинама вештачке интелигенције која користи методе обраде природног језика и машинског учења у задацима откривања скривених знања из велике количине неструктуираних података, препознаје осећања изражена у тексту са тачношћу од 80%, а поседује и способност препознавања и разумевања сарказма, хиперболе и скраћеног изражавања. IBM *Watson*[288] се примењује у области образовања, медицинске дијагностике, здравствене заштите као и у предикцији и праћењу сложених мултидисциплинарних процеса. Најзад, не могу се изоставити социолошки и друштвени аспекти (Cogburn, Hanson & Wozniak, 2012) примене алата анализе осећања у погледу побољшања безбедносних механизама и унапређења процедура заштите природе, друштва и добара.

[285] Seth Grimes
[286] https://www.socialmediaexplorer.com/social-media-monitoring/sentiment-analysis/
[287] http://www.9lenses.com/the-future-of-sentiment-analysis
[288] http://www.ibm.com/smarterplanet/us/en/ibmwatson/

ЛИТЕРАТУРА

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classication in web forums. *ACM Transactions on Information Systems, 26*(3), 12:1–12:34.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Languages in Social Media*, 30–38.

Ahlgren, O., Malo, P., Sinha, A., Korhonen, P. J., & Wallenius, J. (2012). A Dimensionality Reduction Approach for Semantic Document Classification. In *Proceedings of the CEUR Workshop*, 114–121.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, 347–354.

Alm-Arvius, C. (2003). *Figures of Speech*. Studentlitteratur, Sweden. Retrieved from: https://www.studentlitteratur.se/product_images/imagecache/pcatimg/_pm_w-800_h-800/9789144024912.jpg (September 20, 2015.)

Almuallim, H., & Dietterich, T. G. (1991). Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, 547–552. AAAI Press.

Alpaydin, E. (2010). *Introduction to Machine Learning, Second Edition*. The MIT Press, Cambridge, Massachusetts, London, England.

Alsharif, O., Alshamaa, D., & Ghneim, N. (2013). Emotion Classification in Arabic Poetry using Machine Learning. *International Journal of Computer Applications, 65*(16), 10–15.

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014). Learning sentiment from students' feedback for real-time interventions in classrooms. Adaptive and Intelligent Systems (pp. 40-49). *Lecture Notes in Computer Science 8779*. Springer.

Aman, S., & Szpakowicz, S. (2007). Identifying Expression of Emotion in Text. In *Text, Speech and Dialogue*, *4629*, 196–205. Springer.

Asghar, M. Z., Khan, A., Shakeel, A., & Kundi, F. M. (2014). A Review of Feature Extraction in Sentiment Analysis. *Journal of Basic and Applied Scientific Research 4*(3), 181–186.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. ELRA.

Bagić, K. (2012). *Rječnik stilskih figura*. Školska knjiga, Zagreb.

Литература је дата АПА стилом.

Balahur, A., & Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing and Management, 51*, 547–556.

Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language 28*, 56–75.

Barbieri, F., & Saggion, H. (2014). Modelling Irony in Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, {EACL}*, 56–64.

Barbieri, F., Ronzano, F., & Saggion, H. (2015). UPF-taln: SemEval 2015 Tasks 10 and 11. Sentiment Analysis of Literal and Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015),* 704–708.

Bateman, J., & Farrar, S. (2004). Towards a generic foundation for spatial ontology. In *International Conference on Formal Ontology in Information Systems (FOIS-2004),* 237–248. Trento, Italy.

Batista, F., & Ribeiro, R. (2013). Sentiment Analysis and Topic Classification based on Binary Maximum Entropy Classifiers. *Journal Procesamiento del Lenguaje Natural, 50*, 77–84.

Beck, R C. (2000). *Motivation: Theories and Principles* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *COLING 2004 Workshop on Multilingual Linguistic Resources,* 101–108. Geneva, Switzerland.

Berger, A. L., Della Pietra, V. J., & Della Pietra, S. A. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22*(1), 39–71. MIT Press.

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications, 42*, 3105-3114.

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In Lüdeling, A. and Kytö, M. (Eds.), *Corpus Linguistics - an International Handbook*, *2*, 1286–1304. Berlin: Walter de Gruyter.

Bjekić, J., Lazarević, B. Lj., Živanović, M., & Knežević, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis – LIWCse. *Psihologija*, *47*(1), 5–32.

Bjekić, J., Lazarević, Lj., Erić, M., Stojimirović, E., & Đokić, T. (2012). Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser). *Psihološka Istraživanja, 15*(1), 85–110.

Blinov, P., Klekovkina, M., Kotelnikov, E., & Pestov, O. (2013). Research of Lexical Approach and Machine Learning Methods for Sentiment Analysis. In *Proceedings of Dialog*, *2*, 51–61.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2009). A Relational Model of Semantic Similarity Between Words Using Automatically Extracted Lexical Pattern Clusters from the Web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2*, 803–812.

Brickley, D. (1999). WordNet in RDF/XML: 50,000+ RDF class vocabulary. Message to RDF Interest Group. Retrieved from: http://lists.w3.org/Archives/Public/www-rdf-interest/1999Dec/0002 (September 20, 2015.)

Brković, Aleksa. (2011). *Razvojna psihologija*. Regionalni centar za profesionalni razvoj zaposlenih u obrazovanju. Čačak.

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooollllllllllllll!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011)*, 562–570.

Bruce, F. R., & Wiebe, J. M. (1999). Recognizing Subjectivity: A Case Study of Manual Tagging. *Natural Language Engineering*, *5*, 187–205.

Buche, A., Chandak, M. B., & Zadgaonkar, A. (2013). Opinion Mining and Analysis: A survey. *International Journal on Natural Language Computing (IJNLC) 2*(3).

Buitelaar, P. (2005). Ontology-based semantic lexicons: mapping between terms and object description. In *Ontologies and Lexical Resources: IJCNLP-05 Workshop*, 1–12.

Cambria, E., & Hussain, A. (2012). *Sentic Computing Techniques, Tools, and Applications 2nd Edition*. SpringerBriefs in Cognitive Computation.

Cambria, E., & Hussain, A. (2015). *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Springer.

Cambria, E., Hussain, A., & Havasi, C. (2010). Towards the Crowd Validation of the British National Health Service. *In Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC, US.

Cambria, E., Livingstone, A., & Hussain, A. (2012). The Hourglass of Emotions. In *Proceedings of the 2011 International Conference on Cognitive Behavioural Systems, COST'11*, 144–157. Springer.

Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI*, 1515-1521.

Cambria, E., Rajagopal, D., Olsher, D., & Das, D. (2013). Big social data analysis. In R. Akerkar (Ed.) *Big Data Computing*. Chapman and Hall/CRC.

Caragea, C., Squicciarini, A., Stehle, S., Neppalli, K., & Tapia, A. H. (2014). Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy. In *Proceedings of the 11th International Conference on Information Systems for*

Crisis Response and Management (ISCRAM 2014). University Park, Pennsylvania, USA.

Carenini, G., Ng, R. T., & Zwart, E. (2005). Extracting Knowledge from Evaluative Text. In *Proceedings of the 3rd International Conference on Knowledge Capture* (*K-CAP '05*), 11–18.

ChandraKala, S., & Sindhu, C. (2012). Opinion Mining and Sentiment Classification: A survey. *Ictact Journal on Soft Computing, 03*(01).

Chandrasekaran, B., Josephson, J.R., & Benjamins, V.R. (1998). Ontology of Tasks and Methods. Ohio-State University. Retrieved from: http://ksi.cpsc.ucalgary.ca/KAW/KAW98/chandra/index.html (September 30, 2015.)

Charniak, E. (2000). A Maximum-entropy-inspired Parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference,* 132–139. ACL.

Choi, Y., & Wiebe, J. (2014). +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-2014)*, 107–112. ACL.

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, *16*(1), 22–29. MIT Press.

Cogburn, D. L., Hanson, M. E., & Wozniak, A. (2012). Accelerating social sciences for the new age. Moving from traditional methods for analyzing large scale textual data to socially high performance computational methods. *Computer Supported Cooperative Work J2*. Seattle. Washington. USA

Courtois, B., & Silberztein, M. (1990). Dictionnaires électroniques du français. *Langue française, 87*(1),3-4. Larousse, Paris.

Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. *Cambridge University Press*. New York, NY, USA.

Dang, Y., Zhang, Y., & Chen, H. (2009). A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews. *Intelligent Systems, IEEE, 25*(4), 46 –53.

Darroch, J. N., & Ratcliff, D. (1972). Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics, 43*(5), 1470-1480.

Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, 519–528.

Davidov, D., Tsur, O., & Rappoport, A. (2010). A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM-2010)*.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of Coling 2010*.

de Albornoz, J. C., Plaza, L., & Gervás, P. (2012). SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. ELRA.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing Features of Random Fields. In *IEEE Trans. Pattern Anal. Mach. Intell, 19*(4), 380–393. IEEE Computer Society.

Denecke, K. (2008). Using SentiWordNet for Multilingual Sentiment Analysis. *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on data engineering workshop*, 507–512.

Devedžić, V. (2002). Understanding Ontological Engineering. *Communications of the ACM, 45*(4), 136–144.

Devedžić, V. (2006). *Semantic Web and Education, Monograph*. Springer, Berlin Heidelberg New York.

Dias, H. G., Hasanuzzaman, M., Ferrari, S., & Mathet, Y. (2014). TempoWordNet for Sentence Time Taging. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, 833–838.

Dobrić, N. (2012). Language Corpora in The West Balkans – History, Current State and Future Perspective. *Slavistična revija 60*(4), 677–692.

Dragićević, R. (2010). *Leksikologija srpskog jezika*. Zavod za udžbenike, Beograd.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification* (2nd Edition). Wiley-Interscience.

Durand, J. (1987). Rhetorical Figures in the Advertising Image. In *Marketing and Semiotics: New Directions in the Study of Signs for Sale*, 295–31. New York: Mouton de Gruyter.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200.

Elkan, C. (2012). Evaluating Classifiers. Retrieved from http://cseweb.ucsd.edu/~elkan/250Bwinter2012/classifiereval.pdf

Esuli, A., & Sebastiani, F. (2006). Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET:A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 417–422.

Evert, S., Proisl, T., Greiner, P., & Kabashi, B. (2014). SentiKLUE: Updating a Polarity Classifier in 48 Hours. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 551–555.

Farrar, S., & Langendoen, T. D. (2003). A linguistic ontology for the Semantic Web. *GLOT International, 7*(3), 97–100.

Fellbaum, C. (Ed). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Feng, S., Bose, R., & Choi, Y. (2011). Learning general connotation of words using graph-based algorithms. In *Proceedings of Confernece on Empirical Methods in Natural Language Processing (EMNLP-2011)*, 1092–1103. ACL.

Filatova, E. (2012). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Franzoni, V., Poggioni, V., & Zollo, F. (2013). Automated Classification of Book Blurbs According to the Emotional Tags of the Social Network Zazie. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI, ESSEM@AI*IA*, 83–94.

Gamallo, P., & Garcia, M. (2014). Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 171–175.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, 841–847.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2003). Restructuring WordNet's Top-level. *AI Magazine*, *40*.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2005). Interfacing WordNet with DOLCE: toward OntoWordNet. In *Ontologies and Lexical Resources: IJCNLP-05 Workshop*, 1–12.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2010). Interfacing WordNet with DOLCE: towards OntoWordNet. *Ontology and the Lexicon*, 36-52. http://dx.doi.org/10.1017/CBO9780511676536.004

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2003). Sweetening WORDNET with DOLCE. *AI Magazine, 24*(3), 13–24.

Gašević, D., Djurić, D., & Devedžić, V. (2006). *Model Driven Architecture and Ontology Development*. Springer-Verlag New York, Inc.

Gaudette, L., & Japkowicz, N. (2011). Compact Features for Sentiment Analysis. *Canadian Conference on AI*, 146–157. Lecture Notes in Computer Science 6657. Springer.

Gawryjolek, J. J. (2009). Automated Annotation and Visualization of Rhetorical Figures (Master thesis). Universty of Waterloo.

Gawryjolek, J., Di Marco, C., & Harris, R. A. (2009). An Annotation Tool for Automatically Detecting Rhetorical Figures – System Demonstration. In *Proceedings of the IJCAI-09 workshop on Computational Models of Natural Argument*. Pasadena, CA.

Gerber, S. M. (2013). WordNetEngine.cs, Synset.cs. Retrieved from: https://ptl.sys.virginia.edu/msg8u/NLP/Source/ResourceAPIs/WordNet/WordNet/ (September 02, 2013)

Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., & Tserpes, K. (2012). Representation Models for Text Classification: A Comparative Analysis over Three Web Document Types. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, (pp. 13:1-13:12). ACM.

Gianti, A., Bosco, C., Patti, V., Bolioli, A., & Di Caro, L. (2012). Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 1–7.

Gibbs, R. W. Jr., & Colston, H. L. (2012). *Interpreting Figurative Meaning*. Cambridge University Press.

Giunchiglia, F., Maltese, V., Farazi, F., & Dutta, B. (2010). GeoWordNet: A Resource for Geo-spatial Applications. *The Semantic Web: Research and Applications,* 121–136. Lecture Notes in Computer Science 6088. Springer-Verlag, Berlin, Heidelberg.

Glasgow, K., Fink, C., & Boyd-Graber, J. (2014). Our grief is unspeakable: Measuring the community impact of a tragedy. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphors to Idioms*. Oxford University Press.

Glushko, R. J. (Ed.). (2013). *The Discipline of Organizing*. Cambridge, MA: MIT Press.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Technical report*. Stanford Digital Library Technologies Project.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011)*.

Goodman, J. (2001). A bit of progress in language modeling. *Technical Report MSR-TR-2001-72.* Machine Learning and Applied Statistics Group Microsoft Research.

Graovac, J. (2012). Serbian text categorization using byte level n-grams. In Z. Budimac, M. Ivanović and M. Radovanović (Eds.), *Proceedings of CloBL 2012: Workshop on Computational Linguistics and Natural Language, 5th Balkan Conference in Informatics* (pp. 93–97). Faculty of Sciences, Department of Mathematics and Informatics, Novi Sad, Serbia.

Graovac, J. (2014). A variant of n-gram based language-independent text categorization. *Intelligent Data Analysis, 18*(4).

Graves, A., & Gutierrez, C. (2006). Data representations for WordNet: A case for RDF. In Proceedings of the 3rd International WORDNET Conference, 165–169.

Grishman, R., Macleod, C., & Meyers, A. (1994). Comlex Syntax: Building a Computational Lexicon. In *Proceedings of the 15th Conference on Computational Linguistics, 1*: 268–272.

Gross, M. (1989). La construction de dictionnaires électroniques. *Annales des télécommunications, 44*(1-2):4-19.

Gross, M. (1988). The Use of Finite Automata in the Lexical Representation of Natural Languages. In Gross & Perrin (Eds.), *Electronic Dictionaries and Automata in Computational Linguistics*, LNCS, 337, 34-50. Springer, Berlin.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199–220. Retrieved from: http://tomgruber.org/writing/ ontolingua- kaj-1993.htm (September 20, 2015.)

Gupta, R., Nisheeth, J., & Mathur, I. (2013). Analysing Quality Of English-Hindi Machine Translation Engine Outputs Using Bayesian Classification. *International Journal of Artificial Intelligence & Applications (IJAIA), 4*(4).

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *The Journal of Machine Learning Research, 3*(26), 1157–1182.

Habernal, I., Ptáček, T., & Steinberger, J. (2013). Sentiment analysis in czech social media using supervised machine learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 65–74.

Habernal, I., Ptàček, T., & Steinberger, J. (2014). Supervised Sentiment Analysis in Czech Social Media. *Journal Information Processing and Management, Pergamon Press*, *50*(5), 693–707.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Journal Decision Support Systems*, *55*(3), 685–697.

Hajmohammadi, M. S., Ibrahim, R., & Ali Othman, Z. (2012). Opinion Mining and Sentiment Analysis: A Survey. *International Journal of Computers & Technology, 2*(3), 171–178.

Hao, Y., & Veale, T. (2010). An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes. *Journal Minds and Machines, 20*(4), 635–650.

Harris, R. A. (2011). A Handbook of Rhetorical Devices. Retrieved from: http://www.virtualsalt.com/rhetoric.htm (September 20, 2015.)

Harris, R. A., & Di Marco, C. (2009). Constructing a rhetorical figuration ontology. *Symposium on Persuasive Technology and Digital Behavior Intervention,*

Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), 47–52. Edinburgh.

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 174–181.

Hatzivassiloglou, V., & Wiebe, M. J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the 18th Conference on Computational Linguistic, 1*, 299–305.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures, 1*(1), 77–89.

Hobbs, J. R., & Gordon, A. S. (2011). The Deep Lexical Semantics of Emotions, In, *Affective Computing and Sentiment Analysis, Text, Speech and Language Technology 45*, 27–34.

Hristea, T. F. (2013). The Naïve Bayes Model in the Context of Word Sense Disambiguation. *Springer Briefs in Statistics*, 9–16.

Hu, H., Du, X., Liu, D., & Ouyang, J. H. (2006). Ontology Learning using WordNet Lexicon. *Computational Methods*, 1249–1253. Springer Netherlands.

Hu, M., & Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 168–177.

Hu, M., & Liu, B. (2004b). Mining Opinion Features in Customer Reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence, AAAI'04*, 755–760. AAAI Press.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Journal Decision Support Systems, 52*(3), 674–684.

Huang, X., & Zhou, C. (2007). An OWL-based WordNet lexical ontology. *Journal of Zhejiang University SCIENCE, 8*,(6), 864–870.

Ikeda, M., Seta, K., & Mizoguchi, R. (1997). Task Ontology Makes It Easier To Use Authoring Tools. *IJCAI-97*, 342–347. Nagoya Japan.

Inkpen, D., Keshtkar, F., & Ghazi, D. (2009). Analysis and Generation of Emotion in Texts. In *Proceedings of KEPT 2009 Knowledge Engineering-Principles and Techniques*, 3–13.

Jacob, E. K. (1991). Classification and categorization: Drawing the Line. *Advances in Classification Research Online, [S.l]*, 63–80.

Jacob, E. K. (2004). Classification and categorization: a difference that makes a difference. *Library Trends, 52*(3), 515–540.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer Series in Statistics.

Janičić, P., & Nikolić, M. (2010). Veštačka inteligencija. Retrieved from: http://poincare.matf.bg.ac.rs/~janicic//courses/vi.pdf (September 20, 2015)

Jaynes, E. T. (1991). Notes On Present Status And Future Prospects. In W. T. Grandy, Jr. and L. H. Schick (Eds.), *Maximum Entropy and Bayesian Methods*. Kluwer, Dordrecht.

Jiliang, T., Salem, A., & Huan, L. (2014). Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications*, 37–64. CRC Press.

Johansson, R., & Moschitti, A. (2013). Relational Features in Fine-Grained Opinion Analysis. *Computational Linguistics, 39*(3), 473–509.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In William W. Cohen and Haym Hirsh, (Eds.), *Machine Learning: Proceedings of the Eleventh International Conference*, 121–129. Morgan Kaufmann.

Johnson, C., Shukla, P., & Shukla, S. (2012). On classifying the political sentiment of tweets. Retrieved from: https://www.cs.utexas.edu/~cjohnson/TwitterSentimentAnalysis.pdf (September 20, 2015.)

Jolić, N. (2014). Klasifikacija sentimenta u twitter postovima korišćenjem udaljenog nadzora. Elektrotehnički fakultet. Retrieved from: http://home.etf.rs/ ~vm /os/ dmsw/2014/ (September 20, 2015.)

Jolliffe, I. (2002). Principal Component Analysis. *Springer Series in Statistics* (2nd Edition). Springer.

Jovanović, R., & Atanacković, L. (1980). *Sistematski rečnik srpskohrvatskoga jezika*. Matica srpska, Novi Sad.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (2nd ed.). Prentice Hall.

Kaji, N., & Kitsuregawa, M. (2007). Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1075–1083. ACL.

Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, *4*, 1115–1118. Lisbon, PT.

Kanayama, H., & Nasukawa, T. (2006). Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 355–363. ACL.

Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and Improved Naive Bayes Algorithms for Sentiment Analysis of Restaurant Reviews. *Expert Systems with Applications, 39*(5), 6000–6010.

Kaushik, C., & Mishra, A. (2014). A Scalable, Lexicon Based Technique for Sentiment Analysis. *International Journal in Foundations of Computer Science & Technology (IJFCST), 4*(5) .

Kelly, A. R., Abbott, N. A., Harris, R. A., Di Marco, C., & Cheriton, D. R. (2010). Toward an ontology of rhetorical figures. In *Proceedings of the 28th ACM International Conference on Design of Communication SIGDOC '10*, 123–130.

Kennedy, A., & Inkpen, D. (2006). Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence (special issue), 22*(2), 110–125.

Khairnar, J., & Kinikar, M. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classication. *International Journal of Scientic and Research Publications (IJSRP), 3*(6).

Khairnar, J., & Kinikar, M. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classification. *International Journal of Scientific and Research Publications, 3*(6).

Kim, H. D., Ganesan, K., Sondhi, P., & Zhai, C. (2011). Comprehensive Review Of Opinion Summarization (Survey). *Technical report*. University of Illinois at Urbana-Champaign.

Kim, S., & Hovy, E. (2004). Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1367–1373.

Kirtsis, N., Tzekou, P., Besharat, J., & Stamou, S. (2013). Identifying Polarized Wikipedia Articles. In P.O. de Pablos, H. O. Nigro, R. D. Tennyson & S. E. Gonzalez Cısaro (Eds.), *Advancing Information Management through Semantic Web Concepts and Ontologies*. IGI Global.

Koeva, S., Krstev, C., & Vitas, D. (2008). Morpho-semantic Relations in WordNet - a Case Study for two Slavic Languages. In *Proceedings of Global WordNet Conference*, 239–253. University of Szeged.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based Sentiment Analysis of Twitter Posts. *Expert Systems with Applications 40*, 4065–4074.

Kostić, A. (2014). Electronic Corpus Of Sebian Language From 12th To 18th Century. *Pregled NCD, 24*.

Kraychev, B., & Koychev, I. (2011). Classification of Online Reviews by Computational Semantic Lexicons. *Third International Conference on Software, Services and Semantic Technologies S3T 2011. Advances in Intelligent and Soft Computing*, *101*, 9–16.

Krikorian, R. (2013). New Tweets per second record, and how! Twitter Official Blog. Retrieved from: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

Krstev, C. (2008). *Processing of Serbian: Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology,

Krstev, C., & Vitas, D. (2005). Corpus and Lexicon - Mutual Incompletness. In Pernilla Danielsson and Martijn Wagenmakers (Eds.), *Proceedings of the Corpus Linguistics Conference*. Birmingham, UK. Retrieved from: http://www.corpus.bham.ac.uk/PCLC/

Krstev, C., & Vitas, D. (2009). An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds. In *Proceedings of The 28th Conference on Lexis and Grammar*, 204–212.

Krstev, C., Đorđević, B., Antonić, S., Ivković-Berček, N., Zorica, Z., Crnogorac, V., & Macura, Lj. (2008). Cooperative Work in Further Development of Serbian WordNet. *INFOtheca 9*(1-2), 59a–78a.

Krstev, C., Pavlović-Lažetić, G., Vitas, D., & Obradović, I. (2004). Using Textual and Lexical Resources in Developing Serbian Wordnet. In *Romanian Journal of Information Science and Technology, 7*(1-2), 147–161. Romanian Academy, Publishing House of the Romanian Academy,

Krstev, C., Stanković, R., Obradović, I., Vitas, D., & Utvić, M. (2010). Automatic Construction of a Morphological Dictionary of Multi-Word Units. *LNCS, 6233*, 226–237. Springer.

Krstev, C., Stanković, R., Vitas, D., & Obradović, I. (2006). WS4LR - a Worksation for Lexical Resources. In *Proceedings of the Fifth Interantional Conference on Language Resources and Evaluation, LREC 2006*, 1692–1697. Genoa, Italy.

Krstev, C., Vitas, D., Stanković, R., Obradović, I., & Pavlović-Lažetić, G. (2004). Combining Heterogeneous Lexical Resources. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, *4*, 1103–1106. Lisabon, Portugal.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282–289). Morgan Kaufmann.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. The University of Chicago Press, Chicago, IL.

Lau, R., Rosenfeld, R., & Roukos, S. (1993). Adaptive Language Modeling Using the Maximum Entropy Principle. In *Proceedings of the Workshop on Human Language Technology*, 108–113. ACL.

Lee, H. Y., & Renganathan, H. (2011). Chinese Sentiment Analysis Using Maximum Entropy. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011,* 89–93.

Lertsuksakda, R., Pasupa, K., & Netisopakul, P. (2015). Sentiment Analysis on Thai Children Stories with Support Vector Machine. In *Proceeding of the 20th International Symposium on Artificial Life and Robotics (AROB 20th 2015)*, 138–142. Japan.

Lewis, D. D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, 4–15. Springer-Verlag, London, UK, UK.

Liu, B. (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. *Data-Centric Systems and Applications* (2nd ed.). Springer.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, B. (2010). Sentiment Analysis and Subjectivity. In Nitin Indurkhya and Fred J. Demerau (Ed.), *Handbook of Natural Language Processing, Second Edition* (pp. 627-666). CRC Press, Taylor and Francis Group.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating Classification and Association Rule Mining. *Knowledge Discovery and Data Mining*, 80–86.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th International Conference on World Wide Web* (*WWW '05*), 342–351.

Liu, C., & Fujisawa, H. (2005). Classiffication and Learning for Character Recognition: Comparison of Methods and Remaining Problems. *Int. Workshop on Neural Networks and Learning in Document Analysis and Recognition*. Seoul.

Liu, J., & Seneff, S. (2009). Review Sentiment Scoring via a Parse-and-paraphrase Paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, *1*, 161–169.

Lombard, M., Snyder-Duchand, J., & Campanella Bracken, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604.

Long, C., Zhang, J., & Zhut, X. (2010). A Review Selection Approach for Accurate Feature Rating Estimation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 766–774. ACL. (Poster paper).

Lönneker-Rodman, B. (2008). The Hamburg Metaphor Database project: issues in resource creation. *Language Resources and Evaluation 42*(3), 293–318.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 142–150. ACL.

Maggetti, M. (2013). Regulation in Practice: The defacto Independence of Regulatory. *Swiss Political Science Review, 19*(1), 111–113.

Magnini, B., & Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. ELRA.

Manco, G., Masciari, E., Ruffolo, M., & Tagarelli, A. (2002). Towards An Adaptive Mail Classifier. *Italian Association for Artificial Intelligence Workshop Su Apprendimento Automatico: Metodi Ed Applicazioni*.

Manning, C D., Raghavan, P., & Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Maron, M.E., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216–244.

Marrafa, P., Amaro, R., Chaves, P. R., Lourosa, S., Martins, C., & Mendes, S. (2006). WordNet.PT new directions. In *Proceedings of the 3th International Global Wordnet Conference (GWC2006)*, 319–321.

Martineau, J., & Finin, T. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media*. The AAAI Press. San Jose, CA. (Poster paper).

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, 41–48. Technical Report WS-98-05. AAAI Press.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 591–598. Morgan Kaufmann Publishers Inc.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal 5*(4), 1093–1113.

Mehra, N., Khandelwal, S., & Patel, P. (2002). Sentiment Identification Using ME Analysis of Movie Review. *Working paper*. Retrieved from: http://web.stanford.edu/class/cs276a/projects/reports/nmehra-kshashi-priyank9.pdf.

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, 171–180. ACM, NewYork, NY, USA.

Melnik, S., & Decker, S. (2001). Wordnet RDF Representation. Retrieved from: http://www.semanticweb.org/library/ (September 20, 2015.)

Mendes, S. (2006). Adjectives in WordNet. In *Proceedings of the 3th International Global Wordnet Conference (GWC2006)*, 225–230.

Meng, L., Gu, J., & Zhou, Z. (2012). A New Model of Information Content Based on Concept's Topology for Measuring Semantic Similarity in WordNet. *International Journal of Grid and Distributed Computing, 5*(3).

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM 38*(11), 39–41.

Milošević, N. (2012). Mašinska analiza sentimenta rečenica na srpskom jeziku (Master rad). Elektrotehnički fakultet, Beograd. Retrieved from: www.inspiratron.org/Analiza_sentimenta_recenica_na_srpskom_jeziku-Nikola_Milosevic.pdf (September 20, 2015.)

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of the 1st Workshop on Stylistic Analysis Of Text For Information Access, 19*.

Missen, M. M. S., Boughanem, M., & Cabanac, G. (2009). Challenges for Sentence Level Opinion Detection in Blogs. *ACIS-ICIS, IEEE Computer Society*, 347–351.

Mitchel, T. M. (1997). *Machine Learning*. McGraw Hill.

Mitrović, J. (2014). Electronic Tools and Resources for Multi-Word Unit Detection and Research in Serbian. *The 2th General Meeting of The IC1207 COST Action, PARSEME*. Athens, Greece, 10-11 March, 2014. (Poster paper)

Mitrović, J., Mladenović, M., & Krstev, C. (2015). Adding MWEs to Serbian Lexical Resources Using Crowdsourcing. Poster at *The 5th PARSEME general meeting*. Iași, Romania, 23-24 September 2015.

Mladenović, M., & Mitrović, J. (2013). Ontology of Rhetorical Figures for Serbian. *Text, Speech and Dialogue, 8082*, 386–393, LNCS, Springer.

Mladenović, M., & Mitrović, J. (2014). Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. In G. Pavlović Lažetić, C. Krstev, I. Obradović & D. Vitas *Natural Language Processing for Serbian – Resources and Application*, 1-11. Matematički fakultet, Beograd.

Mladenović, M., Mitrović, J., & Krstev, C. (2014). Developing and Maintaining a WordNet: Procedures and Tools. In H. Orav, C Fellbaum & P Vossan (Eds.), *Proceedings of Seventh Global WordNet Conference 2014*, 55–62. University of Tartu, Tartu, Estonia.

Mladenović, M., Mitrović, J., Krstev, C., & Vitas, D. (2015). Hybrid Sentiment Analysis Framework For A Morphologically Rich Language. *Journal of Intelligent Information Systems* (In Press, Available online 15 August 2015).

Mladenović, M., Mitrović, J., & Krstev, C. (2016). Introducing a Language-independent Model for Adding a New Semantic Relation Between Adjectives and Nouns in a WordNet. In *Proceedings of Eight Global WordNet Conference 2016,* 218-225. Romanian Academy Library, Bucharest, Romany.

Mladenović, P. (2008). *Verovatnoća i statistika*. Matematički fakultet, Beograd.

Mohammad, M. S., & Yang, T. W. (2011). Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 70–79.

Mohammad, S. (2011). From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105–114.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence, 29*(3), 436–465.

Mohammad, S., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Mohtarami, M., Amiri, H., Lan, M., Tran, T. P., & Tan, C. L. (2012). Sense Sentiment Similarity: An Analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Mohtarami, M., Lan, M., & Tan, C. L. (2013). *From Semantic to Emotional Space in Probabilistic Sense Sentiment Analysis*. AAAI Press.

Moravec, P., Kolovrat, M., & Snásel, V. (2004). LSI vs. Wordnet Ontology in Dimension Reduction for Information Retrieval. In *Proceedings of the Dateso 2004 Annual International Workshop on Databases, Texts, Specifications and Objects*, 18–26.

Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 412–418.

Mullen, T., & Malouf, R. (2006). A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 159–162.

Mullen, T., & Malouf, R. (2008). Taking sides: User classification for informal online political discourse. *Internet Research, 18*, 177–190.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts, London, England.

Myung, Jae. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Nenadić, G. (2004). Creating digital language resources. *Review Of The National Center For Digitization, 5*, 19–30.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Compositionality Principle in Recognition of Fine-Grained Emotions from Text. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM*. The AAAI Press.

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. In T.G. Dietterich and S.

Becker and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14* (pp. 841–848). MIT Press. Pages.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum etropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, 61–67.

Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, 412–416.

Obradović, I., & Stanković, R. (2008). Software tools for Serbian lexical resources. *INFOtheca, 9*(1–2), 43a–57a.

Obrst, L. (2010). Ontological Architectures. In Roberto Poli, Michael Healy and Achilles Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications*. Springer Publishing Company, Incorporated.

Ohana, B., & Tierney, B. (2009). Sentiment Classification of Reviews Using SentiWordNet. *9th. IT & T Conference*.

Ortega Bueno, R., Fonseca Bruzon, A., Muniz Cuza, C., Gutierrez, Y., & Montoyo, A. (2014). UO_UA: Using Latent Semantic Analysis to Build a Domain-Dependent Sentiment Resource. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 773–778.

Ortigosa, A., Martín, M. J., & Carro, M. R. (2014). Sentiment analysis in Facebook and its application to e-learning. J*ournal Computers in Human Behavior, 31*, 527–541.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge University Press.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The Referential Structure of the Affective Lexicon. *Cognitive Science, 11*, 341–364.

Øye, J. A. (2015). Sentiment Analysis of Norwegian Twitter Messages (Master thesis). Norwegian University of Science and Technology. Retrieved from: https://daim.idi.ntnu.no/masteroppgave?id=12125 (September 20, 2015.)

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC 2010*.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics* (pp. 271–278).

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL-05, 43rd Meeting of the Association for Computational Linguistics*, 115–124.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1–135. Retrieved from: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing (EMNLP)*, *10*, 79–86.

Pavlović-Lažetić, G., & Tomašević, J. (2010). Ontology-driven conceptual document classification. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 383–386.

Peacham H. (1593). *The Garden of Eloquence*. Perseus Digital Library. Retrieved from: http://rhetoric.byu.edu/primary%20texts/Peacham.htm (September 10, 2015.)

Pease, A. (2011). *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.

Pennebaker, J. W., Booth, R J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: A computerized text analysis program. Retrieved from: www.liwc.net

Perera, R.D.W., Anand, S., Subbalakshmi, K.P., & Chandramouli, R. (2010). Twitter Analytics: Architecture, Tools and Analysis. *MILCOM 2010*, 2186–2191.

Petrović, M. (1967). *Metafore i alegorije*. Srpska književna zadruga, Beograd.

Plutchik, R., & Hope, R. C (1997). *Circumplex Models of Personality and Emotions*. American Psychological Association.

Poli, R. (2003). Descriptive, formal and formalized ontologies. In *Husserl's Logical Investigations Reconsidered*, *48*, 193–210. Springer Netherlands.

Popescu, A., & Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 339–346.

Popescu, A., Yates, A., & Etzioni, O. (2004). Class Extraction from the World Wide Web. In *Proceedings of AAAI 2004 Workshop on Adaptive Text Extraction and Mining (ATEM'04)*, 68–73.

Poria, S., Cambria, E., Hussain, A., & Hunag, G. (2015). Towards an intelligent framework for multimodal affective data analysis. *Journal Neural Networks, 63*, 104–116.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined Approach. *Informetrics 3*(2), 143–157.

Prévot, L., Borgo, S., & Oltramari, A. (2005). Interfacing ontologies and lexical resources. In *Ontologies and Lexical Resources: IJCNLP-05 Workshop*, 1–12.

Ptaszynski, M., Rzepka, R., Araki, K. & Momouchi, Y. (2011). Research on Emoticons: Review of the Field and Proposal of Research Framework. In *Proceedings of The Seventeenth Annual Meeting of The Association for Natural Language Processing (NLP-2011), Organized Session on Un-Natural Language Processing*, 1159–1162. Toyohashi, Japan.

Qu, L., Ifrim, G., & Weikum, G. (2010). The Bag-of-opinions Method for Review Rating Prediction from Sparse Text Patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 913–921.

Radunović, D. P. (2003). *Numeričke metode*. Akademska misao. Beograd.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP-96*.

Ratnaparkhi, A., Roukos, S., & Ward, T. (1994). A Maximum Entropy Model For Parsing. *The 3rd International Conference on Spoken Language Processing, (ICSLP)*, 803–806.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems* (In Press, Available online 29 June 2015).

Rentoumi, V., Petrakis, S., Klenner, M., Vouros, A. G., & Karkaletsis, V. (2010). United we stand - improving sentiment analysis by joining machine learning and rule based methods. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1*, 448—453. Morgan Kaufmann Publishers Inc.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research, 11*, 95–130.

Reyes, A., & Rosso, P. (2012a). Building Corpora for Figurative Language Processing: The Case of Irony Detection. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals*, 94–98.

Reyes, A., & Rosso, P. (2012b). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems, 53*(4), 754–760.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 105–112. ACL.

Riloff, E., Patwardhan, S., & Wiebe, J. (2006). Feature Subsumption for Opinion Analysis. *EMNLP, COLING '04*, 440–448. ACL.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning Subjective Nouns using Extraction Pattern Bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 4*, 25–32.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. *Text Mining: Applications and Theory*, 1–20. John Wiley and Sons.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language*, *10*(3), 187–228.

Russell, S., & Norvig, P. (2011). *Veštačka inteligencija - Savremeni pristup*. CET Computer Equipment and Trade, Beograd.

Saif, H., He, Y., & Alani, H. (2012). Semantic Sentiment Analysis of Twitter. In *Proceedings of the 11th International Conference on The Semantic Web, 1*, 508–524.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM, 26*(11), 1022–1036.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Samardžić, T. (2011). Language corpora as a source of new data for lexicographic description of Serbian. *International Slavic conference* (paper 40/1). Belgrade: International Slavic centre. (in Serbian).

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*, 211–229.

Scherer, K., & Ekman, P. (1984). Expression And The Nature Of Emotion. In Scherer, K. & Ekman, P. (Eds.), *Approaches to Emotion*, 319-343. Hillsdale, NJ: Lawrence Erlbaum.

Schulz, A., Thanh, T. D., Paulheim, H., & Schweizer, I. (2013). A Fine-Grained Sentiment Analysis Approach for Detecting Crisis Related Microposts. In *Proceedings of 10th Internation Conference on Information Systems for Crisis Response and Management*, 846–851.

Schwartz, D. B. (2009). Figurative Language and Rhetorical Devices. Retrieved from: http://archive.is/http://cla.calpoly.edu/~dschwart/engl331/ figurative. html (September 20, 2015)

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Seerat, B., & Azam, F. (2012). Opinion Mining: Issues and Challenges (A survey). *International Journal of Computer Applications, 49*(9).

Sloane, T. O. (2001). *Encyclopedia of Rhetoric*. Oxford University Press, New York, USA.

Smith, P., & Lee, M. (2012). A CCG-based Approach to Fine-Grained Sentiment Analysis. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology, COLING 2012*, 3-16.

Solar, M. (1987). *Teorija književnosti*. Školska knjiga, Zagreb.

Somasundaran, S., Ruppenhofer, J., & Wiebe, J. (2007). Detecting arguing and sentiment in meetings. *In Proceedings of the SIGdial Workshop on Discourse and Dialogue*.

Sowa, J. F. (1992). Semantic Networks. In *S. C. Shapiro Encyclopedia of Artificial Intelligence.* Wiley, New York 1987, (revised and extended for the second edition, 1992).

Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Tufis, D., Koeva, S., Totkov, G., Totkov, D., & Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for Balkan Languages. In *Proceedings of the 1st International Global WordNet Conference*. Mysore, India.

Stanimirović, Z. (2014). *Nelinearno programiranje*. Matematički fakultet, Univerzitet u Beogradu.

Stanković, R., Krstev, C., Obradović, I., Trtovac, A., & Utvić, M. (2012). A tool for enhanced search of multilingual digital libraries of e-journals. In *Proceedings of the 8th International Conference on Language Resources and Evaluation,* 1710-1717. ELRA.

Stone, P. J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Strapparava, C., & Mihalcea, R. (2008). Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, 1556–1560.

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* 1083–1086. ELRA.

Strongman, K. T. (2003). *The Psychology of Emotion: From Everyday Life to Theory, Fift Edition*. John Wiley &Sons, England.

Su, C., & Srihari, S. N. (2011). Generative Models and Probability Evaluation for Forensic Evidence. In P. Wang (Ed.), *Pattern Recognition, Machine Intelligence and Biometrics*. Springer.

Surendran, C., A., Platt, J. C., & Renshaw, E. (2005). Automatic Discovery of Personal Topics to Organize Email. *CEAS 2005 - Second Conference on Email and Anti-Spam*, 1–6. Stanford University, California, USA. Retrieved from: http://ceas.cc/2005/ (September 10, 2015.)

Suttles, J., & Ide, N. (2013). Distant Supervision for Emotion Classication with Discrete Binary Values. In *Proceedings of the 14th International Conference, CICLing 2013, 7817*, 121-136. Springer.

Ševa, N., & Kostić, A. (2003). Anotirani korpus i evaluacija procene verovatnoća gramatičkih oblika. *Psihologija, 36*(3), 255–270.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis, *Computational. Linguistics, 37*(2), 267–307.

Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, ECIR '09*, 337–349. Springer-Verlag.

Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2014). Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Paper*, 172–182. Dublin, Ireland

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. In Charu Aggarwal (Ed.) *Data Classification: Algorithms and Applications*, 37–64. CRC Press.

Tartalja, I. (2003). *Teorija književnosti*. Zavod za udžbenike i nastavna sredstva, Beograd.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

TenHouten, W. D. (2007). *A General Theory of Emotions and Social Life*. Routledge, Taylor and Francis Group, London.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63*(1), 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558.

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In *Proceedings of the Workshop on Operational Text Classification (OTC)*.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kubler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., & Tounsi, L. (2010). Statistical Parsing of Morphologically Rich Languages (SPMRL): What, How and Whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, SPMRL '10*, 1–12.

Turner, M. (2002). The Cognitive Study of Art, Language, and Literature. *Poetics Today Spring 23*(1), 9–20.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, 417–424.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning, EMCL '01*, 491–502. Springer.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, *21*(4), 315–346.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 141–188.

Utvić, M. (2011). Annotating the Corpus of Contemporary Serbian. *INFOtheca, 12*(2), 36a–47a.

Utvić, M. (2014a). *Construction of a reference corpus of contemporary Serbian language* (doctoral dissertation). Faculty of Philology, Belgrade, Serbia,

Utvić, M. (2014b). Liste učestanosti Korpusa savremenog srpskog jezika [Corpus of Contemporary Serbian – Frequency Lists]. *Naučni sastanak slavista u Vukove dane*, 241–262.

van Assem, M., Gangemi, A., & Schreiber, G. (2006). RDF/OWL Representation of WordNet. Editor's Draft. Retrieved from: http://www.w3.org/2001/sw /bestpractices/wnet/wn-conversion.html (September 20, 2015.)

van Assem, M., Menken, M. R., Schreiber, G., Wielemaker, J., & Wielinga, B. J. (2004). A Method for Converting Thesauri to RDF/OWL. *International Semantic Web Conference*, 17–31.

van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, Newton, MA, USA.

Veale, T., & Hao, Y. (2008). Enriching WordNet with folk knowledge and stereotypes. In *Proceedings of the 4th International Global Wordnet Conference (GWC2008)*, 453–461.

Veale, T., & Hao, Y. (2009). Support structures for linguistic creativity: a computational analysis of creative irony in similes. In *Proceedings of CogSci 2009, the 31st annual meeting of the cognitive science society*, 1376–1381.

Vinodhini, G., & Chandrasekaran, R. M. (2013). Effect of Feature Reduction in Sentiment Analysis of Online Reviews. *International Journal of Advanced Research in Computer Engineering & Technology 2*(6), 2165–2172.

Vitas, D. (1980). Generisanje imeničkih oblika u srpskohrvatskom jeziku. *Informatica 3*, 34–39. Ljubljana.

Vitas, D., & Krstev, C. (2009). Srpski jezik i SNTPI. U Đuro Kutlača (ur.), *SNTPI'09 - Naučno-stručni skup Sistem naučnih, tehnoloških i poslovnih informacija*, (pp 87–90). Fakultet informacionih tehnologija, Beograd.

Vitas, D., & Krstev, C. (2013). Derivational Morphology in E-Dictionaries of Serbian. In *Proceedings of the 32nd International Conference on Lexis and Grammar*, 177–184.

Vitas, D., Krstev, C., Obradović, I., Popović, Lj., & Pavlović-Lažetić, G. (2003). A Processing Serbian Written Texts: An Overview of Resources and Basic Tools. In S. Piperidis and V. Karkaletsis (Eds), *Workshop on Balkan Language Resources and Tools*, 97–104. Thessaloniki, Greece.

Vitas, D., Krstev, C., Pavlović-Lažetić, G., & Nenadić, G. (2000). Recent Results in Serbian Computational Lexicography. In Neda Bokan (Ed.), *Proceedings of the Symposium "Contemporary Mathematics"*, 113–130. Faculty of Mathematics, Belgrade.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*.

Vossen, P. (1998a). Introduction to EuroWordNet. *Computers and the Humanities 32*(2-3), 73–89.

Vossen, P. (ed.). (1998b). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2012). A Framework for Sentiment Anaysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish. *ISCIS*, 437–445.

Wallace, C. B. (2015). Computational Irony: A Survey and New Perspectives. *Artificial Intelligence Review, 43*(4), 467–483.

Wang, B., & Wang, H. (2008). Bootstrapping Both Product Features and Opinion Words from Chinese Customer Reviews with Cross-Inducing. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, *1*, 289–295.

Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems 57*, 77–93.

Wang, S., Li, D., Wei, Y., & Li, H. (2009). A Feature Selection Method Based on Fisher's Discriminant Ratio for Text Sentiment Classification. *Web Information Systems and Mining*, *5854*, 88–97. LNCS, Springer.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the 17th Conference of the AAAI*.

Wiebe, J., Bruce, R., & O'Hara, T. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 246–253.

Wiebe, J., Wilson, T., & Bell, M. (2001). Identifying Collocations for Recognizing Opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, 24–31.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Computational Linguistics 30*(3), 277–308. MIT Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354. ACL.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis. *Computational Linguistics, 35*(3), 399–433.

Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, *22*, 73–99.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval*, *1*(1-2), 69–90. Kluwer Academic Publishers, Hingham, MA, USA.

Yano, T., & Smith, A. N. (2010). What's Worthy of Comment? Content and Comment Volume in Political Blogs. In *Proceedings of ICWSM 2010*. The AAAI Press.

Yessenalina, A., & Cardie, C. (2011). Compositional Matrix-Space Models for Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 172–182.

Yessenov, K., & Misailovic, S. (2009). Sentiment analysis of movie review comments. *6.863 Spring 2009 final project*. CSAIL, MIT.

Yi, J., Nasukawa, T., Bunescu, R., & Wayne, N. (2003). Sentiment Analyzer: Extracting Sentiments About a Given Topic Using Natural Language Processing Techniques. In *Proceedings of the Third IEEE International Conference on Data Mining*, 427–434.

Yousefpour, A., Ibrahim, R., Nuzly, H., & Hamed, A. (2014). A Novel Feature Reduction Method in Sentiment Analysis. *International Journal of Innovative Computing 4*(1), 34–40.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 129–136.

Zhai, C. (2009). *Statistical Language Models for Information Retrieval*. Morgan & Claypool Publishers.

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2010). Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints. In *Proceedings of the 23rd International Conference on Computational Linguistics COLING '10*, 1272–1280. ACL.

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Clustering Product Features for Opinion Mining. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 347–354.

Zhang, Q., Huang, X., & Wu, L. (2008). Sentiment Classification Using Supervised and Semi-Supervised Conditional Maximum Entropy Modeling. *International Journal of Computer Processing Oriental Languages, 21*(4), 295–308.

Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find Product Weakness from Chinese Reviews by Using Aspects Based Sentiment Analysis. *Expert Systems with Applications 39*(11), 10283–10291.

Zhang, Y., & Zhu, W. (2013). Extracting Implicit Features in Online Customer Reviews for Opinion Mining. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, 103–104.

Zhang, Z., & Li, X. (2010). Controversy is Marketing: Mining Sentiments in Social Media. *HICSS,* 1–10. IEEE Computer Society.

Zhao, J., Dong, L., Wu, J., & Xu, K. (2012). MoodLens: An Emoticon-based Sentiment Analysis System for Chinese Tweets. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1528–1531. ACM.

Zhao, P., Li, X., & Wang, K. (2013). Feature Extraction from Micro-blogs for Comparison of Products and Services. *Web Information Systems Engineering, 8180*, 82–91. LNCS, Springer.

Zou, F., Wang, F. L., Deng, X., Han, S., & Wang, S. L. (2006). Automatic Construction of Chinese Stop Word List. In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science ACOS'06*, 1009–1014.

ПРИЛОЗИ

Прилог 3.1

Табела лексикографских датотека PWN и синтаксних категорија којима оне припадају

Бр. дат.	Назив лексик. датотеке	Опис лексикографске датотеке	Кат *
00	adj.all	сви скупови (кластери) придева	3
01	adj.pert	релациони придеви (pertainyms)	3
02	adv.all	сви прилози	4
03	noun.Tops	именице које означавају опште појмове	1
04	noun.act	именице које означавају радње и дешавања	1
05	noun.animal	именице које означавају животиње	1
06	noun.artifact	именице које означавају објекте створене људском руком	1
07	noun.attribute	именице које означавају атрибуте људи и ствари	1
08	noun.body	именице које означавају делове тела живих бића	1
09	noun.cognition	именице које означавају когнитивне процесе и садржаје	1
10	noun.communication	именице које означавају комуникационе процесе и садржаје	1
11	noun.event	именице које означавају догађаје у природи	1
12	noun.feeling	именице које означавају осећања и расположења	1
13	noun.food	именице које означавају храну и пиће	1
14	noun.group	именице које означавају имена група људи и објеката	1
15	noun.location	именице које означавају појмове просторног позиционирања	1
16	noun.motive	именице које означавају мотиве, циљеве, сврху	1
17	noun.object	именице које означавају објекте које није створио човек (природне)	1
18	noun.person	именице које се односе на људе	1
19	noun.phenomenon	именице које означавају природне феномене и појаве	1
20	noun.plant	именице које означавају биљке	1
21	noun.possession	именице које се односе на поседовање и пренос власништва	1
22	noun.process	именице које означавају природне процесе	1
23	noun.quantity	именице које се односе на величине и мерне јединице	1
24	noun.relation	именице које означавају односе међу људима, стварима и идејама	1
25	noun.shape	именице које означавају дво или вишедимензионалне облике	1
26	noun.state	именице које означавају трајне облике ствари	1
27	noun.substance	именице које означавају супстанце	1
28	noun.time	именице које означавају време и временске релације	1
29	verb.body	глаголи који се односе на дотеривање, одржавање тела и облачење	2
30	verb.change	глаголи који се односе на промену величине и интензивирање	2
31	verb.cognition	глаголи који се односе на мишљење, закључивање, анализу и сумњу	2
32	verb.communication	глаголи који означавају говор, питања, наређења, певања	2
33	verb.competition	глаголи који се односе на борбе и спортске активности	2
34	verb.consumption	глаголи који се односе на конзумирање пића и хране	2
35	verb.contact	глаголи који се односе на физичке контакте између живих бића	2
36	verb.creation	глаголи који се односе на креативне активности	2
37	verb.emotion	глаголи који се односе на исказивање осећања	2
38	verb.motion	глаголи који се односе на кретање	2
39	verb.perception	глаголи који се односе на радње које се откривају чулима	2
40	verb.possession	глаголи који се односе на процесе поседовања и промене власништва	2
41	verb.social	глаголи који се односе на друштв. и политичке активности и догађаје	2
42	verb.stative	глаголи који се односе на процесе постојања и простор.позиционир.	2
43	verb.weather	глаголи који се односе на метеоролошке прилике и појаве	2
44	adj.ppl	придеви изведени из партиципа	3

* Синтаксна категорија – POS (физичка датотека у којој се лексикографска датотека налази)
1- именице 2- глаголи 3 – придеви 4 - прилози

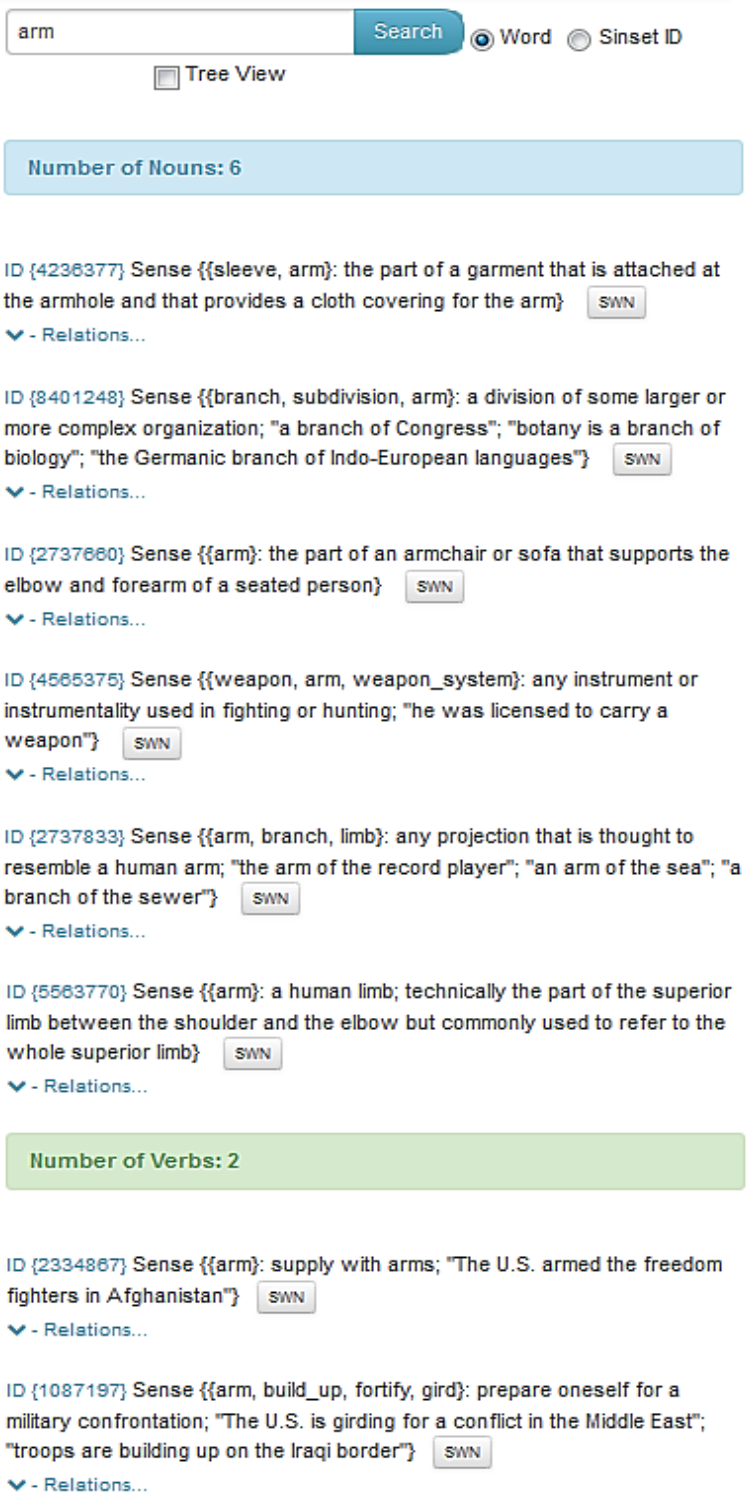
Прилог 3.2

Табела симбола показивача (*pointer_symbols*) у PWN.

Симбол показивача	Релација	Врста релације	Синтаксна категорија
!	Antonym	лексичка	**n, v, a, r**
@	Hypernym	концептуална	**n, v**
@i	Instance Hypernym	концептуална	**n**
	Hyponym	концептуална	**n, v**
i	Instance Hyponym	концептуална	**n**
#m	Member holonym	концептуална	**n**
#s	Substance holonym	концептуална	**n**
#p	Part holonym	концептуална	**n**
%m	Member meronym	концептуална	**n**
%s	Substance meronym	концептуална	**n**
%p	Part meronym	концептуална	**n**
=	Attribute	концептуална	**n, a**
+	Derivationally related form	лексичка	**n, v**
;c	Domain of synset - TOPIC	концептуална	**n, v, a, r**
-c	Member of this domain - TOPIC	концептуална	**n**
;r	Domain of synset - REGION	концептуална	**n, v, a, r**
-r	Member of this domain - TOPIC	концептуална	**n**
;u	Domain of synset - USAGE	концептуална	**n, v, a, r**
-u	Member of this domain - USAGE	концептуална	**n**
*****	Entailment	концептуална	**v**
>	Cause	концептуална	**v**
^	Also see	лексичка	**v, a**
$	Verb Group	концептуална	**v**
&	Similar to	концептуална	**a**
<	Participle of verb	лексичка	**a**
****	Pertainym (pertains to noun)	лексичка	**a**
****	Derived from adjective	лексичка	**r**
;	Domain of synset	концептуална	**n, v, a, r**
-	Member of this domain	концептуална	**n**

Прилог 3.3

Пример претраге речи *arm* у *PWN* где се могу уочити синсетови који садрже дату реч, подељени по синтаксним категоријама



Прилог 3.4

Листе симетричних и инверзних релација у *PWN*

Симетрична релација
Antonym
Similar to
Attribute
Verb Group
Derivationally Related
Also see

Релација	Инверзна релација
Hyponym	Hypernym
Hypernym	Hyponym
Instance Hyponym	Instance Hypernym
Instance Hypernym	Instance Hyponym
Holonym	Meronym
Meronym	Holonym
Domain of	Member of Doman

Прилог 3.5

Листа описа концептуалних релација установљених у PWN

Релација	Опис релације	Синтаксна категорија у којој се примењује
хипероними	именица A је хипероним именице B ако је сваки примерак B подврста од A (*B is a A, B kind of A*)	n, v
хипоними	именица A је хипоним именице B ако је сваки примерак B надврста A (*A is a B, A kind of B*)	n, v
холоними	именица *A* је холоним именице *B* ако је сваки примерак *B* део од *A* (*B part of A*)	n
мероними	именица *A* је мероним именице *B* ако је сваки примерак *A* део од *B* (*A part of B*)	n
координирани изрази	изрази A и B су координирани изрази ако деле исти хипероним C (*A is a C and B is a C*)	n, v
тропоними	глагол *A* је тропоним глагола *B* ако је радња глагола *A* врста радње глагола *B*	v
атрибути	именица *A* је атрибут уколико се њена вредност исказује придевом *B (именица тежина чији су атрибути лак и тежак)*	n
узрочни изрази (entailment)	глагол *A* је узрокован од стране глагола *B* ако се извршавањем радње глагола *B* мора вршити и радња глагола *A*	v
зависни изрази (cause)	глагол *A* је зависан од стране глагола *B* ако се извршавањем радње глагола *B* мора вршити и радња глагола *A*	v
именично-односни изрази (pertainym)	придев *A* се односи на именицу *B* уколико се из значења именице *B* може закључити значење придева *A* (придев *A*, у том случају, нема антоним)	a
слични изрази	придев *A* сличан је придеву *B* ако имају исто значење и придев *A* нема директни антоним	a
партиципи глагола (глаголски придеви)	придев *A* означава резултат акције или процеса глагола *B*	a
корен придева	придев *A* има корен којим се гради прилог *B*	r

Прилог 3.6

Листа генеричких глаголских оквира установљених у PWN

Број оквира	Генерички облик реченичке употребе глагола
1	Something ----s
2	Somebody ----s
3	It is ----ing
4	Something is ----ing PP
5	Something ----s something Adjective/Noun
6	Something ----s Adjective/Noun
7	Somebody ----s Adjective
8	Somebody ----s something
9	Somebody ----s somebody
10	Something ----s somebody
11	Something ----s something
12	Something ----s to somebody
13	Somebody ----s on something
14	Somebody ----s somebody something
15	Somebody ----s something to somebody
16	Somebody ----s something from somebody
17	Somebody ----s somebody with something
18	Somebody ----s somebody of something
19	Somebody ----s something on somebody
20	Somebody ----s somebody PP
21	Somebody ----s something PP
22	Somebody ----s PP
23	Somebody's (body part) ----s
24	Somebody ----s somebody to INFINITIVE
25	Somebody ----s somebody INFINITIVE
26	Somebody ----s that CLAUSE
27	Somebody ----s to somebody
28	Somebody ----s to INFINITIVE
29	Somebody ----s whether INFINITIVE
30	Somebody ----s somebody into V-ing something
31	Somebody ----s something with something
32	Somebody ----s INFINITIVE
33	Somebody ----s VERB-ing
34	It ----s that CLAUSE
35	Something ----s INFINITIVE

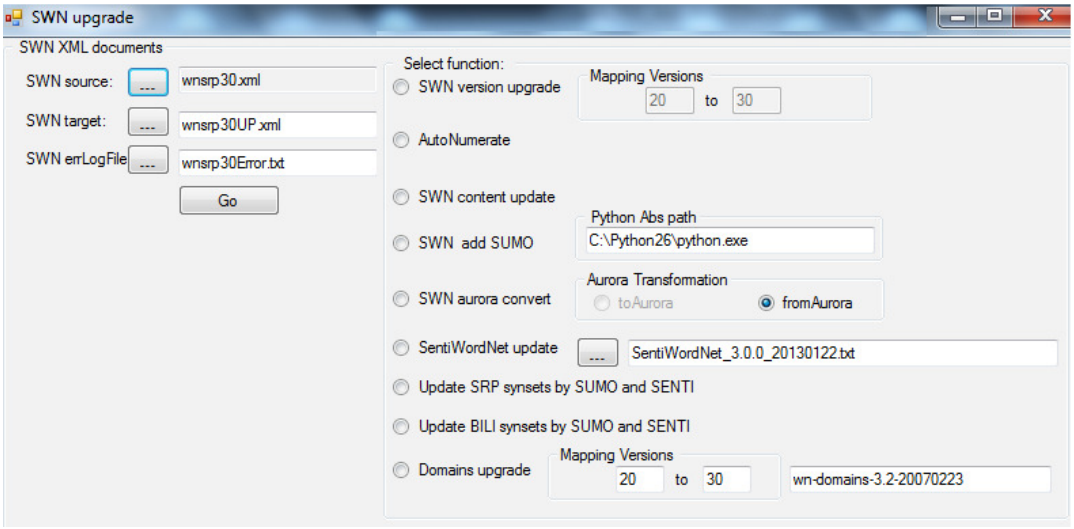
Прилог 3.7

XSD схема Српског ворднет XML документа

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified" attributeFormDefault="unqualified">
    <xs:element name="SRPWN" type="ROOT">
        <xs:unique name="ID">
            <xs:selector xpath="root"/>
            <xs:field xpath="root/SYNSET/ID"/>
        </xs:unique>
    </xs:element>
    <xs:complexType name="ROOT" mixed="true">
        <xs:sequence>
            <xs:element name="SYNSET" type="synsetType" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="synsetType">
        <xs:sequence>
            <xs:element name="ID" type="xs:string"/>
            <xs:element name="POS" type="posType"/>
            <xs:element name="SYNONYM" type="synonymType"/>
            <xs:element name="DEF" type="xs:string" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element name="BCS" type="bcsType" minOccurs="0" maxOccurs="1"/>
            <xs:element name="ILR" type="ilrType" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element name="NL" type="xs:string" minOccurs="0" maxOccurs="1"/>
            <xs:element name="USAGE" type="xs:string" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element name="SNOTE" type="xs:string" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element name="STAMP" type="xs:string" minOccurs="0" maxOccurs="1"/>
            <xs:element name="SUMO" type="sumoType" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element name="SENTIMENT" type="sentimentType" minOccurs="0" maxOccurs="1"/>
            <xs:element name="DOMAIN" type="xs:string" minOccurs="0" maxOccurs="1"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="synonymType">
        <xs:sequence>
            <xs:element name="LITERAL" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType mixed="true">
                    <xs:sequence>
                        <xs:element name="SENSE" type="xs:string" nillable="true" minOccurs="0" maxOccurs="1"/>
                        <xs:element name="LNOTE" type="xs:string" minOccurs="0" maxOccurs="1"/>
                    </xs:sequence>
                </xs:complexType>
            </xs:element>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="posType">
        <xs:restriction base="xs:string">
            <xs:enumeration value="a"/>
            <xs:enumeration value="b"/>
            <xs:enumeration value="n"/>
            <xs:enumeration value="v"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:simpleType name="bcsType">
        <xs:restriction base="xs:integer">
            <xs:enumeration value="1"/>
            <xs:enumeration value="2"/>
            <xs:enumeration value="3"/>
            <xs:enumeration value="4"/>
            <xs:enumeration value="5"/>
        </xs:restriction>
    </xs:simpleType>
    <xs:complexType name="ilrType" mixed="true">
        <xs:sequence>
            <xs:element name="TYPE" type="xs:string"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="sumoType" mixed="true">
        <xs:sequence>
            <xs:element name="TYPE" type="sumotypType"/>
        </xs:sequence>
    </xs:complexType>
    <xs:simpleType name="sumotypType">
        <xs:restriction base="xs:string">
            <xs:enumeration value="="/>
            <xs:enumeration value="+"/>
            <xs:enumeration value="@"/>
            <xs:enumeration value="["/>
            <xs:enumeration value=":"/>
            <xs:enumeration value="]"/>
        </xs:restriction>
    </xs:simpleType>
  <xs:complexType name="sentimentType">
    <xs:sequence>
      <xs:element name="POSITIVE" type="xs:decimal"/>
      <xs:element name="NEGATIVE" type="xs:decimal"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

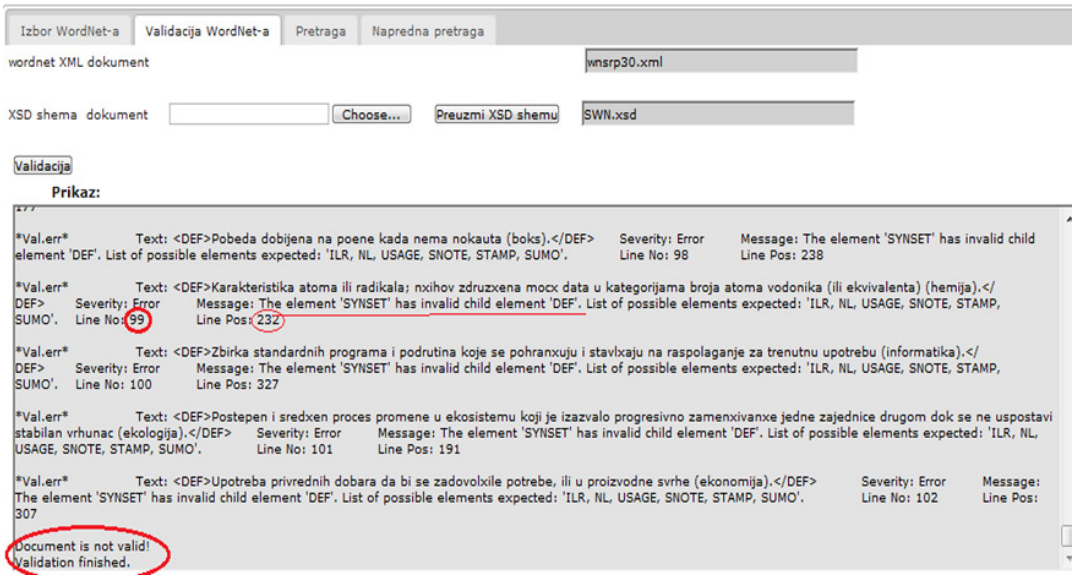
Прилог 3.8

Алат *SWNUpgrade* којим се Српски ворднет 1) подигао (upgraded) са верзије 2.0 на 3.0 и могућношћу будућих унапређења сагласно новим верзијама PWN-а; 2) конвертовао са кодног распоред *Аурора* у *Јуникод (unicode)* (може и обрнуто); 3) проширио додатним скуповима ознака и паралелизовао са онтологијом *SUMO*, лексичким ресурсом *SentiWordNet* и таксономијом *Domains*

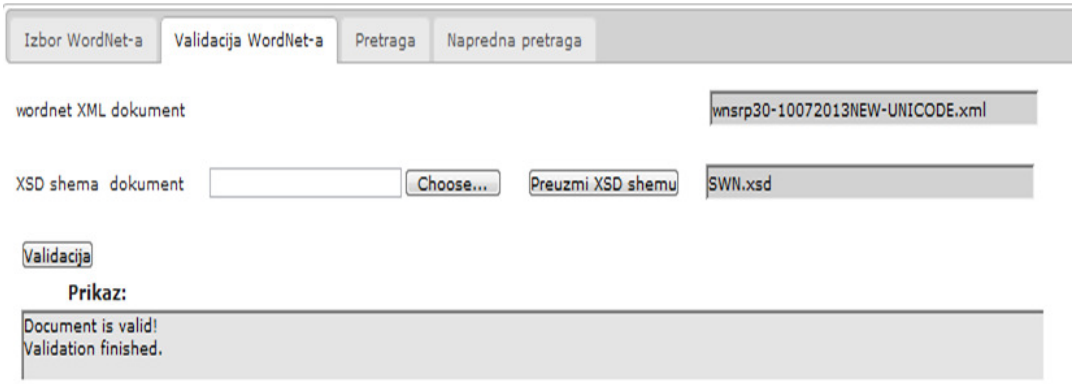


Прилог 3.9

Алат за валидацију *SWN XML* документа у односу на *XSD SWN* схему. У општем случају се може валидирати било који *XML* документ у односу на произвољну схему.



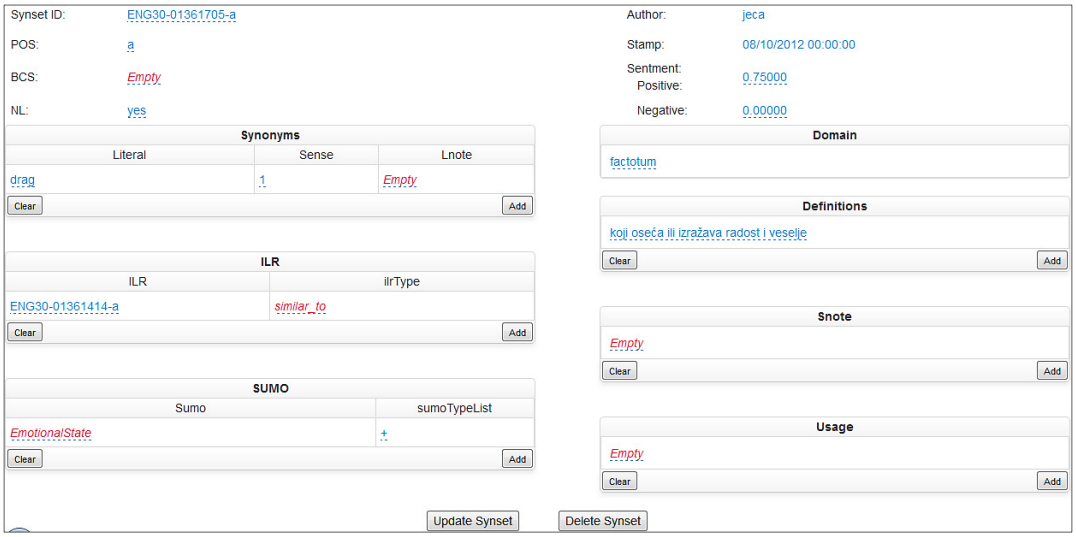
а) пример инвалидног документа са резултатима који корисника упућују на врсту и локацију грешке



б) пример валидног *SWN XML* документа

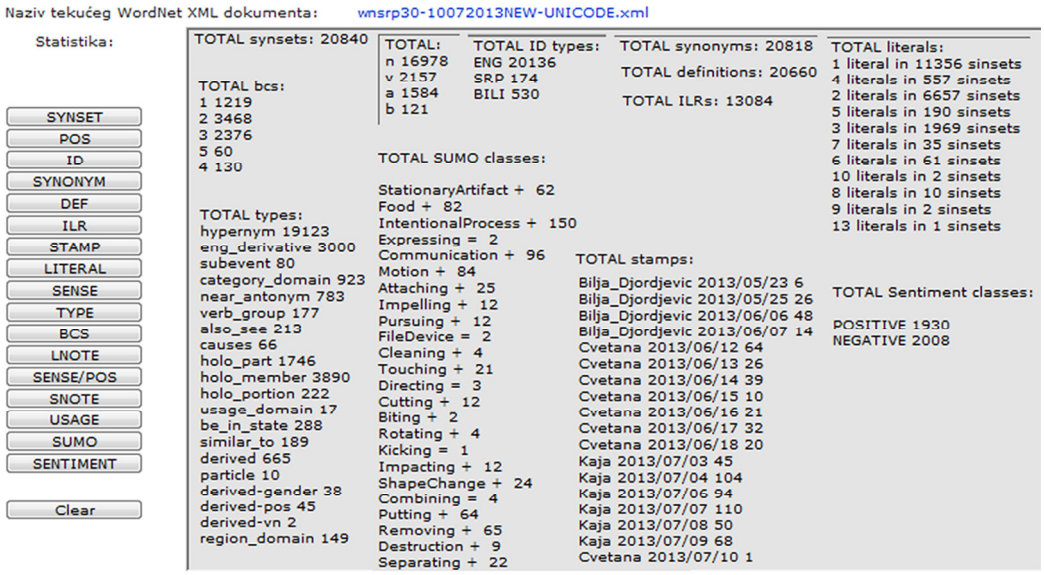
Прилог 3.10

Онлајн едитор Српског ворднета – веб апликација SWNE



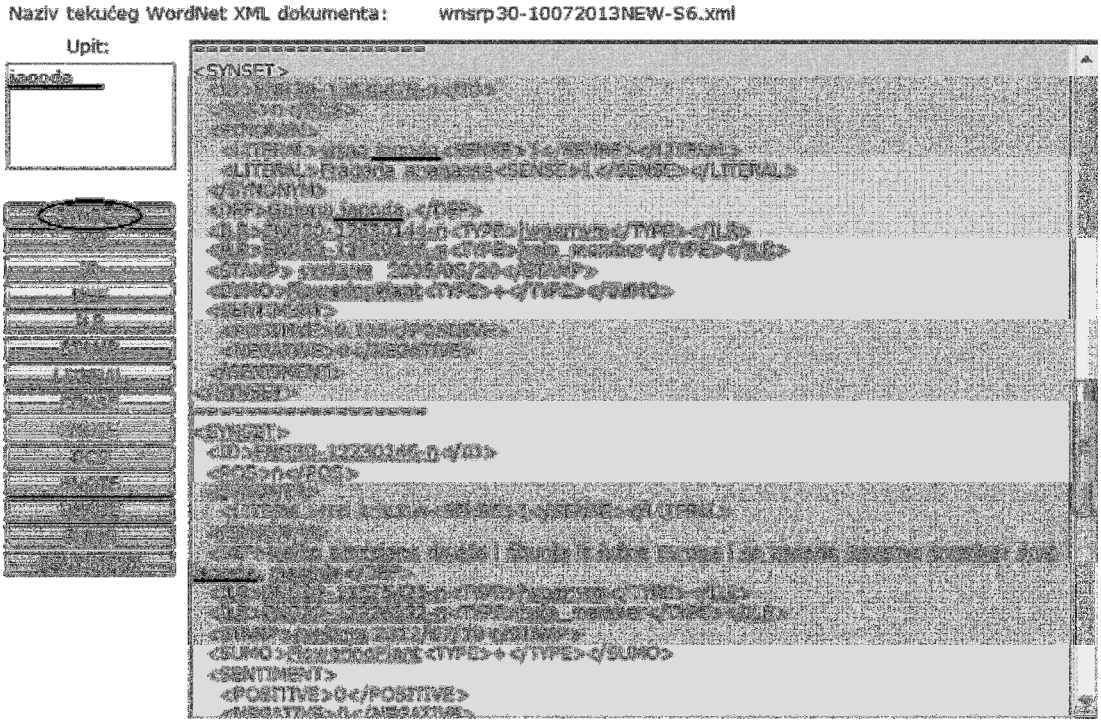
Прилог 3.11

Статистичка извештавања о структури Српског ворднета - веб апликација SWNE



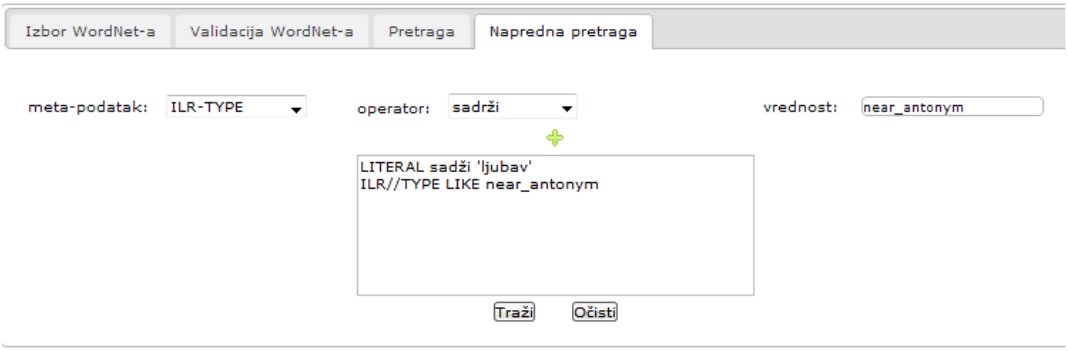
Прилог 3.12

Претрага *SWN* XML документа аутоматским генерисањем *XPATH* упита - веб апликација SWNE

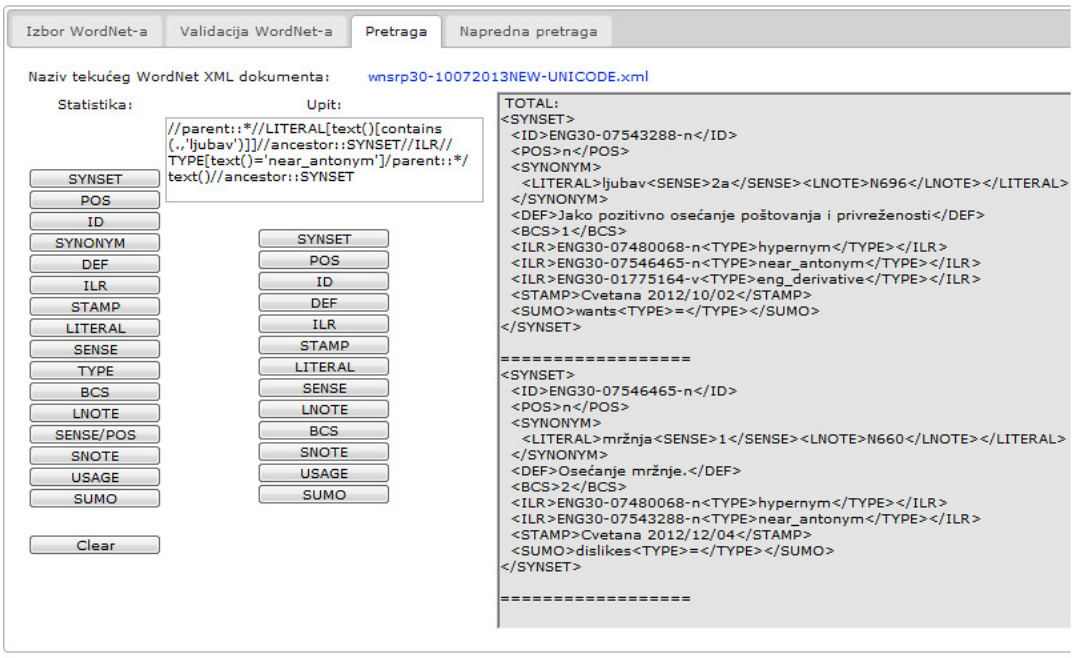


Прилог 3.13

Модул за формирање сложених логичких *XPATH* упита над *SWN* XML ресурсом постављањем упита корисника којима *XPATH* није близак, одабиром поља и вредности за филтрирање и генерисањем *XPATH* упита на основу упита корисника и приказ резултата - веб апликација SWNE



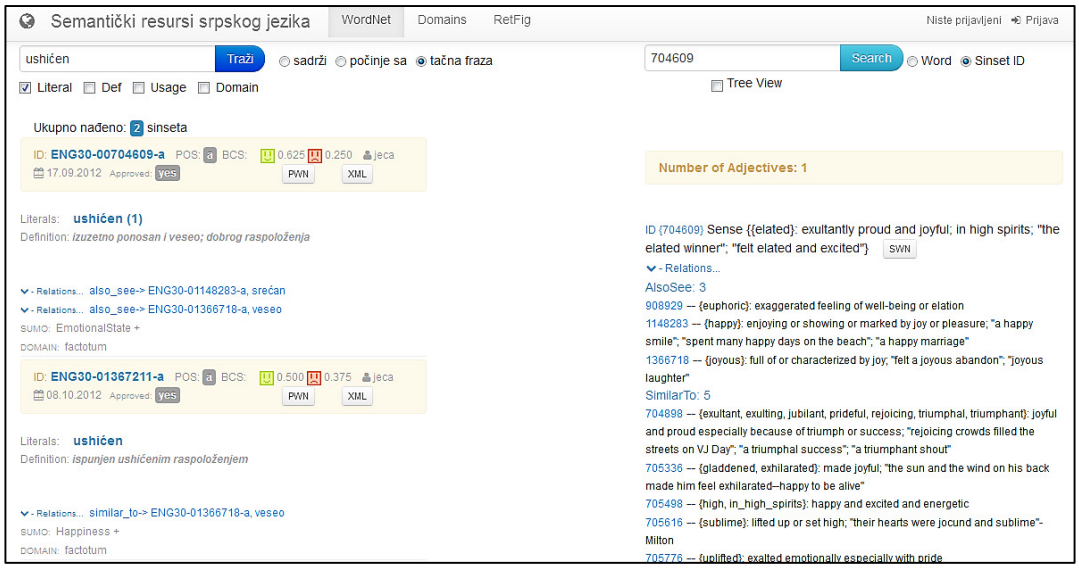
а) упит којим корисник тражи скуп синсетова који представљају антоним синсету чији бар један литерал садржи реч *ljubav*



б) резултат упита а) је *XPATH* структура којом се приказује тражени антоним одговарајућег синсета и примена тог упита над *SWN* XML (синсетови који задовољавају постављене услове приказани у десном делу веб стране)

Прилог 3.14

Модул за паралелну претрагу SWN и PWN на основу литерала, дефиниције, примера употребе или домена - веб апликација SWNE



Прилог 3.15

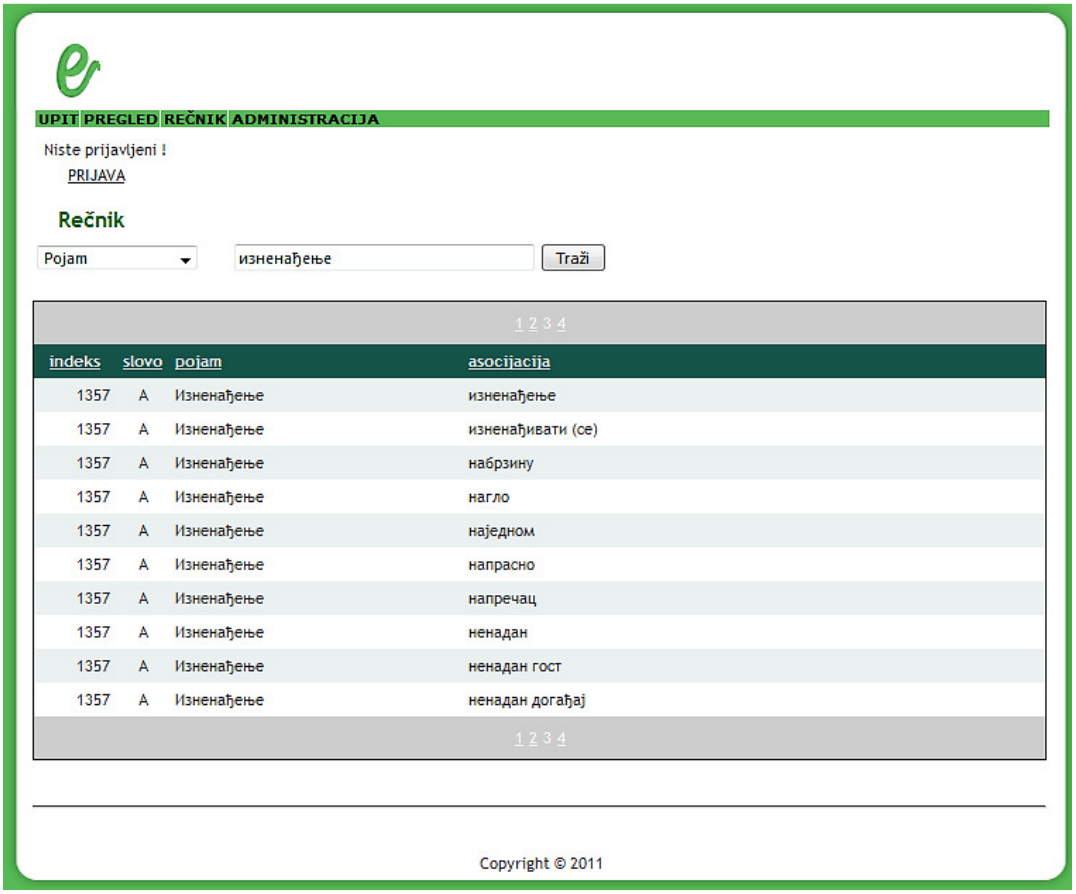
Упоредни приказ релација Принстонског ворднета, Еуроворднета и Српског ворднета

Врста релације	Релација	Врсте речи у релацији	Ознака релац. у PWN	Ознака релације у EWN	Ознака релације у SWN
Syn◄►Syn	hypernymy	N → N	@	HAS_HYPERONYM	hypernym
Syn◄►Syn	troponymy	V → V	@	HAS_HYPERONYM	hypernym
Syn◄►Syn	hyponymy	N → N	~	HAS_HYPONYM	hyponym
Syn◄►Syn	hyponymy	V → V	~	HAS_HYPONYM	hyponym
Syn◄►Syn	cross POS hypernymy	N → V N → Adj,Adv V → Adj,Adv V → N Adj,Adv → N Adj,Adv → V		HAS_XPOS_HYPERONYM	
Syn◄►Syn	meronymy	N → N	#m	HAS_MERO_MEMBER	mero_member
Syn◄►Syn	meronymy	N → N	#p	HAS_MERO_PART	mero_part
Syn◄►Syn	meronymy	N → N	#s	HAS_MERO_PORTION	mero_portion
Syn◄►Syn	meronymy	N → N		HAS_MERO_MADEOF	substanceMeronym
Syn◄►Syn	meronymy			HAS_MERO_LOCATION	
Syn◄►Syn	holonymy	N → N	%m	HAS_HOLO_MEMBER	holo_member
Syn◄►Syn	holonymy	N → N	%p	HAS_HOLO_PART	holo_part
Syn◄►Syn	holonymy	N → N	%s	HAS_HOLO_PORTION	holo_portion
Syn◄►Syn	holonymy	N → N		HAS_HOLO_MADEOF	substanceHolonym
Syn◄►Syn	holonymy	N → N		HAS_HOLO_LOCATION	
Syn◄►Syn	antonymy	N ◄► N V ◄► V		NEAR_ANTONYM	near_antonym
Syn◄►Syn	antonymy	Adj ◄► Adj			near_antonym
WS◄►WS	antonymy	N ◄► N V ◄► N Adj ◄► Adj Adv ◄► Adv	!	ANTONYM	antonym
Syn◄►Syn	domain Category	N → N V → N Adj → N Adv → N	;c		category_domain
Syn◄►Syn	domain Category Member	N → N V → N Adj → N Adv → N	-c		
Syn◄►Syn	domain Region	N → N Adj → N	;r		region_domain
Syn◄►Syn	domain Region Member	N → N Adj → N	-r		
Syn◄►Syn	domain Usage	N → N	;u		usage_domain
Syn◄►Syn	domain Usage Member	N → N	-u		
WS◄►WS	derivate	N, V, Adj, Adv (all pairs)		IS_DERIVED_FROM	
WS◄►WS	derivate	N, V, Adj, Adv (all pairs)		HAS_DERIVED	
WS◄►WS	derivate	N, V, Adj, Adv (all pairs)		DERIVATION	
WS ◄►WS	derivationally related form	N → V V → N	+		eng_derivative
WS◄►WS	derived	Adv → Adj	\		derived

WS ◄►WS	derived	N ► N			derived
WS ◄►WS	derivative	V ► N N ► V			derived-vn
WS ◄►WS	derivative	N ► N			derived-gender
WS ◄►WS	derivative	Adj ► N	\		derived-pos
WS ◄►WS	participle of verb	Adj ► V	<		particle
Syn ◄►Syn	be in state	N ► Adj,Adv V ► Adj,Adv		BE_IN_STATE	be_in_state
Syn ◄►Syn	entailment	V ► V	*		subevent
Syn ◄►Syn	entailment	V>V, N>V, N>N, V>N		IS_SUBEVENT_OF	
Syn ◄►Syn	causes	V ► V	>		causes
Syn ◄►Syn	causes	V ► V N ► V N ► N V ► N V ► Adj,Adv N ► Adj,Adv		CAUSES	
Syn ◄►Syn	verb group	V ◄► V	$		verb_group
WS ◄►WS	also see	N ► N V ► V Adj ► Adj	^		also_see
WS ◄►WS	similar	Adj ◄► Adj	&		similar_to
WS ◄►WS	attribute	N ► Adj Adj ► N	=		
WS ◄►WS	pertainym	Adj ► N	\	PERTAINS_TO	

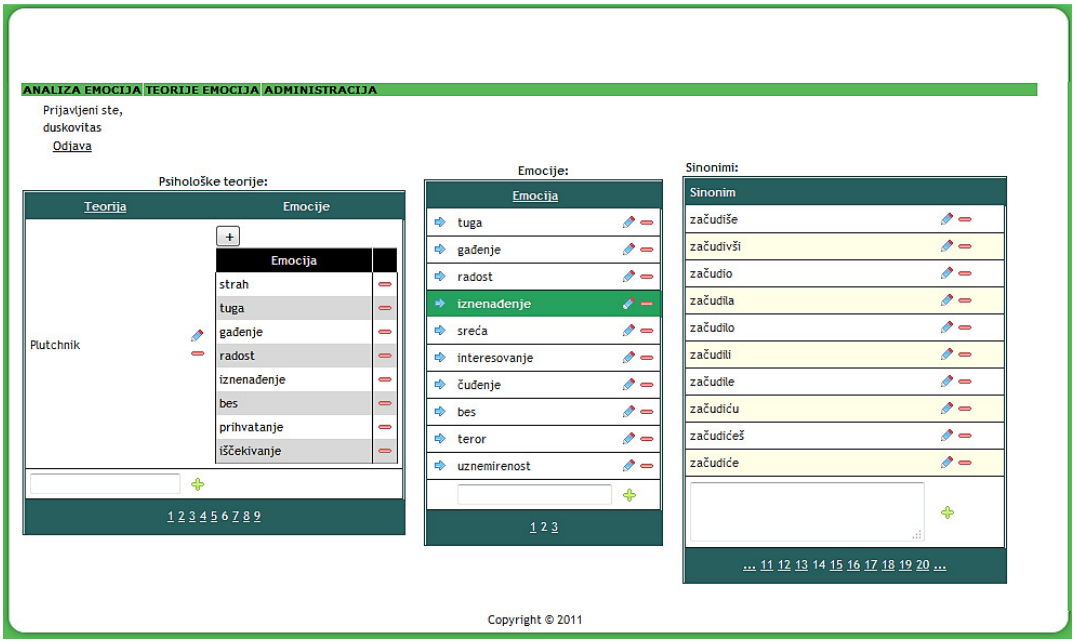
Прилог 5.1

Веб апликација за развој и претрагу асоцијативних појмова којима се може проширити лексикон сентименталних речи и израза, заснован на Систематском речнику српског језика у издању Матице српске



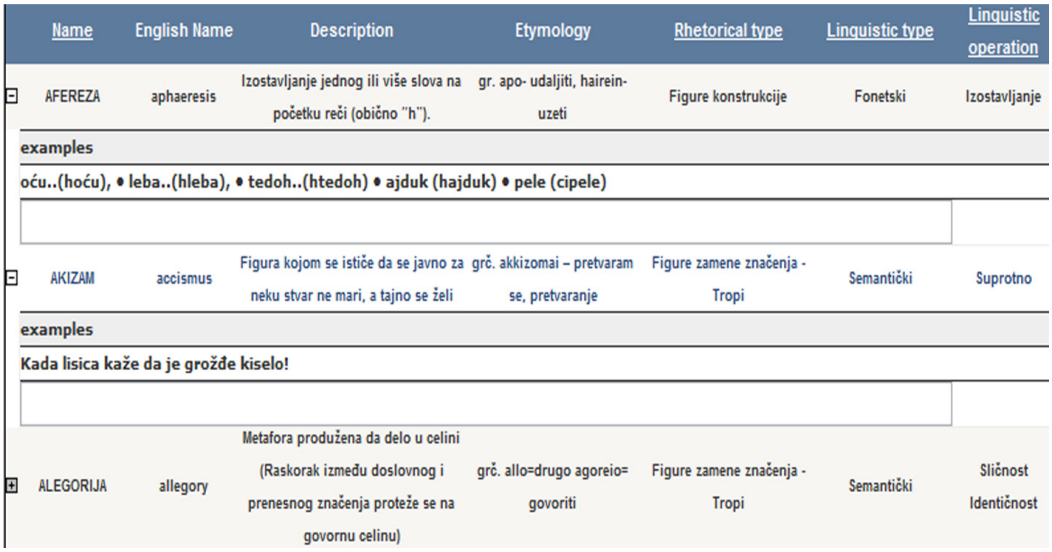
Прилог 5.2

Веб апликација за развој и проширење лексикона сентименталних речи и израза



Прилог 6.1

Веб страна за развој и проширење базе података о реторичким фигурама и њиховим примерима примене у српском језику - веб апликација SWNE



Name	English Name	Description	Etymology	Rhetorical type	Linquistic type	Linquistic operation
AFEREZA	aphaeresis	Izostavljanje jednog ili više slova na početku reči (obično "h").	gr. apo- udaljiti, hairein- uzeti	Figure konstrukcije	Fonetski	Izostavljanje
examples						
oću..(hoću), • leba..(hleba), • tedoh..(htedoh) • ajduk (hajduk) • pele (cipele)						
AKIZAM	accismus	Figura kojom se ističe da se javno za neku stvar ne mari, a tajno se želi	grč. akkizomai – pretvaram se, pretvaranje	Figure zamene značenja - Tropi	Semantički	Suprotno
examples						
Kada lisica kaže da je grožđe kiselo!						
ALEGORIJA	allegory	Metafora produžena da delo u celini (Raskorak između doslovnog i prenesnog značenja proteže se na govornu celinu)	grč. allo=drugo agoreio= govoriti	Figure zamene značenja - Tropi	Semantički	Sličnost Identičnost

и резултат њене серијализације у XML формат на примеру фигуре *афереза*

```
<RETFIG>
 <figure>
  <id>79684247-404e-4aee-baeb-aeafa4e5bdb3</id>
  <name>AFEREZA</name>
  <engname>aphaeresis</engname>
  <description>Izostavljanje jednog ili više slova na početku reči  (obično “h”).</description>
  <etymology>gr. apo- udaljiti, hairein- uzeti</etymology>
  <notice></notice>
  <rhettype>Figure naglašavanja</rhettype>
  <lingtype>Fonetski</lingtype>
  <lingops>Izostavljanje</lingops>
  <examples>
        <example>oću (hoću)</example>
        <example>leba (hleba</example>
        <example>tedoh (htedoh)</example>
        <example>ajduk (hajduk)</example>
        <example>pele (cipele)</example>
  </examples>
 </figure>
 ...
</RETFIG>
```

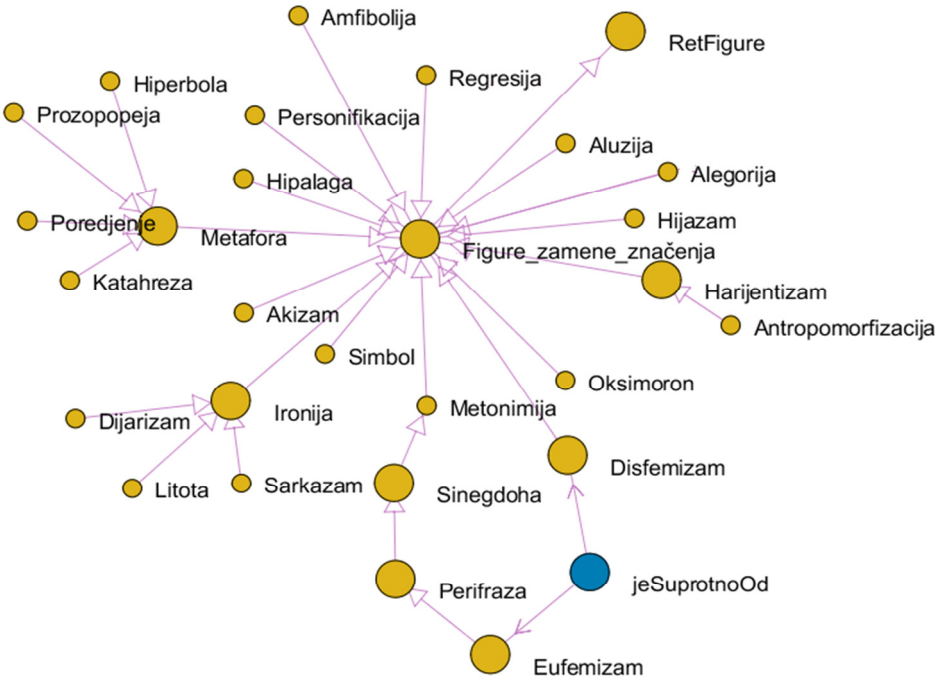
Прилог 6.2

Декларација класе *RetorickaFigura* у онтологији *RetFiguresOnto.owl*

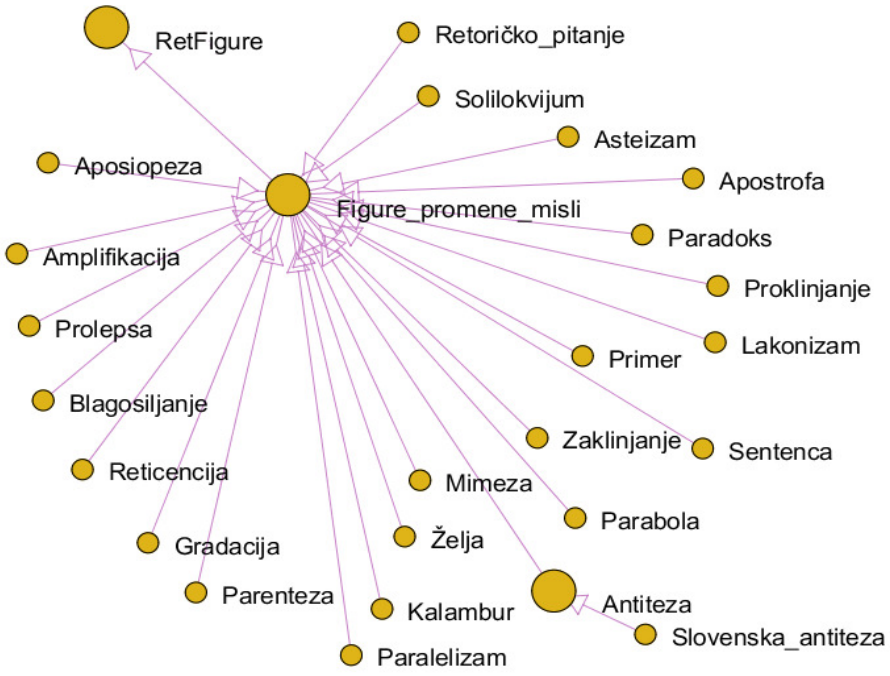
```
Declaration(Class(ont:RetorickaFigura))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:jeNaPoziciji ont:LingvistickaPozicija))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:jeNadObjektom ont:LingvistickiObjekat))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:jeNadOpsegom ont:LingvistickiOpseg))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:jeSimetricno ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura ObjectAllValuesFrom(ont:seDodaje
ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:seIzostavlja ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:sePonavlja ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:sePremesta ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:seRazdvaja ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura ObjectAllValuesFrom(ont:seSpaja
ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura
ObjectAllValuesFrom(ont:seZamenjuje ont:LingvistickiElement))

EquivalentClasses(ont:RetorickaFigura ObjectMinCardinality(1
ont:jeLingvistickaGrupa ont:LingvistickaGrupa))

EquivalentClasses(ont:RetorickaFigura ObjectExactCardinality(1
ont:jeRetorickaGrupa ont:RetorickaGrupa))

SubClassOf(ont:RetorickaFigura ont:LingvistickiEntitet)
SubClassOf(ont:RetorickaFigura ont:RetorickiEntitet)
```

Прилог 6.3

Инстанце класе *RetorickaFigura* које се односе на лингвистичку групу *FigureZameneZnacenja-tropi*

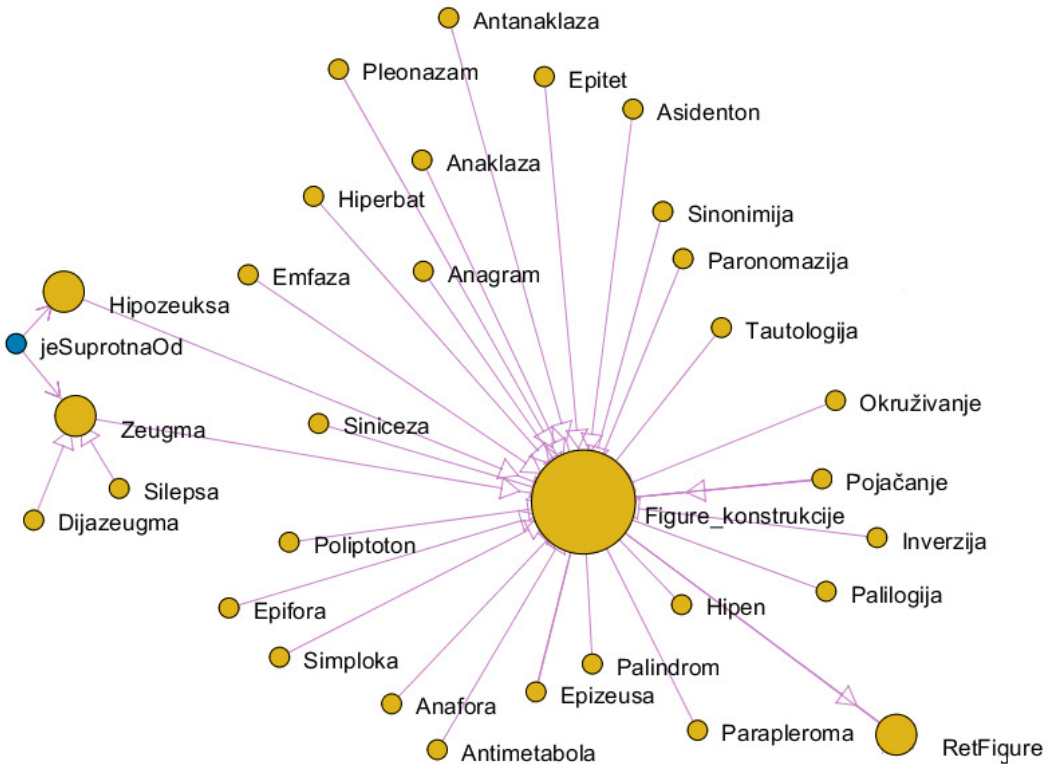


Инстанце класе *RetorickaFigura* које се односе на лингвистичку групу *FigurePromeneMisli*

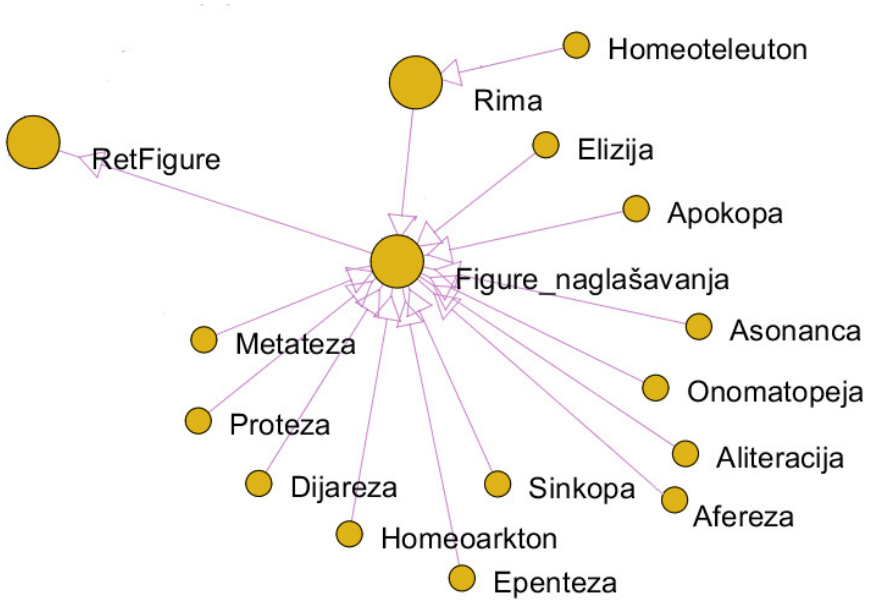


Прилог 6.4

Инстанце класе *RetorickaFigura* које се односе на лингвистичку групу *FigureKonstrukcije*



Инстанце класе *RetorickaFigura* које се односе на лингвистичку групу *FigureNaglasavanja*



Прилог 6.5

Веб алат креиран тако да се могу упаривати литерали свих врста речи у SWN помоћу неке од релација - веб апликација SWNE

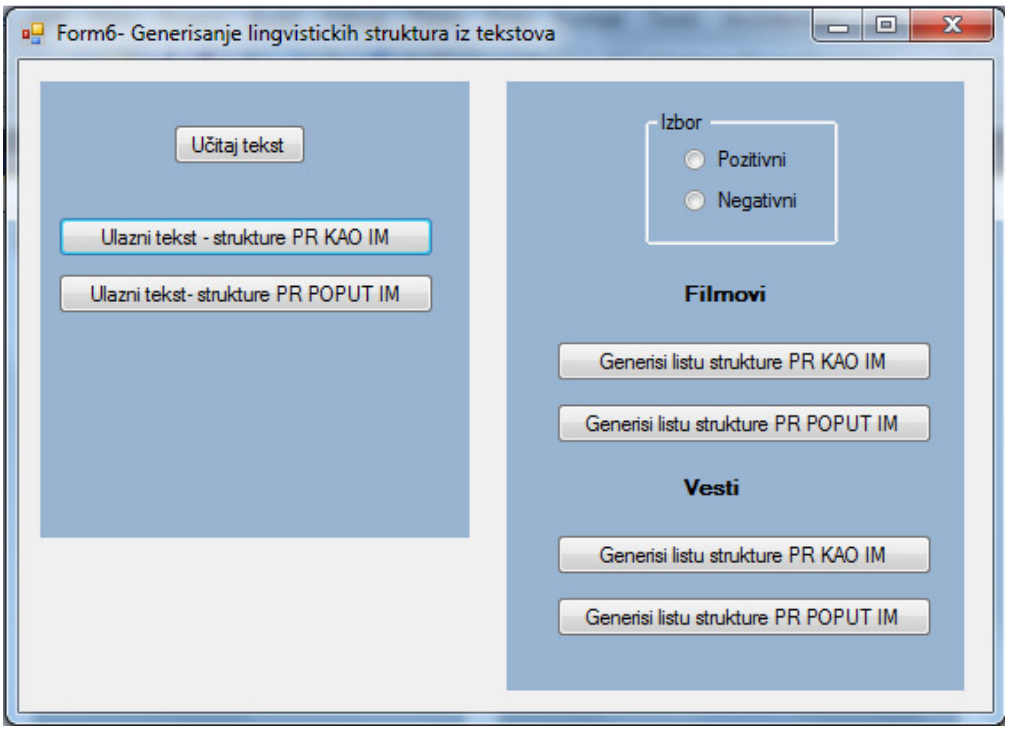
Прилог 6.6

Веб алат за оцену сагласности резултата испитаника неке анкете Крипендорфовим α-коефицијентом, на адреси

http://sm.jerteh.rs/Alati.aspx

Прилог 6.7

Алат за екстракцију лингвистичких структура кандидата за инстанцирање класа *Poredjenje* и *Ironija* у онтологији *SemRetFig*



Прилог 6.8

Резултати проналажења фигуре поређења у роману „Витез седам краљевстава“ Џорџа Мартина онтологијом *SemRetFig*

Структура откривена регуларним изразом	Структура откривена онтологијом
beo kao kost	1
beo kao kreč	1
beo kao mesečina	
bled kao kost	1
crn kao noć	1
crn kao ugalj	1
crven kao ribizla	1
crven kao bulka	1
crven kao krv	1
crven kao nar	1
crven kao oganj	1
crven kao rana	
čvrst kao štit	
glup kao tocilo	
hladan kao kamen	1
jarko kao vatra	1
lagan kao rosa	
lagan kao vazduh	1
mršav kao komarac	
mršav kao koplje	1
mudar kao meštar	
prav kao bodež	
prav kao koplje	1
prokažen kao buntovnik	
prokažen kao izdajnik	
sjajan kao dragulj	
sjajan kao zlato	1
snažan kao bivo	
spor kao bivo	
suv kao dren	
taman kao lešnik	
tup kao buzdovan	
tup kao tocilo	
tup kao topuz	
tvrd kao bobica	
tvrd kao cigla	1
tvrd kao drvo	1
tvrd kao kamen	1
ukočen kao drvo	1
umoran kao pas	1
zelen kao trava	1
žilav kao koren	
Укупно : 42	Откривено:22

Биографија аутора

Миљана Младеновић (рођ. Јосифов) рођена је 11. 12. 1963. године у Врању. Основну школу „Вук Караџић“ и Гимназију „Бора Станковић“ у Врању завршила је 1982. као ђак генерације. Школске 1982/1983. године уписала је студије на Електронском факултету у Нишу (Одсек за рачунарску технику и информатику) и дипломирала 1986/1987. године са просечном оценом 9,02. Школске 2009/2010. године уписала је докторске студије на Математичком факултету, студијски програм Информатика.

Радно искуство стицала је у компанијама: Симпо холдинг компанија у Врању на пословима програмер – Приправник од децембра 1987. до новембра 1988, YUMCO НК у Врању на пословима програмер – аналитичар од новембра 1988. до јануара 1992., Дуванска индустрија Врање на пословима Computer programmer од јануара 1992. до септембра 1992., Симпо НК у Врању, на пословима Шеф рачунског центра од септембра 1992. до новембра 1998. Од новембра 1998. до данас је главни менаџер Агенције "Milenijum III", Врање. У истом периоду радила је и у Техничкој школи у Врању као професор рачунарства и информатике, у периоду од новембра 1998. до септембра 2000. Такође и у организацијама "UNDP", Врање као Office Applications Trainer - од априла 2004. до новембра 2004. Development Alternatives Inc. USAID као Office Applications Trainer – од октобра 2005. до фебруара 2006. и у компанији Guidance AD Beograd као Microsoft Certified Trainer for MCPD – од јуна 2007. до фебруара 2011. Додатна усавршавања: Microsoft Certified Professional Developer MCPD for Windows, Web and Enterprise applications, Achievement Date 2006. Certification Number: C775-9095, Microsoft Certified Trainer, Achievement Date 2008. Certification Number: D242-9741, Microsoft Certified Technology Specialist, Achievement Date 2009. Certification Number: C775-9102.

Основне области интересовања су јој обрада природног језика, класификација текста, класификација текста према осећањима, семантички веб, XML технологије, машинско учење и системи засновани на знању. Коаутор је дигиталног речника говора југа Србије у 2013. години уз подршку Регистра националног Интернет домена Србије у оквиру пројекта 4ПИ.

Прилог 1.

Изјава о ауторству

Потписани-а Милана Младеновић

број индекса 2023/09

Изјављујем

да је докторска дисертација под насловом

Информатички модели у анализи осећања засновани на језичким ресурсима

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 22.02.2016.

М Младеновић

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Милоана Младеновић

Број индекса 2023/09

Студијски програм Информатика

Наслов рада Информатички модели у анализи осећања заснованих на језичким ресурсима

Ментор проф. др Душко Витас

Потписани/а Милоана Младеновић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 22.02.2016.

М Младеновић

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Информатички модели у анализи осећања заснованих на језичким ресурсима

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 22. 02. 2016.

М Младенић