Računalni postupci za modeliranje i analizu medijske agende temeljeni na strojnome učenju

Korenčić, Damir

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Damir Korenčić

RAČUNALNI POSTUPCI ZA MODELIRANJE I ANALIZU MEDIJSKE AGENDE TEMELJENI NA STROJNOME UČENJU

DOKTORSKI RAD

Zagreb, 2019.

Sveučilište u Zagrebu
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Damir Korenčić

RAČUNALNI POSTUPCI ZA MODELIRANJE I ANALIZU MEDIJSKE AGENDE TEMELJENI NA STROJNOME UČENJU

DOKTORSKI RAD

Mentori: Izv. prof. dr. sc. Jan Šnajder
Dr. sc. Strahil Ristov

Zagreb, 2019.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Damir Korenčić

COMPUTATIONAL METHODS FOR MODELLING AND ANALYSIS OF THE MEDIA AGENDA BASED ON MACHINE LEARNING

DOCTORAL THESIS

Supervisors: Associate Professor Jan Šnajder, PhD
Strahil Ristov, PhD

Zagreb, 2019

Doktorski rad izrađen je na Institutu Ruđer Bošković, na Zavodu za elektroniku, te na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave, u Laboratoriju za analizu teksta i inženjerstvo znanja (TakeLab).

Mentori: Izv. prof. dr. sc. Jan Šnajder
Dr. sc. Strahil Ristov

Doktorski rad ima: 216 stranica

Doktorski rad br.: _____

O mentorima

Jan Šnajder

Jan Šnajder diplomirao je, magistrirao i doktorirao u polju računarstva na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva (FER), 2002., 2006. odnosno 2010. godine. Od 2002. godine radio je kao znanstveni novak, od 2011. godine kao docent, a od 2016. godine kao izvanredni profesor na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave FER- a. Usavršavao se na Institutu za računalnu lingvistiku Sveučilišta u Heidelbergu, Institutu za obradu prirodnog jezika Sveučilišta u Stuttgartu, Nacionalnome institutu za informacijske i komunikacijske tehnologije u Kyotu te Sveučilištu u Melbourneu. Sudjelovao je na nizu znanstvenih i stručnih projekata iz područja obrade prirodnog jezika i strojnog učenja. Voditelj je uspostavnog projekta HRZZ-a i projekta provjere koncepta HAMAG-BICRO-a te je istraživač na projektu UKF-a. Autor je ili suautor više od 100 znanstvenih radova u časopisima i zbornicima međunarodnih konferencija u području obrade prirodnog jezika i pretraživanja informacija te je bio recenzentom za veći broj časopisa i konferencija iz tog područja. Nositelj je šest predmeta na FER-u te je bio mentorom ili sumentorom studentima na više od 100 preddiplomskih i diplomskih radova. Član je stručnih udruga IEEE, ACM, ACL, tajnik Hrvatskoga društva za jezične tehnologije te suosnivač i tajnik posebne interesne skupine za obradu prirodnog jezika za slavenske jezike pri udruzi za računalnu lingvistiku (ACL SIGSLAV). Član je Znanstvenog centra izvrsnosti za znanost o podacima i kooperativne sustave te je pridruženi urednik časopisa Journal of Computing and Information Technology (CIT). Dobitnik je Srebrne plakete "Josip Lončar" 2010. godine, stipendije Hrvatske zaklade za znanost 2012. godine, stipendije Japanskog društva za promicanje znanosti 2014. godine te stipendije australske vlade Endeavour 2015. godine.

Strahil Ristov

Strahil Ristov rođen je u Zagrebu 1959. godine. Diplomirao je 1986. godine u polju ekektrotehnike, a magistrirao i doktorirao u polju računarstva na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva (FER), 1991. odnosno 1997. godine. Od 1990. godine radi na Zavodu za elektroniku Instituta Ruđer Bošković. Bio je gostujući istraživač u kratkom boravku na Sveučilištu Marne-la-Valle, Francuska, u 1998. i 2001. godini. 2009. godine izabran je u zvanje višeg znanstvenog suradnika. Sudjelovao je na šest znanstvenih projekata Ministarstva znanosti, obrazovanja i sporta Republike Hrvatske. Trenutno je voditelj istraživačkog projekta: "Napredni deterministički i hibridni algoritmi na nizovima, sljedovima i stablima s primjenama u tehničkim znanostima i znanostima o životu" koji financira Hrvatska zaklada za znanost. Objavio je 25 znanstvenih radova u časopisima i zbornicima konferencija u podru-

čju algoritama na nizovima, kompresije podataka, prepoznavanja uzoraka, te dubinske analize teksta. Dr.sc. Ristov sudjelovao je u međunarodnom programskom odboru znanstvene konferencije FSMNLP, te sudjeluje kao recenzent u većem broju inozemnih časopisa.

About the Supervisors

Jan Šnajder

Jan Šnajder has received his BSc, MSc, and PhD degrees in Computer Science from the University of Zagreb, Faculty of Electrical Engineering and Computing (FER), Zagreb, Croatia, in 2002, 2006, and 2010, respectively. From September 2002 he was working as a research assistant, from 2011 as Assistant Professor, and from 2016 as Associate Professor at the Department of Electronics, Microelectronics, Computer and Intelligent Systems at FER. He was a visiting researcher at the Institute for Computational Linguistics at the University of Heidelberg, the Institute for Natural Language Processing at the University of Stuttgart, the National Institute of Information and Communications Technology in Kyoto, and the University of Melbourne. He participated in a number of research and industry projects in the field of natural language processing and machine learning. He is the principal investigator on a HRZZ installation grant project and a HAMAG-BICRO proof-of-concept project, and a researcher on a UKF project. He has (co-) authored more than 100 papers in journals and conferences in natural language processing and information retrieval, and has been reviewing for major journals and conferences in the field. He is the lecturer in charge for six courses at FER and has supervised and co-supervised more than 100 BA and MA theses. He is a member of IEEE, ACM, ACL, the secretary of the Croatian Language Technologies Society, the co-founder and secretary of the Special Interest Group for Slavic NLP of the Association for Computational Linguistics (ACL SIGSLAV). He is a member of the Centre of Research Excellence for Data Science and Advanced Cooperative Systems and the associate editor of the Journal of Computing and Information Technology. He has been awarded the Silver Plaque “Josip Lončar” in 2010, the Croatian Science Foundation fellowship in 2012, the fellowship of the Japanese Society for the Promotion of Science in 2014, and the Endeavour Fellowship of the Australian Government in 2015.

Strahil Ristov

Strahil Ristov was born in Zagreb in 1959. He received B.Sc. degree in electrical engineering in 1986., and M.Sc. and Ph.D. degrees in computer science in 1991. and 1997., respectively, from the University of Zagreb, Faculty of Electrical Engineering and Computing (FER), Zagreb, Croatia. From 1990. he is working at the Department of Electronics at Ruđer Bošković Institute in Zagreb. He was a visiting researcher at the Universite Marne-la-Valle, France, for short terms in 1998. and 2001. In June 2009. he was promoted to Senior Associated Researcher. He participated in 6 scientific projects financed by the Ministry of Science, Education and Sports of the Republic of Croatia. Currently he is a project leader of the research project: "Advanced deterministic and hybrid algorithms on strings, sequences and trees with applications in techni-

cal and life sciences" financed by the Croatian Science Foundation. He published 25 papers in journals and conference proceedings in the area of string algorithms, data compression, pattern recognition, and text analysis. Dr. Ristov was a member of program committee of FSMNLP conference, and serves as a reviewer for numerous international journals.

Zahvale

Hvala Janu na vremenu i trudu uloženom u suradnju na ovom istraživanju te na pozitivnoj energiji i svoj pomoći. Hvala Staši na suradnji vezanoj uz doktorat, suradnji na algoritmima kompresije te na pomoći i strpljenju.

Hvala svim kolegama koji su obilježili moj dosadašnji istraživački rad, prije svega ekipi s Ruđerovog ZEL-a i ekipi iz TakeLaba. Hvala za sve rasprave, seminare i predavanja, interne recenzije, suradnje, razgovore i druženja, kave i pive. Hvala i za svu stručnu, birokratsku i moralnu podršku.

Hvala obitelji i prijateljima na strpljenju pri slušanju svih kritika raznih aspekata procesa doktorskog obrazovanja. Hvala Ivi na mnogim razgovorima o borbi i ostalim zaista važnim stvarima. Hvala (super) Mariu na potpori i pozitivi u nekim od težih trenutaka.

Hvala bezbrojnim inženjerima i znanstvenicima koji su stoljećima gradili znanja nužna da bi ovo istraživanje bilo izvedivo i zamislivo.

Hvala bezbrojnim programerima koji su izradili alate korištene tijekom ovog istraživanja i koji su učinili te alate slobodno dostupnim a njihov programski kod otvorenim i slobodnim.

Sažetak

Rad se bavi računalnim postupcima analize medijske agende (engl. *Media Agenda*) temeljenima na tematskim modelima (engl. *Topic Models*) te metodama vrednovanja tematskih modela. Analiza medijske agende provodi se radi stjecanja uvida u strukturu i zastupljenost medijskih tema, što je od interesa za društvenoznanstvena istraživanja te za medijsku industriju i druge komercijalne i političke aktere. Računalni postupci analize medijske agende omogućuju automatsko otkrivanje tema u velikim skupovima tekstova i mjerenje njihove zastupljenosti. Ovi postupci pružaju analitičaru uvid u teme prisutne u medijima i uvid u zastupljenost tema u pojedinim medijima i vremenskim razdobljima te omogućuju analizu korelacije zastupljenosti tema sa podacima poput ljudske percepcije njihove važnosti.

Cilj istraživanja bio je razvoj računalnih postupaka za eksplorativnu analizu i mjerenje medijske agende temeljenih na tematskim modelima, klasi modela strojnog učenja pogodnih za analizu tematske strukture teksta. Istraživanje obuhvaća razvoj postupaka primjene tematskih modela na otkrivanje medijskih tema i mjerenje njihove zastupljenosti te razvoj računalnih alata za unaprijeđenje i provedbu tih postupaka. Ti alati obuhvaćaju metode vrednovanja tematskih modela te programsku potporu za implementaciju postupaka analize agende i vrednovanja modela. Primjena postupaka na analizu medijskih tekstova brzo je pokazala potrebu za razvojem novih metoda vrednovanja tematskih modela radi povećanja efikasnosti na modelima temeljenih postupaka. Iz tog je razloga poseban naglasak istraživanja bio na razvoju i analizi metoda vrednovanja tematskih modela.

Prvo je provedeno istraživanje postupaka primjene tematskih modela na analizu medijske agende. Na temelju istraživanja postojećih postupaka predložen je poboljšani postupak koji se sastoji od tri koraka: koraka otkrivanja tema, koraka definicije tema i koraka mjerenja tema. Predloženi postupak otklanja uočene nedostatke ranijih metoda: upotrebu samo jednog modela za otkrivanje tema, nemogućnost prilagodbe i definicije novih tema te izostanak kvantitativnog vrednovanja metoda mjerenja. Postupak je primijenjen u dvije analize medijske agende provedene na zbirkama američkih i hrvatskih političkih vijesti. Na temelju opažanja i podataka iz tih analiza uočena je potreba za mjerom interpretabilnosti tema modela te za metodom mjerenja pokrivenosti skupa koncepata od strane modela.

Drugi istraženi problem bio je problem mjerenja interpretabilnosti tema modela. Standardni pristup ovom problemu je mjerenje semantičke koherentosti tema, a postojeće mjere koherentnosti temelje se na računanju koherentosti skupa uz temu vezanih riječi. Ove mjere pokazale su se nepogodnima u slučaju prolaznih medijskih tema karakteriziranih semantički nepovezanim riječima. Predložena je nova klasa mjera koherentosti medijskih tema temeljenih na uz teme vezanim dokumentima. Vrednovanje niza predloženih mjera na skupovima engleskih i hrvatskih medijskih tema otkrilo je najbolju mjeru koja računa koherentnost agregacijom lokalne

povezanosti grafa dokumenata. Provedena je kvantitativna i kvalitativna usporedba razvijenih mjera dokumentne koherentosti s postojećim mjerama koherentnosti riječi koja je otkrila komplementarnost ova dva tipa mjera.

Treći istraženi problem je problem pokrivenosti tema, motiviran podacima iz primjene postupka analize medijske agende, koji su pokazali da jedan tematski model pokriva samo dio svih otkrivenih koncepata. Problem pokrivenosti nadilazi domenu medijskih tekstova i unatoč važnosti ovog problema dosadašnja istraživanja na tu temu su rudimentarana. Problem pokrivenosti razmotren je u općenitosti i definiran kao problem mjerenja poklapanja između skupa automatski naučenih tema modela i skupa referentnih tema koji sadrži od ljudi uočene koncepte. Predložena je metoda izrade skupa referentnih tema i dvije metode mjerenja pokrivenosti temeljene na računanju poklapanja tema. Predložene mjere vrednovane su na dva raznorodna skupa podataka, medijskom i biološkom, te primijenjene na analizu četiri različite klase standardnih tematskih modela.

Završni korak istraživanja postupka analize medijske agende bio je poboljšanje postupka na temelju predloženih metoda vrednovanja tematskih modela i iskustava iz primjena postupka na analizu hrvatskih i američkih medija. Glavna poboljšanja odnose se na korak eksplorativne analize odnosno otkrivanja tema i temelje se na razvijenim mjerama pokrivenosti i dokumentne koherentosti tema. Ova poboljšanja imaju za cilj brže otkrivanje većeg broja koncepata. Ostala poboljšanja odnose se na povećanje efikasnosti postupka interpretacije tema modela.

Tijekom istraživanja postupka analize medijske agende i metoda vrednovanja tematskih modela uočen je niz problema vezanih uz upotrebu, izgradnju, pohranu i dohvat tematskih modela i vezanih resursa. Ovi problemi javljaju se kod implementacije grafičkog korisničkog sučelja za provedbu postupka i kod provedbe eksperimenata vrednovanja. Rješavanju ovih problema pristupilo se sustavno i oblikovan je radni okvir za izgradnju i upravljanje resursima u tematskom modeliranju. Arhitektura okvira temelji se na četiri načela koja u kombinaciji definiraju općenitu i fleksibilnu metodu izrade programske potpore za primjenu i vrednovanje tematskih modela. Razvijeni su i grafičko korisničko sučelje za eksplorativnu analizu i potporu mjerenju zastupljenosti tema te aplikacija namijenjena izradi zbirki medijskih tekstova koja tijekom duljeg vremenskog razdoblja sakuplja tekstove iz niza web-izvora.

Ključne riječi: Medijska agenda, tematski modeli, vrednovanje tematskih modela, koherentnost tema, pokrivenost tema, mjere udaljenosti tema, izgradnja tematskih modela, nenadzirano učenje, nadzirano učenje.

Summary

Computational Methods for Modelling and Analysis of the Media Agenda Based on Machine Learning

This thesis focuses on computational methods for media agenda analysis based on topic models and methods of topic model evaluation. The goal of a media agenda analysis is gaining insights into the structure and frequency of media topics. Such analyses are of interest for social scientists studying news media, journalists, media analysts, and other commercial and political actors. Computational methods for media agenda analysis enable automatic discovery of topics in large corpora of news text and measuring of topics' frequency. Data obtained by such analyses provides insights into the type and structure of topics occurring in the media, enables the analysis of topic cooccurrence, and analysis of correlation between topics and other variables such as text metadata and human perception of topic significance.

The goal of the research presented in the thesis is development of efficient computational methods for the discovery of topics that constitute the media agenda and methods for measuring frequencies of these topics. The proposed methods are based on topic models – a class of unsupervised machine learning models widely used for exploratory analysis of topical text structure. The research encompasses the development of applications of topic models for discovery of media topics and for measuring topics' frequency, as well as development of methods for improvement and facilitation of these applications. The improvement and facilitation methods encompass methods of topic model evaluation and software tools for working with topic models. Methods of topic model evaluation can be used for selection of high-quality models and for accelerating the process of topic discovery. Namely, topic models are a useful tool, but due to the stohasticity of the model learning algorithms the quality of learned topics varies. For this reason the methods of topic model evaluation have the potential to increase the efficiency of the methods based on topic models.

Media agenda consists of a set of topics discussed in the media, and the problem of media agenda analysis consists of two sub-tasks: discovery of the topics on the agenda and measuring the frequencies of these topics. The first contribution of the thesis is a method for media agenda analysis based on topic models that builds upon previous approaches to the problem and addresses their deficiencies. Three notable deficiencies are: usage of a single topic model for topic discovery, lack of possibility to define new topics that match the analyst's interests, and the lack of precise evaluation of methods for measuring topics' frequency. In addition to addressing the identified deficiencies, the method also systematizes the previous approaches to the problem and is evaluated in two case studies of media agenda analysis. The proposed experimental method for media agenda analysis consists of three steps: topic discovery, topic

definition, and topic measuring steps.

In order to achieve better topic coverage, the discovery step is based not on a single model but on a set of topic models. The type and number of topic models used depends on available model implementations and the time available for topic annotation, while the hyperparameter defining the number of model topics depends on the desired generality of learned topics. Reasonable default settings for model construction are proposed based on the existing agenda analysis studies and an iterative procedure for tuning the number of topics is described. After the topic models are constructed, topic discovery is performed by human inspection and interpretation of the topics. Topic interpretation produces semantic topics (concepts) that are recorded in a reference table of semantic topics that serves as a record of topics and as a tool for synchronization of human annotators. After all the model topics are inspected, annotators can optionally perform the error correcting step of revising the semantic topics, as well as the step of building a taxonomy of semantic topics. Topic discovery is supported with a graphical user interface developed for topic inspection and annotation.

The step of topic definition is based on semantic topics obtained during topic discovery. The purpose of topic definition is to define new semantic topics that closely match the analyst's exact interests. The possibility of defining new semantic topics is an important difference between the proposed and the existing media agenda analysis approaches. Namely, the existing approaches base the analysis only on model-produced topics, although there is no guarantee that these topics will match the concepts of interest to the analyst. During topic definition, the analysts infers definitions of new semantic topics based on previously discovered topics and describes these topics with word lists. Discovered semantic topics that already closely match the concepts of interest are used without modification.

During the step of topic measuring the frequencies of semantic topics obtained during the discovery and definition steps are measured. Topic frequency is defined as the number of news articles in which a topic occurrs, and the measuring problem is cast as the problem of multi-label classification in which each news article is being tagged with one or more semantic topics. This formulation allows for precise quantitative evaluation of methods for measuring topic frequency. Two measuring methods are considered. The baseline is a supervised method using the method of binary relevance in combination with a linear kernel SVM model. The second method is a newly proposed weakly supervised approach, in which the measured semantic topics are first described by sets of highly discriminative words, after which a new LDA model is constructed in such a way that the topics of the model correspond to measured topics, which is achieved via prior probabilities of model topics. The method for selecting words highly discriminative for a semantic topic represents the main difference between the proposed and the previous weakly supervised approaches. This method consists of inspecting, for each measured semantic topic, closely related model topics, and selecting words highly discriminative for the topic by means

of inspecting word-related documents and assessing their correspondence with the topic.

The proposed three-step method for media agenda analysis is applied to two media agenda analyses: the analysis of mainstream US political news and the analysis of mainstream Croatian political news in the election period. The applications of the proposed method show that the topic discovery step gives a good overview of the media agenda and leads to the discovery of useful topics, and that the usage of more than one topic model leads to a more comprehensive set of topics. The two analyses also demonstrate the necessity of the proposed topic definition step – in the case of US news new sensible topics corresponding to issues are pinpointed during this step, while in the case of Croatian election-related news the analysis is based entirely on newly defined semantic topics that describe the pre- and post-election processes. Quantitative evaluation of topic frequency measuring shows that the proposed weakly supervised approach works better than the supervised SVM-based method since it achieves better or comparable performance with less labeling effort. In contrast to the supervised method, weakly supervised models have a higher recall and work well for smaller topics. Qualitative evaluation of measuring models confirms the quality of the proposed approach – measured topic frequency correlates well with real-world events and the election-related conclusions based on measuring models are in line with conclusions drawn from social-scientific studies.

Observations from two media agenda analysis studies and the analysis of collected topic data underlined two problems related to methods of topic model evaluation. The first is the problem of measuring topic quality – the studies both confirmed variations in topic quality and indicated the inadequacy of existing word-based measures of topic coherence. The second is the problem of topic coverage – while the data confirms the limited ability of a single topic model to cover all the semantic topics, no available methods for measuring topic coverage exist, so it is not possible to identify the high-coverage models. These observations motivated the development of new methods of topic model evaluation – document-based coherence measures and methods for topic coverage analysis.

As described, the analysis of topics produced during the applications of topic discovery confirmed variations in topics' quality and underlined the need for better measures of topic quality. The analysis also indicated that existing word-based measures of topic coherence are inadequate for evaluating quality of media topics often characterized by semantically unrelated word sets. Based on the observation that media topics can be successfully interpreted using topic-related documents, a new class of document-based topic coherence measures is proposed.

The proposed measures calculate topic coherence in three steps: selection of topic-related documents, document vectorization, and computation of the coherence score from document vectors. Topic-related documents are selected using a simple model-independent strategy – a fixed number of documents with top document-topic weights is selected. Two families of document vectorization methods are considered. The first family consists of two standard methods

based on calculation of word and document frequencies: probabilistic bag-of-words vectorization and tf-idf vectorization. Methods in the second family vectorize documents by aggregating either CBOW or GloVe word embeddings. Three types of methods are considered for coherence score computation: distance-based methods that model coherence via mutual document distance, probability-based methods that model coherence as probabilistic compactness of document vectors, and graph-based methods that model coherence via connectivity of the document graph. The space of all the coherence measures is parametrized and sensible parameter values are defined to obtain a smaller set of several thousand measures. Then the selection and evaluation of the coherence measures is performed, using model topics manually labeled with document-based coherence scores and using the area under the ROC curve (AUC) as the performance criterion. The measures are partitioned in structural categories and the best measure from each category is selected using AUC on the development set as a criterion. These best measures are then evaluated on two test sets containing English and Croatian news topics.

The evaluation of document-based coherence measures shows that the graph-based measures achieve best results. Namely, best approximators of human coherence scores are the graph-based measures that use frequency-based document vectorization, build sparse graphs of locally connected documents and calculate coherence by aggregating a local connectivity score such as closeness centrality. Quantitative evaluation of word-based measures confirms the observations that word-based measures fail to approximate document-based coherence scores well and qualitative evaluation of coherence measures indicates that document- and word-based coherence measures complement each other and should be used in combination to obtain a more complete model of topic coherence.

Motivated by the data from the topic discovery steps performed in two media agenda analyses and by the obvious need to increase the number of topics discovered by a single topic model, the problem of topic coverage is defined and solutions are proposed. This problem occurrs in application of topic models to any text domain, i.e., it is domain-independent and extends beyond applications to media text. The problem of topic coverage consists of measuring how automatically learned model topics cover a set of reference topics – topical concepts defined by humans. Two basic aspects of the problem are the reference topics that represent the concepts topic models are expected to cover and the measures of topic coverage that calculate a score measuring overlap between the model topics and reference topics. Finally, the third aspect encompasses evaluation of a set of topic models using a reference set and coverage measures.

The coverage experiments are conducted using two datasets that correspond to two separate text domains – news media texts and biological texts. Each dataset contains a text corpus, a set of reference topics, and a set of topic models. Reference topics consist of topics that standard topic models are expected to be able to cover. These topics are constructed by human inspection, selection, and modification of model-learned topics. Both sets of reference topics

are representative of useful topics discovered during the process of exploratory text analysis.

Two approaches to measuring topic coverage are developed – an approach based on supervised approximation of topic matching and an unsupervised approach based on integrating coverage across a range of topic-matching criteria. The supervised approach is based on building a classification model that approximates human intuition of topic matching. A binary classifier is learned from a set of topic pairs annotated with matching scores. Four standard classification models are considered: logistic regression, support vector machine, random forest, and multilayer perceptron. Topic pairs are represented as distances of topic-related word and document vectors using four distinct distance measures: cosine, hellinger, L_1, and L_2. Model selection and evaluation shows that the proposed method approximates human scores very well, and that logistic regression is the best-performing model. The second proposed method for measuring coverage uses a measure of topic distance and a distance threshold to approximate the equality of a reference topic and a model topic. The threshold value is varied and for each threshold coverage is calculated as a proportion of reference topics that are matched by at least one model topic at a distance below the threshold. Varying the threshold results in a curve with threshold values on the x-axis and coverage scores on the y-axis. The final coverage score is calculated as the area under this curve. This unsupervised measure of coverage, dubbed area under the coverage-distance curve, correlates very well with the supervised measures of coverage, while the curve itself is a useful tool for visual analysis of topic coverage. This measure enables the users to quickly perform coverage measurements on new domains, without the need to annotate topic pairs in order to construct a supervised coverage measure.

Using the proposed coverage measures and two sets of reference topics, coverage experiments in two distinct text domains are performed. Experiments consist of measuring coverages obtained by a set of topic models of distinct types constructed using different hyperparameters. In addition to demonstrating application of coverage methods, the experiments show that the NMF model has high coverage scores, is robust to domain change and able to discover topics on a high level of precision. Nonparametric model based on Pitman-Yor priors achieves the best coverage for news topics.

Two proposed methods of topic model evaluation – document-based coherence measures and methods devised for solving the coverage problem – are applied in order to improve the previously proposed topic-model-based method of media agenda analysis. The improvements refer to the step of topic discovery and lead to quicker discovery of a larger number of concepts. This is achieved by using more interpretable models with higher coverage, and by ordering model topics, before human inspection, in the descending order of their coherence. These improvements conclude the contribution of the thesis related to the methods of computational media agenda analysis. The first improvement is based on the analysis of the coverage and document-based coherence scores measured for a large number of different topic models. The main result

is the recommendation for using the NMF model as the default model for topic discovery, due to the fact that NMF proved as a robust, interpretable, and a high-coverage model with the additional advantage of being fast to train. In addition, the nonparametric topic model based on Pitman-Yor priors also proved as a good choice for exploratory analysis of news texts since it achieves a very high coverage. The second improvement is model-agnostic and consists of ordering model topics, inspected during the topic discovery step, by descending topic coherence. This results in low-quality topics being pushed towards the end of the topic inspection queue. The experiments show that applying the best graph-based coherence measure in the described way significantly improves the discovery rate of semantic topics. Several other improvement recommendations are given based on the experience gained in the course of application of the media agenda analysis methods. These improvements include: improving the topic inspection and interpretation process by way of discarding the shared reference table of semantic topics, improving the step of measuring topics' frequency, and using tools that lead to quick guided discovery of topics of interest.

Research of the media agenda analysis methods and the methods of topic model evaluation revealed a number of technical problems related to usage, construction, storage, and retrieval of topic models. Namely, in topic modelling experiments it is often necessary to construct a large number of models by varying model types and various parameters of the construction process such as hyperparameters, low-level resources, and preprocessing components. A systematic solution to these problems is proposed – a framework for resource building and management in topic modeling. The framework's architecture is based on four principles, which in combination define a general and flexible method for designing and building code for evaluation and application of topic models. In addition, an application for building corpora of media text by collecting texts from a set of web news feeds was developed, as well as a graphical user interface that supports the topic discovery and topic frequency measurement.

The topic modelling framework, dubbed *pytopia*, is an object-oriented Python framework that can be viewed as a middleware framework located between application-level code and the algorithm-level frameworks such as TensorFlow. The framework's architecture is based on four design principles: the principle of standard interfaces and the adaptation of various components to these interfaces, the principle of component identifiability, the principle of using the abstraction dubbed Context to organize and retrieve components, and the principle of hierarchical compositionality that reflects the structure of text-mining components and facilitates their design and implementation. The framework contains core functionality that supports the four design principles, and functionality for component building, saving, and loading. The framework also contains a set of components related to topic modeling, ranging form basic resources such as dictionaries and corpora to complex components such as sets of vectorized texts and topic models. Finally, the framework includes several tools for topic model evaluation as well as

logging and testing functionality.

A corpus of media texts is a basis for media agenda analysis based on topic models. Motivated by the need for a tool that enables maximum flexibility in defining and building such a corpus, an application for collecting texts from a set of news feeds is developed. This application, dubbed *feedsucker*, enables the user to build a corpus of news texts containing texts from a set of sources corresponding to the exact user interest. The user specifies a set of news feeds in a text format and runs the application that continuously collects new texts and stores them in a database. The application is Java-based, object-oriented, and extensible.

This thesis describes the research of computational methods for media agenda analysis, which enable discovery and measuring of topics in large news corpora and find applications in a range of scientific and commercial analyses of media text. The researched methods are based on topic models, standard machine learning models for analysis of topical text structure. In the first phase of the research, an analysis of existing media agenda methods is performed and a new method that improves and systematizes the existing ones is proposed. The application of the proposed method in two use-cases underlined the need for new methods of topic model evaluation that would improve the efficiency of topic-model based tools. Consequently, two new methods of topic model evaluation are proposed – document-based measures of topic coherence and methods for analysis of topic coverage. These evaluation methods are then applied to improve the initially proposed method for media agenda analysis. In addition, research of topic model applications and methods of topic model evaluation led to a framework for resource building and management in topic modelling. The four main contributions of the thesis are: a method for computational analysis of the media agenda based on topic models, document-based measures of topic coherence, methods for analysis of topic coverage, and the framework for resource building and management in topic modelling.

The research described in the thesis led to an improved method for media agenda analysis and new methods of topic model evaluation. The evaluation methods find applications more general than the media agenda analysis – the measures of document-based coherence are applicable to any topic-model-based analysis of news text while the methods related to the problem of topic coverage are domain-independent. These evaluation methods represent new approaches with a potential to provide new insights about topic models, a class of widely-used machine learning models of text. The topic modelling framework could serve the same purpose since it facilitates complex experiments.

Keywords: Media agenda, Topic models, Topic model evaluation, Topic coherence, Topic coverage, Topic distance measures, Topic model construction, Unsupervised learning, Supervised learning.

Sadržaj

Poglavlje 1

Uvod

Ovaj rad bavi se računalnim postupcima analize medijske agende koji se temelje na tematskim modelima. *Medijska agenda* definirana je kao skup tema o kojima mediji govore, a *tematski modeli* su standardni modeli strojnog učenja namijenjeni analizi tematske strukture teksta. Analiza medijske agende od interesa je za društvenoznanstvena istraživanja medija te za komercijalne i političke analize medija. Alati za analizu medijske agende temeljeni na tematskim modelima omogućuju automatsko otkrivanje tema u velikim zbirkama tekstova i automatiziraju mjerenje zastupljenosti tih tema. Podaci dobiveni tom analizom pružaju uvid u tip i strukturu tema koje se javljaju u medijima, omogućuju analizu supojavljivanja tema te mogu poslužiti za koreliranje zastupljenosti tema s drugim varijablama poput mjera ljudske percepcije tema i metapodataka o tekstovima.

Cilj ovog rada je razvoj učinkovitih računalnih postupaka za otkrivanje i mjerenje tema koje čine medijsku agendu. Početna točka za razvoj ovih postupaka su tematski modeli – klasa nenadziranih modela strojnog učenja široko korištena za eksplorativnu analizu tematske strukture teksta. Metode vrednovanja tematskih modela koriste se za analizu i vrednovanje modela te za odabir kvalitetnih modela. Naime, tematski modeli su koristan alat, no zbog stohastičnosti svojstvene procesu učenja modela kvaliteta naučenih tema može znatno varirati. Iz tog razloga metode njihovog vrednovanja, primjenjive za odabir kvalitetnih modela i tema, imaju potencijal za povećanje učinkovitosti na modelima temeljenih postupaka. Drugi važan aspekt primjene tematskih modela je razvoj programske podrške koja olakšava izgradnju i upravljanje s modelima i povezanim resursima te olakšava interakciju s modelima. Istraživanje opisano u ovom radu bavi se razvojem postupaka primjene tematskih modela na analizu medijske agende, razvojem i analizom metoda semantičkog vrednovanja tematskih modela te razvojem programske podrške za postupak analize agende i za eksperimente vrednovanja tematskih modela.

Uže područje ovog rada odgovara računalnoj analizi medijske agende i tematskom modeliranju. Pregled računalne analize medijske agende dan je u nastavku uvoda, dok je poglavlje 3 u potpunosti posvećeno toj temi. Tematsko modeliranje obuhvaća razvoj i oblikovanje temat-

skih modela i algoritama njihovog učenja te metode vrednovanja i primjene tematskih modela. Pregled opsežnog područja tematskih modela i metoda njihovog vrednovanja dan je u poglavlju 2. Primjene tematskih modela obuhvaćaju eksplorativnu analizu i vizualizaciju strukture teksta, ekstrakciju tematskih značajki tekstova te zadatke poput preporučivanja i dohvata tekstova. Od specifičnog interesa u ovom radu su primjene tematskih modela na analizu medijske agende, opisane u odjeljcima 3.1.1 i 3.1.2, te primjene na druge vrste analize medijskog teksta opisane u odjeljku 4.2. Razvoj programske podrške za analizu agende i upravljanje tematskim modelima povezan je i s područjem softverskog inženjerstva, odnosno potpodručjem te discipline koje se bavi razvojem programske podrške za izradu, vrednovanje i primjenu alata strojnog učenja.

Šire područje ovog rada obuhvaća presjecište tri potpodručja umjetne inteligencije: područja strojnog učenja (engl. *machine learning*), područja obrade prirodnog jezika (engl. *natural language processing*) te područja pretraživanja informacija (engl. *information retrieval*). Područje strojnog učenja obuhvaća izradu, analizu i vrednovanje računalnih modela koji mogu naučiti razne vrste uzoraka (engl. *patterns*) koji se javljaju u podacima i na taj način aproksimirati uže aspekte ljudskih kognitivnih sposobnosti, poput primjerice uočavanja tema u tekstovima. Alati strojnog učenja temelj su većine suvremenih postupaka analize i pretraživanja teksta. Obrada prirodnog jezika je područje primjene računalne tehnologije na analizu i generiranje prirodnog jezika. Metode obrade prirodnog jezika koriste znanja iz lingvistike i umjetne inteligencije, a moderni pristupi temelje se ponajviše na strojnom učenju. Pretraživanje informacija bavi se metodama dohvata podataka (engl. *retrieval*) koji zadovoljavaju neku informacijsku potrebu korisnika. Područje ovog rada dotiče se metoda za dohvat tekstnih podataka koje se temelje na metodama strojnog učenja i obrade prirodnog jezika. Konkretno, postupci analize medijske agende dohvaćaju podatke o strukturi i definiciji medijskih tema.

1.1 Analiza medijske agende

Analiza medijske agende jest analiza medijskih tema koja obuhvaća analizu strukture i prikaza tema, analizu supojavljivanja tema te analizu odnosa tema i drugih podataka, poput praćenja kretanja zastupljenosti tema u vremenu. Potreba za analizom medijske agende javlja se u nizu zadataka koji podrazumijevaju neku vrstu analize tematske strukture medijskih tekstova. Takvi zadaci mogu se javiti u društvenoznanstvenim istraživanjima medija poput istraživanja postavljanja agende (engl. *agenda setting*) i uokvirivanja (engl. *framing*), pri analizi medija od strane novinara te u komercijalnim analizama medijskog teksta. Razni zadaci analize medijske agende, koji mogu profitirati od metoda računalne analize predloženih u ovom radu, detaljnije su opisani u poglavlju 3.

Dva temeljna koraka analize medijske agende su otkrivanje tema koje se pojavljuju u medijima te mjerenje zastupljenosti tih tema. Izazov pri računalnoj analizi medijske agende je

učinkovito provođenje ova dva koraka – brzo otkrivanje što većeg broja tema te brzo i točno mjerenje zastupljenosti. Standardni alat strojnog učenja za zadatke otkrivanja tema u zbirkama tekstova su tematski modeli – familija nenadziranih modela koji na temelju skupa tekstova uče teme modela reprezentirane listama riječi i tekstova. Mjerenje zastupljenosti se također često provodi na temelju tematskih modela, preciznije na temelju podataka modela o "težini" pojedine teme unutar dokumenta.

Tematski modeli primijenjeni su za otkrivanje i mjerenje tema medijske agende u nizu postojećih istraživanja. U tim primjenama glavni je izazov izgradnja kvalitetnih modela i postojeća istraživanja često izvještavaju o pojavi beskorisnih tema koje su ili neinterpretabilne ili ne otkrivaju teme od interesa za istraživača. Problem povezan s otkrivanjem tema je standardna upotreba samo jednog tematskog modela koji očekivano ne može otkriti sve teme. Važan problem je i nemogućnost definicije novih tema od interesa – analiza je ograničena na skup tema koje model nauči a te teme su naučene stohastičkim metodama i mogu varirati.

Pristup računalnoj analizi agende predložen u ovom radu koristi veći broj tematskih modela u koraku otkrivanja te uključuje mogućnost definicije i mjerenja novih tema od interesa za analitičara. Problemu kvalitete tematskih modela pristupa se primjenom mjera vrednovanja modela. Predložena metoda računalne analize medijske agende opisana je u poglavlju 3, dok se u poglavlju 6 opisuje primjena predloženih metoda vrednovanja tematskih modela na unaprijeđenje te metode.

S računalnom analizom medijske agende povezani su sljedeći dijelovi izvornog znanstvenog doprinosa ovog rada.

- Računalni postupak za analizu medijske agende koji obuhvaća otkrivanje i definiranje tema u zbirci medijskih tekstova te mjerenje agende označavanjem tekstova temama;
- Radni okvir i odgovarajuća programska podrška za računalno potpomognutu analizu medijske agende.

Treba napomenuti da je drugi dio doprinosa, radni okvir, ujedno usko vezan i uz problematiku vrednovanja tematskih modela opisanu u sljedećem odjeljku. Razvoj radnog okvira motiviran je problemima koji se javljaju pri razvoju programske podrške za postupak analize medijske agende i za vrednovanje tematskih modela. Radni okvir temelji se na općenitoj arhitekturi koja definira pristup zadacima upravljanja, izgradnje, pohrane i dohvata tematskih modela i vezanih komponenti.

1.2 Vrednovanje tematskih modela

Tematski modeli klasa su modela strojnog učenja često korištena za eksplorativnu analizu teksta. Međutim, kvaliteta tematskih modela varira zbog složenosti modelirane domene i stohastičnosti postupka učenja modela. Iz tog razloga razvijen je niz metoda vrednovanja tematskih

modela koje računaju ocjene kvalitete modela i njihovih tema, a pregled tih metoda dan je u poglavlju 2. U ovom radu se metode vrednovanja tematskih modela primjenjuju za odabir tema i modela koje dovode do efikasnijeg i bržeg otkrivanja tema na medijskoj agendi. U kontekstu te primjene, od glavnog interesa su metode koje vrednuju semantička svojstva tema poput interpretabilnosti tema i njihovog poklapanja s konceptima.

Najvažnija primjena tematskih modela u analizi medijske agende je primjena u koraku otkrivanja tema. Otkrivanje tema tematskim modelima provodi se izgradnjom jednog ili više modela na temelju zbirke medijskih tekstova, nakon čega analitičari pregledavaju i interpretiraju teme modela. Pri tome neke teme modela odgovaraju semantičkim temama odnosno konceptima dok su neke teme neinterpretabilne i odgovaraju slučajnom šumu ili mješavini većeg broja koncepata. Prva vrsta tema otkriva koncepte koji analitičaru mogu pružiti uvid u medijsku agendu, dok druga vrsta tema troši vrijeme uloženo u njihov pregled i interpretaciju te smanjuje kapacitet modela za otkrivanje korisnih tema. Stoga su od interesa mjere kvalitete modela i tema koje se mogu iskoristiti za odabir interpretabilnijih tema i modela i na taj način ubrzati postupak otkrivanja tema ili dovesti do većeg broja otkrivenih tema. Primjene postupka analize medijske agende, opisane u poglavlju 3, motivirale su razvoj dviju novih metoda koje vrednuju aspekte kvalitete tema i modela važne za postupak otkrivanja tema.

Prva od tih metoda vrednovanja je metoda za računanje dokumentne koherentnosti tema. Metoda je motivirana potrebom za ubrzanjem postupka otkrivanja tema pomoću filtriranja neinterpretabilnih tema modela. Postojeće mjere koherentnosti temeljene su na koherentnosti riječi, no u slučaju medijskih tema dokumentna koherentnost predstavlja bolji model interpretabilnosti. Razvoj i vrednovanje mjera dokumentne koherentnosti opisani su u poglavlju 4. Druga metoda vrednovanja tematskih modela rješava problem pokrivenosti tema. Metoda je motivirana potrebom za odabirom tematskih modela koji otkrivaju što veći broj koncepata. Ovaj važan problem tematskog modeliranja slabo je istražen i u poglavlju 5 se opisuje pristup mjerenju pokrivenosti koji se sastoji od izgradnje skupa koncepata te od nadziranih i nenadziranih mjera pokrivenosti skupa koncepata temeljenih na ocjenjivanju poklapanja tema modela i koncepata. Obadvije predložene metode vrednovanja imaju primjene koje nadilaze analizu medijske agende – metoda dokumentne koherentnosti može se primijeniti u slučaju bilo koje vrste tematskog modeliranja medijskog teksta, dok su metode mjerenja pokrivenosti domenski neovisne.

S vrednovanjem tematskih modela povezani su sljedeći dijelovi izvornog znanstvenog doprinosa ovog rada.

- Postupak vrednovanja koherentnosti tema dobivenih tematskim modelom temeljen na analizi semantičke sličnosti dokumenata;
- Postupak vrednovanja pokrivenosti tema temeljen na usporedbi tema dobivenih tematskim modelom s referentnim skupom tema;

1.3 Struktura rada

Ostatak rada strukturiran je na sljedeći način. U poglavlju 2 dan je sažet pregled područja tematskog modeliranja i metoda vrednovanja tematskih modela. Ovi modeli i metode sadrže temeljne alate koji se u radu koriste za računalnu analizu medijske agende. Poglavlje 3 opisuje problematiku, primjene i metode računalne analize medijske agende te daje prijedlog nove metode koja otklanja neke nedostatke postojećih pristupa. U poglavlju 4 opisana je predložena metoda vrednovanja tematskih modela koja računa semantičku koherentnost tema na temelju dokumenata povezanih s temama. Slijedi opis predložene metode vrednovanja pokrivenosti tema u poglavlju 5. Nakon opisa metoda vrednovanja, u poglavlju 6 opisuju se poboljšanja postupka računalne analize medijske agende temeljena na tim metodama. U poglavlju 7 opisan je radni okvir za upravljanje tematskim modelima te ostala programska podrška za postupke analize medijske agende. Konačno, poglavlje 8 iznosi pregled i zaključke rada te predlaže smjerove za daljnja istraživanja.

Poglavlje 2

Tematski modeli i njihovo vrednovanje

U ovom poglavlju opisuju se tematski modeli (engl. *topic models*), klasa nenadziranih modela za brojnim primjenama u obradi prirodnog jezika. Tematski model je nenadzirani model teksta koji iz skupa tekstnih dokumenata uči teme opisane utežanim skupovima riječi i dokumenata. Tablica 2.1 sadrži primjere tema naučenih iz medijskog teksta. Teme očekivano odgovaraju konceptima i mogu se koristiti kao sažeci za eksplorativnu analizu tekstnih zbirki ili kao značajke za razne zadatke obrade prirodnog jezika. Pristupi analizi medijske agende razmatrani i predloženi u ovom radu temelje se na tematskim modelima koji se koriste za otkrivanje tema prisutnih u medijskim tekstovima te za mjerenje zastupljenosti tih tema.

Glavna prednost tematskih modela je to što su nenadzirani, odnosno za njihovo učenje nisu potrebni od ljudi označeni podaci, te ujedno zahtijevaju vrlo malo jezičnog pretprocesiranja. S druge strane, kvaliteta naučenih tema može dosta varirati, a taj problem dodatno je otežan aproksimativnim i stohastičkim procesom učenja tematskih modela. Zbog toga kvaliteta naučenih tema nije unaprijed zajamčena već je modele potrebno vrednovati i u tu je svrhu razvijeno niz metoda za ocjenu raznih aspekata njihove kvalitete, poput stabilnosti modela i semantičke koherentnosti tema. U nastavku poglavlja slijedi pregled tematskih modela s naglaskom na vjerojatnosni pristup, nakon čega slijedi pregled metoda vrednovanja tematskih modela.

2.1 Tematski modeli

Tematski modeli [1] su nenadzirani modeli teksta s brojnim primjenama u analizi teksta koje uključuju eksploratornu analizu zbirke tekstova [2], pretraživanje informacija [3], ekstrakciju značajki [4] te zadatke obrade prirodnog jezika, poput razlikovanja značenja riječi [5] i analize sentimenta [6]. Dvije prevladavajuće klase tematskih modela su generativni vjerojatnosni modeli, koji modeliraju teme kao vjerojatnosne razdiobe nad riječima te faktorizacijski modeli, koji aproksimiraju dokument-riječ matricu skupom latentnih faktora koji odgovaraju temama.

Generativni vjerojatnosni modeli su prevladavajući pristup tematskom modeliranju. Ti mo-

Tablica 2.1: Teme modela naučenog iz zbirke američkih medijskih tekstova. Za svaku temu prikazane su vezane riječi i naslovi vezanih tekstova.

`oznaka teme`: Obamacare	`riječi teme`: osiguranje, razmjena, vrhovni, pokrivenost, dostupan, subvencija, sudac, tužba, žalba
Vrhovni sud odlučuje o ključnim odredbama Obamacare zakona Sve što trebate znati o osporavanju Obamacare zakona Slučaj Obamacare ponovo pred Vrhovnim sudom	
`oznaka teme`: Neutralnost interneta	`riječi teme`: internet, povezan, propis, neutralan, komisija, FCC, Google, naslov, korisnik
SAD razmatra zabranu kontroverznog prioritiziranja Internet korisnika Predsjednik FCC-a predlaže stroga Internet pravila FCC planira uvođenje strožih Internet pravila	
`oznaka teme`: Cijepljenje	`riječi teme`: cjepivo, roditelj, bolest, ospice, izbijanje, ebola, djeca, imuni, širiti, znanost
Christie izjavio da je cijepljenje izbor roditelja Ospice stavile težak izbor pred Kongres John Boehner: “Svu djecu bi trebalo cijepiti”	

deli opisuju vjerojatnosni proces generiranja teksta a njihova struktura opisana je skupom slučajnih varijabli i veza među njima. U primjenama su varijable od interesa pretežno *teme*, definirane kao vjerojatnosne razdiobe nad riječima rječnika i povezane s tekstovima varijablama koje definiraju vjerojatnosti pojavljivanja tema unutar tekstnih dokumenata. Tablica 2.1 sadrži primjere tema tematskog modela naučenog iz novinskih tekstova. Svaka tema prikazana je skupom najvjerojatnijih riječi te skupom novinskih članaka s najvišom vjerojatnošću pojavljivanja teme u tekstu.

Prvi, i u primjenama najpopularniji, tematski model je latentna Dirichletova alokacija (engl. *Latent Dirichlet Allocation* – LDA) [1]. Model LDA je reprezentativan za većinu tematskih modela, a mnogi tematski modeli su izravna nadogradnja njegove strukture. Model LDA pretpostavlja fiksni broj tema K, pri čemu je svaka tema definirana kao vjerojatnosna razdioba nad riječima rječnika. Teme se modeliraju pomoću vjerojatnosne matrice ϕ, pri čemu je ϕ_{ij} vjerojatnost pojavljivanja riječi j u temi i. Analogno, dokumenti se modeliraju kao vjerojatnosna matrica θ, pri čemu je θ_{ij} vjerojatnost teme j u i-tom dokumentu. Pretpostavka modela je da su tekstovi generirani vjerojatnosnim procesom koji se odvija na sljedeći način. Na početku se svaka tema ϕ_i (multinomijalna razdioba) uzorkuje iz Dirichletove apriorne razdiobe s parametrom $\vec{\beta}$. Zatim se za svaki dokument D_i multinomijalna razdioba θ_i uzorkuje iz Dirichletove

apriorne razdiobe s parametrom $\vec{\alpha}$. Naposlijetku, za riječ na poziciji j unutar i-tog dokumenta prvo se iz razdiobe θ_i uzorkuje pripadna tema z_{ij}, nakon čega se sama riječ uzorkuje iz $\phi_{z_{ij}}$, razdiobe vjerojatnosti svih riječi unutar teme. Opisani generativni proces može se sažeto opisati pomoću izraza koji povezuju vjerojatnost matrice tema ϕ te vjerojatnost pojedinog dokumenta D_i s ostalim varijablama modela:

$$p(\phi) = \prod_{i=1}^{K} \mathrm{Dir}(\phi_i|\vec{\beta})$$
$$p(D_i) = \mathrm{Dir}(\theta_i|\vec{\alpha}) \prod_j \mathrm{Mult}(z_{ij}|\theta_i)\mathrm{Mult}(w_{ij}|\phi_{z_{ij}})$$

Tekstni dokumenti $D = \{D_i\}$ su opažene varijable modela i pomoću algoritama učenja mogu se aproksimirati vjerojatnosti riječi za teme prema aposteriornoj razdiobi $p(\phi|D)$, vjerojatnosti tema za dokumente prema aposteriornoj razdiobi $p(\theta|D)$ te vjerojatnosti pridjeljivanja tema pojedinim riječima prema aposteriornoj razdiobi $p(z|D)$. Učenje se običajeno provodi algoritmima za aproksimativno statističko zaključivanje, poput Gibbsovog uzorkovanja (engl. *Gibbs sampling*) [7] i varijacijskog zaključivanja (engl. *variational inference*) [1, 8]. Predložena su brojna proširenja osnovnog modela LDA poput modeliranja metapodataka tekstova [9], veza među temama [10] te modela s varijabilnim brojem tema [11].

Iako su generativni vjerojatnosni modeli prevladavajući tematski modeli, nisu jedini pristup tematskom modeliranju. Popularnu alternativu predstavljaju modeli temeljeni na matričnoj faktorizaciji, poput latentne semantičke analize (engl. *latent semantic analysis* – LSA) [12] i nenegativne matrične faktorizacije (engl. *non-negative matrix factorization* – NMF) [13]. Ovi modeli uče skup latentnih faktora na način da aproksimiraju matricu težina riječi za dokumente (primjeri težina su vjerojatnosti riječi i tf-idf težine) umnoškom dviju matrica, matrice s težinama riječi za faktor te matrice s težinama faktora za dokumente. Pri tematskom modeliranju faktorizacijskim modelima latentni faktori se smatraju temama čije značenje je određeno pripadnim težinama dokumenata i riječi. Posebno se model NMF pokazao kao popularna alternativa vjerojatnosnim tematskim modelima [14, 15, 16, 17], a nedavno sustavno vrednovanje tematskih modela pokazuje da NMF kvalitetom može parirati vjerojatnosnom modelu LDA [18]. Slično kao i kod generativnih modela, za NMF su razvijene mnoge varijante osnovnog modela i algoritmi učenja [19].

Razvijen je velik broj tematskih modela čije strukture odražavaju različite pretpostavke o tematskoj strukturi teksta. U ovom radu naglasak nije na detaljima strukture modela ili jednom specifičnom tipu modela, već se tematski modeli promatraju iz općenite perspektive – kao modeli koji na temelju skupa tekstova na izlazu daju teme, objekte čija je semantika definirana pripadnim riječima i dokumentima ili, preciznije, utežanom listom riječi i utežanom listom tekstova. Stoga sve u ovom radu predložene metode od podataka sadržanih u modelu koriste samo

težine riječi za temu (engl. topic-word weights) zadane matricom ϕ te *težine tema za dokument (engl. document-topic weights)* zadane matricom θ. U slučaju vjerojatnosnih modela ove težine predstavljaju vjerojatnosti ili druge probabilističke varijable, no za tematske modele poput faktorizacijskih modela težine ne moraju imati vjerojatnosnu interpretaciju. Općenita interpretacija tih težina je da definiraju mjeru povezanosti između tema s jedne te riječi i dokumenata s druge strane. Težine za svaku temu definiraju i poredak riječi i poredak dokumenata prema povezanosti s temom. Teme se uobičajeno prikazuju manjim brojem po težinama najbolje rangiranih riječi i dokumenata (u pravilu 10-20 riječi te nekoliko desetaka dokumenata). Za te riječi i dokumente se u ovom radu koriste nazivi *najbolje rangirane riječ* i *najbolje rangirani dokumenti*.

Ovaj općeniti pogled na teme, te način njihovog prikaza, predočen je u tablici 2.1 u kojoj su teme prikazane najbolje rangiranim riječima i dokumentima, poredanima padajuće po težini. Ovim pristupom postiže se općenitost u radu predloženih metoda analize medijske agende i metoda vrednovanja tematskih modela – metode analize kao svoju ključnu komponentu mogu koristiti mnoštvo trenutno dostupnih i budućih modela, a metode vrednovanja primjenjive su na široki spektar modela.

Tematski modeli imaju brojne primjene u analizi teksta i obradi prirodnog jezika. Često se primjenjuju za eksplorativnu analizu tekstnih zbirki, pri čemu se naučene teme modela (predočene listom riječi i dokumenata) koriste kako bi se dobio pregled tema koje se javljaju u tekstovima. Razvijeni su mnogi vizualizacijski alati koji pružaju dodatne uvide vizualizacijom strukture tema, veza među temama i veza između tema i tekstova. Drugi smjer primjena koristi teme za reprezentaciju objekata od interesa (poput riječi, dokumenata, autora, web-sjedišta itd.) pomoću skupa tema i njihovih težina. Takve tematske reprezentacije koriste se kao značajke za klasifikaciju, za dohvat objekata sa sličnom tematskom strukturom te za prikaz objekata. U ovom radu naglasak je na primjenama tematskih modela na analizu medijskih tekstova, pri čemu glavnu primjenu predstavljaju znanstvena istraživanja medijskog teksta. Pregled tih primjena fokusiranih na otkrivanje tema i mjerenje njihove zastupljenosti dan je u poglavlju 3, s naglaskom na znanstvenoistraživačkim primjenama. U poglavlju 4 dan je pregled raznolikih primjena u medijskoj domeni, dok je u poglavlju 5 opisano nekoliko domenski raznolikih primjena tematskih modela povezanih s idejom tematske pokrivenosti.

2.2 Vrednovanje tematskih modela

Korisnost tematskih modela izravno ovisi o kvaliteti naučenih tema, koja ovisi o nizu faktora i predstavlja protutežu opisanim prednostima tih modela. S jedne strane, uspješna primjena tematskih modela ovisi o faktorima vezanim uz odabir tematskog modela i algoritma učenja, postavljanje hiperparametara i način pretprocesiranja teksta. S druge strane, čak i kada su pret-

hodne odluke donesene, kvaliteta modela može varirati zbog stohastičke prirode algoritama učenja. Metode za automatsko vrednovanje tematskih modela korisne su u oba slučaja: mogu se iskoristiti za donošenje odluka prilikom oblikovanja rješenja, kao i za odabir najkvalitetnijih od većeg broja modela naučenih na temelju istog algoritma i parametara.

Metode vrednovanja tematskih modela mogu se podijeliti na ekstrinzične metode (ovisne o zadatku) i intrinzične metode (neovisne o zadatku). Pri ekstrinzičnom pristupu kvaliteta modela vrednuje se na temelju poboljšanja performansi na zadatku obrade prirodnog jezika koji koristi model, primjerice zadatku pretraživanja informacija [3], razlikovanja značena riječi [20], analize sentimenta [21] te računanja sličnosti riječi i klasifikacije dokumenata [22]. S druge strane intrinzični pristup vrednuje kvalitetu naučenih tema i modela neovisno o primjeni. U ovom radu veći naglasak je na šire primjenjivim metodama intrinzičnog vrednovanja modela.

Intrinzične metode mogu se podijeliti u četiri kategorije: mjere vjerojatnosnog poklapanja, mjere stabilnosti, mjere poklapanja sa zlatnim standardom (engl. *gold standard*, *ground truth*), te mjere kvalitete tema. Pristupi vjerojatnosnog poklapanja na temelju vjerojatnosne strukture tematskih modela formuliraju mjeru poklapanja ili raskoraka između modela i podataka. Najpopularnija mjera iz ove kategorija je mjera zbunjenosti (engl. *perplexity*), obrnuto proporcionalna vjerojatnosti koju naučeni model pridružuje izdvojenom skupu tekstnih podataka [1, 23]. Sofisticiranija vjerojatnosna mjera predložena u [24] mjeri raskorak između empirijski procijenjenih svojstava latentnih varijabli modela i svojstava očekivanih na temelju vjerojatnosne strukture modela.

Mjere stabilnosti motivirane su varijabilnošću naučenih tema i činjenicom da je stabilnost modela poželjno svojstvo u nizu primjena, poput društvenoznanstvenih analiza. Stabilnost skupa tematskih modela računa se uprosječivanjem sličnosti parova modela. Sličnost se može izračunati uparivanjem međusobno sličnih tema dvaju modela [25, 26, 27], ili na temelju usporedbe reprezentacija modela temeljenih na riječima i dokumentima [27].

Vrednovanje tematskih modela može se postaviti i kao računanje poklapanja tema modela ili iz njih izvedenih podataka s referentnim oznakama. U [28] teme se promatraju kao meke grupe (engl. *soft clusters*) i tematski modeli se vrednuju poopćenom metodom za ocjenu kvalitete grupiranja. Pristup predložen u [29] sastoji se od metode uparivanja tema modela se konceptima izrađenima od strane ljudi.

Mjere kvalitete tema računaju ocjene kvalitete pojedinih tema, koje je moguće agregirati kako bi se dobila ocjena modela. U [30] autori mjere kvalitetu tema kao udaljenost između vjerojatnosnih razdioba dokumenata i riječi za temu s jedne, te neinformativnih razdioba (uniformne i "prazne") s druge strane. Autori u [31] pristupaju vrednovanju kvalitete tema kao zadatku pronalaženja riječi "uljeza": ljudima je dan zadatak da prepoznaju riječi uljeze ubačene u skup uz temu vezanih riječi, s idejom da će za interpretabilne teme biti lakše uočiti uljeze. U [32] se predlaže metoda automatizacije zadatka pronalaženja riječi uljeza iz [31]. U [33]

računa se "pojmovna značajnost" tema koja mjeri lakoću pridruživanja pojma nekoj temi. Postupak se izvodi preslikavanjem riječi teme na WordNet[1] pojmove i pronalaženjem maksimalno specifičnih WordNet pojmova koji obuhvaćaju preslikane pojmove.

Novija i sve popularnija metoda računanja kvalitete tema temelji se na ideji *koherentnosti tema* koja se okvirno definira kao poklapanje teme s nekim konceptom [35]. Postojeći pristupi računanju tematske koherentnosti temeljeni su na riječima – pretpostavljaju da koherentnost teme odgovara koherentnosti uz temu vezanih riječi i računaju je temeljem skupa najbolje rangiranih riječi za temu. Detaljniji pregled metoda tematske koherentnosti dan je u poglavlju 4 gdje se predlaže nova metoda računanja tematske koherentnosti na temelju uz temu vezanih dokumenata.

Tematski modeli i metode njihovog vrednovanja važni su za sve metode predložene u ovom radu. U poglavlju 3 opisuju se i predlažu metode analize medijske agende temeljene na upotrebi tematskih modela za otkrivanje tema i mjerenje zastupljenosti tema na medijskoj agendi, pri čemu se metode vrednovanja mogu iskoristiti za ocjenu i povećanje kvalitete korištenih tematskih modela. U poglavlju 4 predlaže se nova metoda vrednovanja tematskih modela koja računa ocjene koherentnosti tema na temelju vezanih dokumenata. U poglavlju 5 predlaže se nova metoda vrednovanja tematskih modela koja mjeri poklapanje tema modela s referentnim skupom semantičkih tema. U poglavlju 6 predlaže se na tematskim modelima temeljena metoda analize medijske agende dobivena poboljšanjem prethodno predložene metode pomoću metoda vrednovanja tematskih modela. U poglavlju 7 opisuje se radni okvir izgrađen sa svrhom učinkovite izgradnje većih skupova tematskih modela u primjenama poput razvoja metoda vrednovanja modela .

[1] WordNet [34] je taksonomija pojmova koji odgovaraju značenjima riječi.

Poglavlje 3

Računalna analiza medijske agende

3.1 Uvod

Pojam medijske agende (engl. *media agenda*) koristi se u istraživanjima postavljanja agende [36, 37] (engl. *agenda setting*), gdje označava skup važnih pitanja (engl. *issues*) o kojima se govori u medijima.[1] Pojam medijske agende koji se koristi u ovom radu nastao je poopćavanjem prethodne definicije i označava općeniti skup tema koje se pojavljuju u medijima. Vezani pojam *analize medijske agende* uveden je kao generalizacija koja obuhvaća niz analiza od znanstvenog i komercijalnog interesa koje se provode na zbirkama medijskih tekstova i koje su detaljnije opisane u nastavku poglavlja.

Medijska agenda sastoji se od skupa tema o kojima mediji govore. *Analiza medijske agende* sastoji se od dva temeljna zadatka: zadatka *otkrivanja skupa tema* koji čine agendu te zadatka *mjerenja zastupljenosti* tih tema. Rezultati koraka otkrivanja tema su skup tema i njihove definicije. Mjerenje pridružuje temama neku mjeru zastupljenosti poput broja članaka, rečenica ili riječi vezanih uz temu.

U nekim analizama je od glavnog interesa otkrivanje što većeg broja tema odnosno precizna karakterizacija medijske agende. S druge strane, ako analitičar unaprijed zna teme koje želi mjeriti, odmah se može pristupiti mjerenju bez prethodnog otkrivanja. Otkrivanje i mjerenje tema tradicionalno provode ljudi čitanjem tekstova i kodiranjem tekstova u tematske kategorije, što su vremenski zahtjevni zadataci i tehnike strojnog učenja omogućavaju automatizaciju ovih koraka koja dovodi do značajnih ušteda vremena i omogućava obradu velikih zbirki tekstova, no uz cijenu smanjenja kvalitete rezultata analize.

Analiza medijske agende ima brojne primjene u znanstvenim i komercijalnim analizama medijskog teksta. Iz općenite perspektive, rezultati analize su skup interpretiranih i označenih tema i izmjerene zastupljenosti tih tema a različite primjene se razlikuju u načinu primjene

[1] Pojam agende u engleskom jeziku označava program, plan ili cilj. Primjerice, govori se o političkoj agendi (engl. *political agenda*), agendi sastanka (engl. *meeting agenda*), ili osobnoj agendi (engl. *personal agenda*).

tih podataka koji se mogu koristiti sami za sebe (npr. pregledno promatranje tema i analiza zastupljenosti) ili ih je moguće korelirati sa podacima o temama dobivenima iz drugih izvora (npr. podacima o javnoj percepciji važnosti tema).

Prevladavajuća klasa modela strojnog učenja u računalnoj analizi medijske agende su tematski modeli [11] (eng. topic models), nenadzirani modeli teksta koji iz skupa tekstova uče teme – utežane skupove riječi i tekstova. U ovom radu predlaže se računalni pristup analizi medijske agende temeljen na tematskim modelima, pri čemu naglasak nije na specifičnom tipu modela čija struktura je prilagođena medijskom tekstu, nego se model promatra kao skup naučenih tema a naglasak pristupa je na operacionalizaciji modela i na metodama validacije modela. Predložena metoda analize agende sastoji se od tri koraka – otkrivanja tema na temelju nekolicine tematskih modela, definicije tema od interesa za analizu, te mjerenja tema pomoću slabo nadziranih tematskih modela. Ovaj pristup nastoji otkloniti uočene nedostatke postojećih pristupa – otkrivanje tema temeljeno na jednom modelu, nedostatak faze definicije i posljedično nerazmatranje problema mjerenja novodefiniranih tema, te nedostatak vrednovanja postupka mjerenja tema. Predložena eksperimentalna metoda testira se kroz primjenjenu na dvije studije slučaja analize medijske agende.

Poglavlje započinje pregledom istraživanja vezanih uz računalnu analizu medijske agende. To su društvenoznanstvena istraživanja medija koje koriste računalne alate ili razmatraju metodologiju njihove primjene te radovi iz područja računarstva koji razmatraju primjene i predlažu pristupe analizi medijske agende. Zatim se opisuje otvoren problem validacije nenadziranih tematskih modela, posebno važan u društvenoznanstvenim primjenama modela. Nakon toga slijedi detaljan opis predložene metode analize agende te opisi primjena metode na analizu američke i hrvatske medijske agende. Na kraju poglavlja analiziraju se kroz primjene uočeni nedostaci postojećih metoda validacije tematskih modela koji motiviraju istraživanja opisana u daljnjim poglavljima rada.

Slijedi pregled nekoliko važnih društvenoznanstvenih područja koja mogu profitirati od računalne analize medijske agende te pregled komercijalnih primjena.

3.1.1 Primjene u znanstvenim istraživanjima medija

Analiza sadržaja Analiza sadržaja [38] je istraživačka metodologija iz društvenih i humantističkih znanosti koja ima za cilj izvođenje replikabilnih i validnih zaključaka iz tekstnog sadržaja. Koraci analize sadržaja uključuju čitanje i analizu teksta, izradu analitičkih kategorija te kategorizaciju tekstova [38]. Kategorizacija koju provode ljudi prema dobro definiranim uputama naziva se kodiranjem (engl. *coding*). Društvenoznanstvena istraživanja koja mogu profitirati od metoda računalne analize medijske agende tradicionalno koriste analizu sadržaja, stoga se te metode može promatrati kao metode za računalnu analizu tekstnog sadržaja.

Tematski modeli, prevladavajući algoritamski alat za analizu medijske agende, primjenjuju

se u analizi sadržaja kako bi se djelomično automatizirali čitanje i kategorizacija teksta [39, 40]. Pritom se korak otkrivanja tema odnosno pregledna analiza tematske strukture tekstova smatra zamjenom za pregledno čitanje tekstova (engl. *initial reading*) te može služiti kao smjernica za podrobnije čitanje (engl. *deep reading*) [40]. Korak mjerenja zastupljenosti tema predstavlja zamjenu za postupak kodiranja [39], pri čemu tematski modeli automatski uče kategorije (teme) te provode kategorizaciju tekstova. Tematski modeli mogu znatno ubrzati analizu i omogućiti rad s velikim zbirkama tekstova, no pouzdanost rezultata analize nužno je osigurati validacijom modela od strane ljudi [39, 40, 41].

Postavljanje agende Postavljanje agende (engl. *agenda setting*) je utjecajna teorija medijskih efekata. Temelji se na ideji da mediji mogu oblikovati javnu percepciju važnosti pojedinih tema [42, 43]. Drugim riječima, postavljanje agende je efekt prijenosa važnosti s medija na medijsku publiku. Istraživanja postavljanja agende tipično nastoje otkriti kako medijska agenda (skup tema koje mediji pokrivaju) utječe na javnu agendu (skup tema koje javnost smatra bitnima), te daju uvide u ulogu masovnih medija i načine na koji oblikuju javno mišljenje.

Alati za računalnu analizu medijske agende mogu se primijeniti u istraživanjima postavljanja agende tako da se u koraku otkrivanja dobije pregled tema o kojima mediji govore, na temelju čega se vrši odabir i analiza tema od interesa [44, 45]. Zatim se u koraku mjerenja automatski mjeri zastupljenost istraživaču zanimljivih tema u zbirci tekstova. Izmjerene zastupljenosti tema zatim se mogu korelirati s mjerama važnosti tema koje daju odraz njihove javne percepcije.

Uokvirivanje Uokvirivanje (engl. *framing*) je teorija medijskih efekata koja proučava kako se stvarnost prikazuje u medijskom tekstu u svrhu promicanja određene perspektive [46]. Uokvirivanje podrazumijeva odabir i naglašavanje važnosti pojedinih aspekata stvarnosti [46]. Jedan aspekt uokvirivanja je tematska struktura teksta [47], odnosno neke teme mogu se dobro poklapati s pojedinim okvirima [48].

Računalna analiza medijske agende može se primijeniti na istraživanja uokvirivanja na način da se na zbirci medijskih tekstova koji govore o jednom društvenom pitanju provede automatsko otkrivanje tema, nakon čega se odaberu i analiziraju teme koje odgovaraju okvirima za to pitanje [48, 49].

Ostala istraživanja Alati za računalnu analizu medijske agende primjenjuju se ili imaju potencijal za primjenu u raznim drugim društvenoznanstvenim analizama medija. Teme automatski otkrivene u zbirkama medijskih tekstova mogu se iskoristiti za kvalitativnu analizu, primjerice za analizu novinskog razgraničavanja znanosti i lažne znanosti [50] ili za analizu povijesnih novinskih zbirki [51]. Pritom tematski modeli imaju potencijal za generiranje tema koje mogu analitičaru sugerirati raznolike perspektive i interpretacije sadržaja [40].

Automatsko otkrivanje i mjerenje zastupljenosti tema moglo bi se primijeniti na istraživanja procesa filtriranja medijskog sadržaja (engl. *gatekeeping*) poput istraživanja televizijskih vijesti [52] koje provodi kvantitativnu analizu zastupljenosti tematskih kategorija.

3.1.2 Primjene u neznanstvenim analizama medija

Alati za otkrivanje tema u zbirkama medijskih tekstova i mjerenje njihove zastupljenosti mogu koristiti profesionalcima poput novinara, poslovnih analitičara i istraživača tržišta. Otkrivanje tema može se koristiti za preglednu eksplorativnu analizu ili podrobniju analizu zbirke vijesti od interesa, primjerice vijesti iz određenog vremenskog razdoblja, vijesti iz određenih medijskih izvora, ili vijesti koje spominju neki entitet od komercijalnog interesa poput tvrtke ili robne marke. Mjerenje zastupljenosti tema može se koristiti za daljnje analize tema od interesa, poput vremenskog kretanja zastupljenosti tema ili korelacije zastupljenosti tema sa događajima, spominjanjima osoba i organizacija u tekstu te drugim varijablama od interesa.

Prethodno opisane znanstvene analize medijskog teksta mogle bi se iskoristiti i u komercijalne svrhe. Primjerice, od interesa bi moglo biti istraživanje kako agenda suparničkih medija korelira međusobno ili s javnom percepcijom važnih pitanja, ili kako ti mediji provode uokvirivanje nekih tema. Pritom u komercijalnim primjenama takvih analiza očekivano nije potrebna ista razina metodološke strogoće kao u znanstvenim istraživanjima.

Alati za računalnu analizu medijske agende potencijalno su koristan alat za analizu velikih baza medijskih tekstova, poput platformi koje objedinjuju kontinuirano sakupljanje vijesti iz niza izvora te alate za pretraživanje i analizu tih vijesti [53, 54, 55]. MediaCloud[2] [54] i MediaViz [55] platforme sadrže alate za pretraživanje te vizualizacijske alate temeljene na pretraživanju po ključnim riječima i metapodacima. Europe Media Monitor [3] [53], platforma koja sakuplja vijesti iz cijele Europe, osim navedenih sadrži i alate za automatsko sažimanje vijesti, detekciju događaja i imenovanih entiteta te za klasifikaciju vijesti u tematske kategorije.

Alati za pretraživanje i vizualizaciju temeljeni na temama iz automatske analize medijske agende mogli bi pružiti uvide komplementarne onima temeljenim na ključnim riječima i metapodacima pošto teme mogu otkriti analitičaru korisne apstraktne koncepte. U [54] su opisane neke primjene tematskih modela na preglednu analizu velikih baza novinskih tekstova i navedene su potencijalne koristi tematski orijentiranih analiza, poput analize načina obrade pojedinih društvenih pitanja od strane medija, za novinare i konzumente vijesti.

[2] https://mediacloud.org/
[3] http://emm.newsbrief.eu

3.1.3 Studije s računalnom analizom medijske agende

Ovdje je dan pregled istraživanja koja se bave primjenom metoda za računalnu analizu medijske agende temeljenih na alatima strojnog učenja. Svi opisani pristupi koriste tematske modele koji su postali široko prihvaćena klasa modela za automatsko otkrivanje i mjerenje zastupljenosti tema u zbirci tekstova. Istraživanja obuhvaćaju društvenoznanstvena istraživanja medija računalnim alatima, istraživanja usmjerena na metodološki aspekt primjene računalnih alata za analizu medija, te istraživanja iz područja računarstva fokusirana na računalne modele i njihove primjene u analizi medija. Ova istraživanja polaze od zbirke tekstova relevantne za istraživanje, grade tematski model koji služi za otkrivanje i mjerenje tema, provode neku vrstu validacije modela odnosno podataka dobivenih pomoću modela, te zatim analiziraju teme i njihove zastupljenosti i korelacije s drugim podacima kako bi izveli zaključke. Pri opisu istraživanja naglasak nije na prirodi i rezultatima istraživanja već na tehničkim aspektima pristupa poput izgradnje zbirke tekstova, izgradnje i validacije modela, te načina korištenja modelom dobivenih podataka. Validacija modela je važan korak nužan za potvrdu korisnosti modela za izvođenje zaključaka [41]. Postoji više metoda validacije no sveprisutna i temeljna metoda je pregled samih tema prikazanih listom najbolje rangiranih riječi i dokumenata za tu temu. Ova vrsta validacije ovdje se naziva *pregled i interpretacija tema* i obuhvaća postupak tijekom kojeg čovjek procjenjuje semantičku koheretnost riječi i dokumenata i nastoji ih interpretirati u terminima nekog koncepta.

Istraživanje postavljanja agende pri obradi teme trgovine ljudima opisano u [45] provedeno na 134 članka i pripadnim korisničkim komentarima (koji se koriste kao indikator javne agende) s web-portala Guardian[4]. Tekstovi su razlomljeni na paragrafe i provedeno je tematsko modeliranje faktorizacijskim LSA modelom. sa 11 tema (broj odabran “arbitrarno”, na temelju intuicije autora). Provedena je validacija tema njihovim pregledom i interpretacijom i svaka od tema detaljno je opisana. Teme su statističkim metodama grupirane u tri skupine: teme zastupljenije u novinskim člancima, teme zastupljenije u korisničkim komentarima te podjedanko zastupljene teme. Na temelju tog grupiranja provedena je daljnja analiza postavljanja agende.

Istraživanje na koji način vlada utječe na medijsku agendu opisano u [44] korelira podatke o temama dobivene računalnom analizom medijske agende sa službenim objavama vlade SAD-a o podizanju razine terorističke prijetnje. Koristi se zbirka od 51.766 novinskih tekstova i transkripata televizijskih vijesti neposredno prije i nakon događaja podizanja razine prijetnje koji su se odvili u razdoblju od 3 godine. Za analizu agende koriste se proširenje modela LDA koje eksplicitno modelira “pažnju” koju pojedini medij pridaje temi. Izgrađen je model sa 24 teme, broj odabran variranjem broja tema i promatranjem odgovarajućih modela. Semantička validacija tema provodi se promatranjem i interpretacijom tema a indirektna validacija tema i njihovih zastupljenosti postiže kroz analizu rezultata koji se pokazuju smislenima – podizanje

[4] https://www.theguardian.com

razine prijetnje pozitivno utječe na pažnju koju mediji pridaju s terorizmom vezanim temama.

U [56] se opisuje studija slučaja analize novinskih tekstova o nuklearnoj tehnologiji. Zbirka tekstova sastoji se od 51.528 članaka iz New York Times baze dobivenih filtriranjem članaka objavljenih između 1945. i 2013. prema pojavljivanju ključnih riječi vezanih uz nuklearnu tehnologiju. Tematsko modeliranje provodi se modelima LDA sa 10 tema (prema broju kategorija korištenih u ranijoj studiji) i sa 25 tema (broj odabran automatski prema kriteriju zbunjenosti). Teme su semantički validirane pregledom i interpretacijom, te pomoću vizualizacijskog alata koji omogućava analizu koherentnosti tema na temelju supojavljivanja uz teme vezanih riječi. Dodatna validacija modela koja je potvrdila smislenost tema provedena je korelacijom zastupljenosti tema s relevantnim događajima. Zaključuje se da je model LDA dobar alat za eksplorativnu analizu velikih zbirki tekstova no da se nije se pokazao dobrim za analizu uokvirivanja zbog lošeg poklapanja tema s okvirima.

Analizu uokvirivanja političkih pitanja vezanih uz zabranu pušenja, s naglaskom na geografski i politički kontekst njihovog širenja kroz savezne države SAD-a, opisana je u [49]. Istraživanje je provedeno na zbirci tekstova dobivenih iz 49 novina reprezentativnih za savezne države. Zbirka je dobivena filtriranjem ključnim riječima i metodama nadziranog učenja te se sastoji od 52.675 paragrafa vezanih uz zabranu pušenja. Korišten je strukturni tematski model STM (engl. *structural topic model*) [57] sa 12 tema (broj odabran promatranjem modela s različitim brojem tema i mjerenjem koherentnosti tema). Teme su validirane pregledom i interpretacijom te analizom korelacije zastupljenosti tema s događajima i drugim varijablama poput postotka pušača u saveznoj državi. Korelacijska validacija pokazala je smislenost tema.

Istraživanje medijskog uokvirivanja javnog financiranja umjetnosti u SAD-u u kontekstu politizacije i problematizacije tog pitanja opisano je u [48]. Korištena je zbirka tekstova iz 5 nacionalnih novina objavljenih između 1986. i 1997. godine. Konačnih 8.000 tekstova vezanih uz istraživanu problematiku dobiveno je filtriranjem po ključnim riječima i metapodacima. Koristi se model LDA s 12 tema. Provedena je validacija promatranjem i interpretacijom tema te dodatna semantička validacija koja na temelju ljudskih ocjena ispituje imaju li iste riječi imaju različita značenja kada ih model svrstava u različite tema. Model i teme su dodatno validirane promatranjem korelacije s događajima. Autori izvještavaju o dobrom poklapanju između tema modela i okvira.

U [58] opisano je istraživanje iz područja računalne sigurnosti koje provodi analizu medijske agende kako bi se otkrilo koje aspekte računalne sigurnosti pokrivaju vijesti i kako bi se napravila usporedba s aspektima o kojima se piše na specijaliziranim web stranicama i o kojima krajnji korisnici međusobno raspravljaju. Zbirka tekstova sastoji se od 1072 članka iz 16 velikih novina iz SAD-a koji su dobiveni filtriranjem prema ključnim riječima povezanima s računalnom sigurnošću, te od 500 tekstova s web stranica i 300 anketom dobivenih tekstova krajnjih korisnika. Agenda se otkriva modelom LDA s deset tema, broj koji je odabran promatranjem

više modela s različitim brojevima tema kako bi teme bile konceptualne a ne specifične. Provedena je validacija promatranjem i interpretacijom tema. Teme su detaljno opisane i provedena je analiza supojavljivanja tema i njihove zastupljenosti u svakom od opisanih izvora tekstova.

Metodološko istraživanje primjene tematskih modela za analizu novinskog teksta opisano u [50] provedeno je kroz analizu načina na koji velike novine razgraničavaju znanost i pseudoznanost. Koristi se zbirka od 15.000 tekstova objavljenih u 30 velikih novina iz SAD-a između 1980. i 2012. godine, dobivena filtriranjem po ključnim riječima vezanim uz temu istraživanja. Koristi se model LDA pri čemu se razmatra 9 modela-kandidata s brojem tema između 15 i 100 tema. Kandidati se validiraju tijekom tri faze, prvo pregledom tema prikazanih riječima, nakon čega se provodi analiza tema na temelju mjerenja koherentnosti te razlika između tema i razdiobe riječi na razini zbirke. Zatim se u trećoj fazi provodi validacija ocjenjivanjem smislenosti supojavljivanja tema te korelacijom zastupljenosti tema i vezanih događaja. Nakon odabira modela sa 45 tema provodi se filtriranje nevažnih i loših tema, nakon čega slijedi analiza usredotočena na prirodu odabranih tema (detaljna interpretacija) i kretanje njihove zastupljenosti kroz vrijeme. Zaključuje se da teme modela odgovaraju temama (engl. *"subjects of discussion"*) koje su razmatrane u prethodnim srodnim studijama te da su otkrivene nove teorijski relevantne teme.

U [59] opisuje se metodološko istraživanje primjene tematskih modela u komunikacijskim znanostima na primjeru analize medijskog pokrivanja građanskog rata u Siriji. Analiza se provodi na zbirci od 2.083 teksta iz novina The Guardian objavljenih između 2011. i 2015. godine. Tekstovi su dobiveni filtriranjem po ključnim riječima i metapodacima s ciljem izdvajanja tekstova na temu rata u Siriji. Modeliranje se provodi modelom LDA sa 8 tema a validacija je provedena pregledom i interpretacijom tema te analizom strukture sličnosti tema dobivene pomoću metode hijerarhijskog grupiranja. Provodi se i validacija analizom korelacija zastupljenosti tema s bitnim događajima unutar proučavanog vremenskog razdoblja. Autori kao glavnu prednost tematskog modeliranja ne ističu mogućnost kvalitativne analize otkrivenih tema nego analize koja omogućavaju iz modela dobiveni podaci o zastupljenosti tema u dokumentima, poput analize supojavljivanja tema i analize skupova tematski srodnih dokumenata.

U [51] opisano je istraživanje povijesnog novinskog teksta temeljeno na tematskim modelima i provedeno na zbirci digitaliziranih novina iz savezne države Texas u SAD-u izdanih između 1829. i 2008. godine. Zbirka sadrži 32.745 izdanja s ukupno 232.567 stranica koje se koriste kao tekstovi pri modeliranju. Autori koriste model LDA sa 10 tema (manji broj odabran kako bi se dobile općenitije teme i grade modele na cijeloj zbirci tekstova te na podskupovima zbirke koji odgovaraju pojedinim vremenskim razdobljima ili sadrže tematski specifične ključne riječi. Provodi se validacija tema pregledom i interpretacijom od strane povjesničara. Autori zaključuju da je modeliranje rezultiralo "visoko korisnim" rezultatima te da se teme poklapaju s očekivanjima povjesničara (naglasak je na ekonomske teme). Također ističu nužnost

interpretacije tema od strane domenskih stručnjaka.

U [60] predlaže se računalni pristup analizi postavljanja agende temeljen na tematskom modeliranju novinskih tekstova i korisničkih komentara. Modeliranje se povodi na zbirci od 17.674 novinska članka i 763.721 komentara sakupljenih s web sjedišta Nacionalne radijske postaje SAD-a[1] u razdoblju od 2011. do 2013. godine. Iz podataka se uči HDP model [61], neparametarski tematski model koji sam procjenjuje broj tema. Nakon pregleda i interpretacije tema modela izdvojena je skupina tema za koje se zatim provodi analiza postavljanja agende usporedbom zastupljenosti tema u medijskim tekstovima (medijska agenda) i korisničkim komentarima (javna agenda) te se izdvajaju primjeri tema s visokom i niskom korelacijom medijske i javne agende. Predlažu se i analize korelacije zastupljenosti tema u korisničkim komentarima i na društvenim mrežama s jedne te automatski izmjerenim sentimentom tekstova s druge strane.

Račnalni pristup analizi medijske agende, uz oblikovanje odgovarajuće programske potpore i rudimentarnu evaluaciju metode, opisuje se u [62] i predstavlja prvo istraživanje koje se bavi računalnom analizom medijske agende. Metoda se temelji na grupiranju tekstova, pri čemu alat pruža izbor između nekoliko različitih metoda grupiranja. Grupe tekstova smatraju se otkrivenim temama a alat omogućava pregled grupa promatranjem vezanih riječi i dokumenata te vizualizaciju međusobne povezanosti grupa i vremenske distribucije dokumenata unutar grupe. Evaluacija metode provodi se usporedbom s rezultatima društvenoznanstvene studije postavljanja agende tijekom koje su novinski članci od strane ljudi kodirani u tematske kategorije. Svaka od alatom dobivenih grupa prvo se ručno mapira na jednu od tih kategorija a zatim se radi usporedba ukupne zastupljenosti kategorija (mjereno brojem tekstova) dobivene studijom i one dobivene automatskom metodom. Razlika od ostalih metoda je pristup korišten u postupku validacije u kojem se računalnim modelom dobivene teme mapiraju na tematske kategorije razvijene od strane domenskih stručnjaka.

U [54] se opisuje upotreba tematskih modela za eksplorativnu analizu velike zbirke medijskih tekstova dobivene pomoću Media Cloud platforme koja automatski sakuplja tekstove iz velikog broja medija. Zbirka se sastoji od 429.042 teksta sakupljena tijekom 3 mjeseca 2014. godine praćenjem 25 najpopularnijih web portala iz SAD-a. Za tematsko modeliranje koristi se model LDA sa 100 tema. Odabrane su 42 teme ocijenjene koherentnima na temelju promatranja vezanih riječi. Autori predlažu nekoliko vizualizacijskih alata temeljenih na tematskoj karakterizaciji web portala (izvora vijesti) – vizualizacija korespondencije tema i izvora vijesti, grupiranje izvora po "tematskom potpisu", vizualizacija tematskog potpisa pojedinog izvora te dodavanje podataka o vezanim izvorima pri promatranju samih tema. Predlaže se niz potencijalnih primjena za novinare i konzumente vijesti te se ističe potreba za validacijom tema. Autori predlažu vizualizacijske alate za validaciju temeljene na međusobnoj korelaciji tema te korelaciji tema s vanjskim podacima poput događaja i spominjanja osoba u tekstu.

[1] https://www.npr.org

Tematski model za istraživanje uokvirivanja predložen je u [63]. Model se temelji na NMF modelu te modelira okvire i njihovu vremensku komponentu. Primjena modela demonstrirana je na zbirci od 13.000 članaka iz razdoblja od 1997. do 2011. godine dohvaćenih ih LexisNexis baze na temelju ključnih riječi vezanih uz ilegalnu imigraciju. Opisuje se nekoliko okvira dobivenih modelom te se vizualizira kretanje njihove zastupljenosti u vremenu. Treba napomenuti da nije napravljena usporedba sa standardno korištenim tematskim modelima kako bi se pokazale relativne prednosti predloženog modela.

U navedenim istraživanjima postupak mjerenja zastupljenosti tema provodi se pomoću težina tekstova za teme koje se očitavaju iz tematskog modela. Pritom se kriterij pridruživanja teme tekstu na temelju težine ponekad niti ne spominje, no najčešće se radi o odabiru teme s najvećom težinom ili odabiru teme čija težina je iznad nekog praga. Pridruživanje tema dokumentima je instanca problema klasifikacije pri čemu se tekstovi klasificiraju u klase koje odgovaraju temama, no niti u jednom istraživanje ne provodi se kvantitativna validacija. Iako korelacije zastupljenosti tema s događajima i drugim varijablama te kvalitativne analize rezultata daju smislene rezultate, bilo bi metodološki ispravno provesti validaciju – društvenoznanstvena istraživanja koja koriste nadzirane modele za klasifikaciju tekstova standardno provode validaciju točnosti tih mjerenja [38, 41].

Sve opisane analize medijske agende ograničene su na učenje samo jednog modela. No proces učenja modela je stohastičan i postoje istraživanja koja ukazuju na visoku varijaciju tema naučenih jednim modelom [64, 65]. Ovo ograničenje ima za posljedicu mogućnost da jedan model ne detektira teme od interesa ili da detektira teme koje se tek aproksimativno poklapaju s koncepatima koje analitičar želi koristiti pri analizi.

Vrlo čest postupak pri izgradnji zbirke tekstova je filtriranje veće zbirke prema određenim ključnim riječima kako bi se dobili tekstovi iz neke uže tematske domene. Iz tog razloga bi programska potpora za računalnu analizu medijske agende trebala podržavati ovu operaciju.

3.2 Vrednovanje tematskih modela u kontekstu analize medijske agende

Struktura prirodnog jezika vrlo je složena i računalni modeli teksta, uključujući i tematske modele, samo su gruba aproksimacija stvarnosti. Stoga za niti jedan model nije moguće apriorno tvrditi da je bolji od drugog već je modele potrebno validirati kako bi se demonstrirala njihova korisnost za određeni zadatak [41]. U kontekstu analize medijske agende metode validacije tematskih modela su alat važan i za otkrivanja tema i za mjerenje zastupljenosti tema – poželjno je da modeli nauče interpretabilne i relevantne teme koje dobro aproksimiraju tematsku strukturu tekstova, te da naučene težine tema za tekstove odgovaraju temama koji se u tekstovima javljaju.

Kvaliteta tematskih modela poželjna je neovisno o primjeni, no posebno je važna kod primjena modela u društvenoznanstvenim istraživanjima jer se tada na naučenim modelima temelje znanstveni zaključci. Problem validacije tematskih modela u primjenama na znanstvenu analizu tekstnog sadržaja je otvoren problem i ne postoje standardno prihvaćane metode validacije [40, 66, 67].

U društvenoznanstvenim primjenama, standardni pristupi validaciji tematskih modela modela su semantička validacija i prediktivna validacija [39, 48, 68]. Semantička validnost tema odnosi se primarno na interpretabilnost tema odnosno dobro poklapanje teme s nekim konceptom. Prediktivna validnost odnosi se na svojstvo teme da na očekivani način korelira sa vezanim događajima, primjerice očekivano je da zastupljenost teme "kriza" znatno poraste u razdoblju nakon što je kriza nastupila.

Neovisno o primjeni, za teme modela uvijek je poželjno da budu semantički interpretabilne i koherentne, no to nije zajamčeno zbog stohastičke i aproksimativne prirode tematskih modela. Primjeri nekvalitetnih tema su teme koje sadrže veću ili manju količinu šuma (slučajnih ili nevezanih riječi i dokumenata) te teme koje spajaju dva različita koncepta [69]. Takve nekvalitetne teme su od ograničene ili nikakve koristi za analitičara pošto je njihova interpretacija teža ili nemoguća, te očekivano vremenski zahtjevnija od interpretacije kvalitetnih tema. Mjere koherentnosti tema opisane u 2.2 i 4.3 razvijene su upravo kao pokušaj automatskog ocjenjivanja ovog aspekta kvalitete tema.

Glavna metoda semantičke validacije tema je njihov pregled i interpretacija na temelju vezanih riječi i dokumenata. Ova metoda može najbolje uvjeriti analitičara da su teme smislene te se uvijek provodi, makar implicitno, pošto je neizbježan dio analize agende otkrivanje tema putem pregleda i interpretacije. Ostale metode uključuju primjenu mjera koherentosti za odabir broja tema u modelu [49, 50] i vizualizacije koje prikazuju neki aspekt koherentnosti tema [56]. Predložena je i procjena koherentnosti tema modela na temelju ljudskih ocjena semantičke sličnosti parova dokumenata [41]. Pri tome se mjera kvalitete modela računa kao razlika između prosječne sličnosti dokumenata vezanih uz istu temu i prosječne sličnosti dokumenata vezanih uz različite teme.

Osim mjerenja inherentne semantičke kvalitete tema, semantička validnost tema može se procijeniti i analizom međusobne sličnosti tema. U [39, 59] teme se grupiraju (na temelju vezanih riječi ili dokumenata) i promatra se da li dobivena struktura sličnosti tema odgovara semantičkoj sličnosti koncepata koji odgovaraju temama. U [48] semantička validnost procjenjuje se, provjerom od strane ljudi, da li iste riječi poprimaju različita značenja kada ih model pridruži različitim temama. Poopćenje ove metode bila bi semantička validacija tema raznim "vanjskim" semantičkim vrednovanjima pri čemu bi bilo poželjno da se ta vrednovanja mogu provesti automatski.

Prediktivna validacija tema provodi se ocjenjivanjem korelacije tema sa stvarnim događa-

jima – promatra se da li promjene u zastupljenosti tema kroz vrijeme koreliraju sa uz temu povezanim događajima [39, 48, 68]. Osim s događajima, validacija se može provesti i promatranjem korelacija tema s drugim varijablama poput demografskih podataka [39, 49].

Relevantnost Za razliku od metoda nadziranog učenja, za nenadzirane tematske modele nije moguće precizno definirati varijable koja se mjere pošto su ti modeli alat za otkrivanje prethodno nepoznatih tema odnosno koncepata. Stoga je jedino pregledom naučenih tema od strane stručnjaka moguće prosuditi njihovu korisnost za određenu analizu. Relevantnost se odnosi na svojstvo tema modela da odgovaraju korisnim pojmovima koji pružaju uvid u tekstove i takve teme su u literaturi nazvane "teorijski zanimljivima" (engl. *theoretically interesting*) [41] te "analitički korisnima" (engl. *analytically useful*) [48]. Kriterij relevantnosti razlikuje se ovisno o primjeni – orijentacijska pregledna analiza zbirke tekstova od strane neupućenog korisnika zahtijeva samo dovoljno dobru aproksimaciju tema u tekstovima dok su za znanstvene analize zbirke tekstova potrebne specifične "teorijski zanimljive" i "analitički korisne" teme.

Relevantnost tematskih modela za određenu primjenu ocjenjuju se promatranjem i interpretacijom tih tema, pri čemu analitičar interpretira teme kao koncepte te procjenjuje njihovu korisnost u kontekstu analize koja se provodi.

3.3 Računalni postupak za analizu medijske agende

U ovom odjeljku opisan je prijedlog metode za analizu medijske agende koja se nadograđuje na postojeće pristupe, sistematizira ih te predlaže rješenja za njihove uočene nedostatke.

Tri su nedostatka postojećih pristupa – problem pokrivenosti tema na agendi, nepostojanje jamstva pojavljivanja tema od interesa za analitičara, te nedostatak vrednovanja mjerenja zastupljenosti tema. Prvi problem odnosi se na činjenicu da jedan tematski model očekivano ne otkriva sve teme u zbirci tekstova, na što ukazuju postojeći eksperimenti u kojima se mjeri preklapanje skupova tema [29, 64]. Predloženo rješenje je otkrivanje medijske agende pomoću većeg broja tematskih modela.

Drugi problem posljedica je toga što se analiza tema u postojećim pristupima ograničava na teme koje nauči jedan tematski model. No teme jednog modela s jedne strane ne moraju nužno pokriti sve postojeće teme, a druge strane mogu radi stohastičnosti procesa učenja semantički varirati odnosno prikazivati srodne ali različite koncepte [64, 65]. Pošto su kod znanstvenih primjena potrebne teme koje precizno odgovaraju konceptima [41, 48] predloženi pristup predviđa preciznu definiciju novih tema na temelju okrivenih tema. Mjerenje zastupljenosti novodefiniranih tema zahtijeva izgradnju novih modela za mjerenje, za što se predlaže pristup temeljen na slabo nadziranim tematskim modelima.

Treći problem odnosi se na mjerenje zastupljenosti tema koje se standardno provodi na te-

melju pridruživanja mjerenih tema dokumentima pomoću iz modela očitanih težina tema za dokumente. Pri tome izostaje precizno vrednovanje pridruživanja što se u predloženom pristupu rješava eksplicitnom formulacijom pridruživanja kao problema klasifikacije s višestrukim oznakama (engl. *multilabel classification*) koja omogućava kvantitativno vrednovanje.

Predložena metoda analize medijske agende sastoji se od tri konceptualno različita koraka koji prirodno proizlaze iz prirode procesa analize medijske agende od strane ljudskog analitičara. Prvi korak je *otkrivanje tema* koje čine medijsku agendu temeljen na tematskim modelima. Korak se provodi pregledom i interpretacijom tema modela prikazanih pomoću vezanih riječi i tekstova. Drugi korak, usko povezan s prvim, je *definicija tema* od interesa, tijekom kojeg analitičar na temelju prethodno otkrivenih tema po potrebi definira nove teme u svrhu preciznije analize. Treći korak je *mjerenje zastupljenosti tema* definiranih u prvom i drugom koraku, postavljeno kao zadatak klasifikacije s višestrukim oznakama (engl. *multilabel classification*) – dokumenti se klasificiraju po temama, odnosno svakom dokumentu se pridružuje jedna ili više tema. Zatim se zastupljenost tema mjeri brojem pridruženih dokumenata.

Predloženi postupak analize medijske agende interaktivan je, odnosno uključuje ljudskog analitičara, u svim koracima koji se rješavaju podrazumijevaju analizu i interpretaciju tekstova i njihovog konteksta. Drugim riječima, metode strojnog učenja nisu zamjena već alati za analitičara [41]. Postupak je opisan općenito, kao radni tijek interakcije čovjeka s modelima strojnog učenja, a sama implementacija može se izvesti na različite načine – od interaktivnog korisničkog sučelja do "ručnog" izvođenja pisanjem i izvršavanjem koda te promatranjem izlaza.

U nastavku su opisani motivacija i provedba svakog od tri koraka, nakon čega slijede dvije studije slučaja analize medijske agende provedene predloženom metodom te rezultati i iskustva proizišla iz tih istraživanja. Na kraju poglavlja opisuju se smjerovi daljnjeg istraživanja motivirani rezultatima studija te se provodi rasprava o samoj metodi.

3.3.1 Otkrivanje tema

Svrha koraka otkrivanja tema je uvid u tematsku strukturu zbirke vijesti. Ovaj korak se u klasičnim analizama tekstnog sadržaja provodi čitanjem tekstova što je vremenski zahtjevan proces nepraktičan ili neizvediv za velike tekstne zbirke. Predloženi pristup temelji se na tematskim modelima koji su prihvaćeni alat za ovakvu vrstu ekplorativne analize i za koje se u nizu primjena na analizu medijske agende, opisanih u 3.1.3, pokazalo da mogu otkriti teme od interesa za istraživača. Otkrivanje tema provodi se učenjem tematskog modela te pregledom i interpretacijom tema prikazanih listom vezanih riječi i listom vezanih dokumenata.

U kontekstu tematskih modela i posebno interpretacije njihovih tema, bitno je naglasiti razliku između *teme modela* – konstrukta tematskog modela koji odgovara utežanoj listi riječi i tekstova, te *semantičke teme* – koncepta koji koji je konstrukt ljudskog uma nastao intrepretacijom teme modela.

Postojeći pristupi otkrivanju tema na medijskoj agendi koriste jedan tematski model. No postoje eksperimenti koji ukazuju na to da samo jedan tematski model ne može pokriti sve semantičke teme u zbirci tekstova. U [29] analizira se preklapanje tema modela sa skupom koncepata konstruiranih od strane domenskih stručanjaka i predlaže metoda vizualne analize koja pokazuje da modeli očekivano otkrivaju samo manji broj koncepata. Drugo istraživanje [64] grupira bliske teme većeg broja modela i pokazuje da pojedini modeli često ne pokrivaju sve grupe, posebice one koje odgovaraju manje zastupljenim temama. U kontekstu otkrivanja tema ovo može rezultirati neotkrivanjem analitičaru potencijalno zanimljivih tema. Stoga se predloženi pristup otkrivanju agende temelji na izgradnji i promatranju tema većeg broja modela. Veći broj modela očekivano otkriva više tema, a osim toga može pružiti i bolji uvid u zbirku tekstova zbog toga što semantičke varijacije tema modela [65] mogu ukazati na raznolike analitičaru zanimljive interpretacije [40].

Cijena koja se plaća je veći utrošak vremena potreban za pregled većeg broja tema. Ostale razlike od postojećih pristupa su tablica semantičkih tema (otkrivenih tematskih koncepata) koja služi za lakšu evidenciju pri obradi većeg broja modela i usklađivanje većeg broja anototora, te neodbacivanje tema modela koje se poklapaju s dvije semantičke – takve teme se u pravilu smatraju šumom no njihova interpretacija može dovesti do otkrivanja većeg broja tema.

Izgradnja tematskih modela Temeljni korak za provedbu otkrivanja tema je izgradnja tematskih modela i sastoji se od pretprocesiranja zbirke tekstova, odabira hiperperametara modela i algoritma učenja, te konstrukcije odnosno učenja modela i njihove pohrane za daljnje korištenje.

Predloženi pristup se ograničava na jedan tip tematskog modela, nego se modeli promatraju općenito, kao objekti koji iz tekstova uče teme – liste riječi i tekstova koje je moguće interpretirati kao semantičke teme. Ovakav pristup je smislen zbog velikog broja tipova modela (koji će se očekivano povećavati) te strukturne složenosti teksta zbog koje ne postoji apriorno najbolji model za pojedini zadatak [41]. Stoga je razumno omogućiti korisniku da odabere model ovisno o primjeni, vlastitim iskustvima i drugim faktorima poput programske podrške za pojedini model. Unatoč opisanom pristupu, kao početna točka za provedbu metode daju se preporuke za odabir i izgradnju modela izvedene iz postojećih eksperimenata s modeliranjem medijske agende.

LDA tematski model [1] je popularan u istraživanjima medijske agende i drugim primjenama i predstavlja dobar početni izbor za tip modela. Prednost modela LDA je postojanje velikog broja implementacija u mnogim programskim jezicima, često kao dio knjižnice za strojno učenje ili tematsko modeliranje. Hiperparametri modela LDA često se postavljaju prema preporukama iz [7].

Broj tema modela je važan hiperparametar velikog broja tematskih modela i ima veliki utjecaj na strukturu naučenih tema. Jedan mogući pristup odabiru broja tema je njegovo variranje

i odabir na temelju neke mjere kvalitete modela ili pregleda tema modela. Međutim, postojeće primjene tematskih modela na analizu medijske agende pokazuju da modeli s manjim brojem tema uče općenitije teme (apstraktne koncepte) dok modeli s većim brojem tema uče specifičnije teme (događaji, osobe, organizacije) . Stoga se može reći da ne postoji optimalni broj tema nego broj tema određuje tip modela. Pri tome sam broj tema koji odgovara pojedinoj razini općenitosti ovisi o veličini zbirke tekstova.

Sljedeće smjernice za odabir broja tema izvedene su iz postojećih istraživanja medijske agende. Za zbirke novinskih tekstova od nekoliko desetaka tisuća dokumenata modeli sa 50 i manje tema uče apstraktne, dok modeli sa 100-njak tema uče specifične teme [50]. Za manje zbirke od 5 do 10 tisuća tekstova razumno je odabrati 10-tak tema za apstraktne i 50-tak za specifične teme . Za velike zbirke od više stotina tisuća tekstova 100-njak tema daje apstraktne teme [54] dok je razumno očekivati da bi oko 1000 tema trebalo dati specifične teme.

Pri odabiru broja tema, intuicija analitičara o očekivanom broju tema u zbirci svakako je bitna odrednica. Ovisno o zbirci tekstova, primjeni, i potrebnoj kvaliteti tema eksperimentiranje s brojem tema može biti neizbježno. Pri tome se sam postupak pregleda i interpretacije tema može započeti s razumnom početnom vrijednosti, i u slučaju da teme nisu dobre, mogu se ponoviti izgradnja i promatranjem modela s različitim brojem tema. Razumna heuristika u tom slučaju može biti eksponencijalno povećanje ili smanjenje broja tema za faktor 2, počevši od početnog broja tema, pri čemu je smanjenje ili povećanje određeno potrebom za općenitijim ili specifičnijim temama.

Dvije značajke predloženog postupka analize agende olakšavaju i skraćuju postupak odabira broja tema u odnosu na dosadašnje pristupe. Prvo, zbog uvedenog koraka definicije tema, teme modela za otkrivanje agende ne moraju se precizno poklapati sa konceptima od interesa za analitičara već je dovoljno da ga usmjere prema dobrim definicijama tih koncepata koji će se zatim moći precizno analizirati na temelju postupka mjerenja. Drugo, pošto sam korak otkrivanja tema predviđa pregled većeg broja tematskih modela moguće je izgraditi i pregledati više modela s različitm brojem tema, primjerice jedan model koju uči apstraktne i jedan koji uči specifične teme.

Razumna preporuka za odabir skupa modela za otkrivanje agende je odabir 2 modela sa manjim brojem tema i 2 modela s većim brojem tema, pri čemu je “manje” i “veće” definirano prethodnim razmatranjima. Važno je napomenuti da je različite modele potrebno izgraditi korištenjem različitih vrijednosti inicijalizacije slučajnih generatora (engl. *random seed*) koji se koriste za inicijalizaciju i učenje modela, u suprotnome će modeli sadržavati identične podatke. Odabir broja modela ograničen je i vremenom koje je analitičarima dostupno za pregled tema a koje ovisi o ukupno dostupnom broju čovjek-sati te vremenu potrebnom za pregled i interpretaciju jedne teme.

Koraci postupka otkrivanja tema Nakon izgradnje tematskih modela, pristupa se postupku otkrivanja tema koji se provodi pregledom i interpretacijom naučenih tema modela od strane ljudskih označivača. Interpretacija tema rezultira semantičkim temama koje se dodaju u referentnu tablicu koja služi za evidenciju o temama te usklađivanje među označivačima. Nakon obrade svih tema modela, opcionalno se provodi korak revizije označenih podataka s ciljem otklanjanja grešaka, te korak izgradnje taksonomije za otkrivene semantičke teme.

Pregled i interpretacija tema modela provode se na temelju uz temu vezanih dokumenata i riječi što je standardni postupak u eksplorativnoj analizi temeljenoj na tematskim modelima. Prikazani dokumenti i riječi trebaju ljudskom označivaču s jedne strane pružiti dovoljno informacija za interpretaciju a s druge strane biti dovoljno sažeti da postupak pregleda teme ne zahtijeva previše vremena. Riječi se prikazuju kao lista do 10 riječi uređenih padajuće prema težinama riječi za temu. Dokumenti se prikazuju kao lista dokumenata uređenih padajuće prema težinama teme za dokument. Temeljem preliminarne analize zaključeno je da je umjesto ograničavanja broja dokumenata bolje prikazati veliki broj dokumenata i informaciju o pripadnoj težini teme za dokument te dopustiti označivaču da procijeni koliko dokumenata s vrha liste je relevantno za temu, uz uputu da nije potrebno pregledati više od 50-tak dokumenata. Ovim pristupom pokriveni su i zastupljenije tema s velikim brojem dokumenata (50 i više) i one manje zastupljene koje se javljaju u manjem broju dokumenata (manje od 20). Pri pregledu dokumenata dokumente je potrebno sažeto prikazati zbog brzine pregleda. U slučaju novinskih članaka to nije problem pošto za članke već postoje informativni i sažeti naslovi. Označivaču je omogućen i pregled cjelovitog teksta koji može biti nužan za interprataciju teme.

Na temelju pregleda teme modela označivač interpretira temu kao *semantičku temu* – koncept koji dobro opisuje dokumente i riječi teme modela. Pri tome se može dogoditi da riječi i dokumenti pripadaju dvama različitim semantičkim temama ili da sadrže određenu količinu šuma (slučajnih riječi i dokumenata) ali su i dalje interpretabilni. Ti slučajevi odgovaraju standardnim greškama tematskih modela [69]. Umjesto odbacivanja ovih tema kao grešaka modela, predloženi postupak predviđa bilježenje svih semantičkih tema te dodatno označavanje teme s labelom za šum. Na ovaj način može se postići otkrivanje većeg broja semantičkih tema. Nakon interpretacije, tema modela se označava s oznakom jedne ili većeg broja semantičkih tema te moguće i oznakom za šum. Ukoliko je tema neinterpretabilna oznaka za šum je jedina oznaka teme.

Pregled i interpretacija tema ujedno predstavljaju i najvažniju metodu validacije tematskih modela – jedino ovim postupkom analitičar se može u potpunosti uvjeriti u kvalitetu tema i njihovu relevantnost za istraživanje. Ukoliko pregled tema ukazuje na nisku kvalitetu (neinterpretabilnost) ili irelevantnost semantičkih tema za analizu (npr. teme su preopćenite), potrebno se vratiti na korak izgradnje tematskih modela.

Postupak pregleda i interpretacije tema modela, odnosno otkrivanja semantičkih tema vođen

je referentnom tablicom semantičkih tema koja je vidljiva svim označivačima. Nakon pregleda teme modela, označivač prvo konzultira tablicu te pokušava pronaći već otkrivenu semantičku temu koja odgovara temi modela. Ukoliko takva tema nije pronađena, potrebno je iznaći oznaku i definiciju nove semantičke teme i upisati ih u tablicu zajedno s identifikatorom teme modela. Ukoliko je odgovarajuća semantička tema pronađena, samo joj se pridružuje identifikator nove teme modela. Korištenjem referentne tablice semantičkih tema nastoji se postići usklađivanje anotatora – spriječiti da iste semantičke teme budu različito imenovane te olakšati imenovanje i interpretaciju prethodno otkrivenih semantičkih tema. Tablica ujedno omogućava lakši pregled svih semantičkih tema što je važno pošto se pregledava veći broj tematskih modela.

Prije početka označavanja, potrebno je uputiti sve označivače u postupak pregleda i interpretacije, te je poželjno provesti kalibracijski postupak označavanja na manjem uzorku tema kako bi se označivačima demonstrirao postupak i kako bi se razjasnile nejasnoće i uskladio postupak interpretacije tema. Pri klasičnoj analizi sadržaji standardno se provodi analogni postupak treniranja i kalibracije označivača kako bi mogli kvalitetno i dosljedno provoditi kodiranje tekstova u definirane analitičke kategorije [38].

Nakon obrade svih tema modela, može se provesti dodatni korak revizije i ispravljanja dobivenih podataka – definicija semantičkih tema te njihovih veza s temama modela. Također, semantičke teme se moguće je organizirati u taksonomiju što dovodi do veće preglednosti te, očekivano, novih uvida u strukturu tema i veza među njima. Sama taksonomija i njena razrađenost ovisi o analizi koja se provodi, a kategorije taksonomije mogu biti ili same semantičke teme (koje obuhvaćaju druge manje općenite teme) ili nove kategorije dobivene poopćenjem tema. Ovaj postupak može biti koristan i za dolazak do definicija novih semantičkih tema od interesa za analizu.

3.3.2 Definicija tema

Korak definicije tema provodi se na temelju semantičkih tema iz prethodnog koraka koje očekivano daju dobar pregled medijske agende. Svrha ovog koraka je definicija semantičkih tema od interesa za istraživača, na temelju kojih se zatim provodi daljnja analiza agende. Analiza se može provesti primjerice mjerenjem zastupljenosti tih tema te korelacijom zastupljenosti s vanjskim varijablama od interesa, ili analizom supojavljivanja tema. Definirane semantičke teme mogu imati i eksplorativnu svrhu – postupkom izgradnje prilagođenih tematskih modela opisanim u koraku mjerenja dobivaju se modeli čije teme odgovaraju definiranim semantičkim temama i te je modele moguće koristiti za eksplorativnu analizu koja može ukazati na nove teme ili pružiti bolje razumijevanje definiranih tema. Novodefinirane semantičke teme često su od interesa u znanstvenim analizama medijske agende no za primjene poput pregledne eksplorativne analize ovaj korak nije potreban.

Mogućnost definicije novih semantičkih tema predstavlja važnu razliku u odnosu na posto-

jeće pristupe koji ograničavaju analizu agende na teme naučene jednim tematskim modelom i u kojima se umjesto definicije provodi samo odabir otkrivenih tema korisnih za analizu. Međutim, postojeći eksperimenti s tematskim modelima ukazuju na to da zbog stohastičke prirode tematskih modela otkrivene teme mogu znantno varirati. Pojedine teme uopće ne moraju biti pokrivene od strane pojedinog modela [29, 64], te je uobičajeno da se javljaju teme koje mogu otkriti srodne no različite semantičke aspekte [64, 65]. Posljedica opisanih svojstava modela je da ne postoji jamstvo da će se teme modela poklapati s konceptima od interesa za analitičara. Primjerice, u dva srodna istraživanja medijske agende koja nastoje primijeniti tematske modele na istraživanja uokvirivanja, autori prvog istraživanja izvještavaju da teme modela ne odgovaraju konceptima okvira [56], dok su iskustva autora drugog istraživanja upravo suprotna [48]. Predloženi korak definicije tema trebao bi otkloniti ove nedostatke i učiniti tematske modele pouzdanijim alatom.

U koraku definicije tema analitičar na temelju prethodno otkrivenih semantičkih tema – njihovih oznaka, definicija te analize vezanih tema modela odnosno pripadnih riječi i tekstova – izvodi definicije novih semantičkih tema koje predstavljaju za analizu korisne koncepte. Semantičke teme otkrivene u prethodnom koraku koje već dobro odgovaraju konceptima od interesa jednostavno se uključuju na listu tema koje će se koristiti za daljnju analizu. Iako konceptualno različit, ovaj korak u provedbi može biti proveden paralelno s korakom otkrivanja, odnosno nove definicije mogu se izvoditi tijekom samog postupka otkrivanja. Korak osim samog procesa definicije podrazumijeva i vođenje evidencije o novodefiniranim temama – izradu liste tema s njihovim oznakama, definicijama te vezama s prethodno otkrivenim semantičkim temama i temama modela.

3.3.3 Mjerenje tema

U koraku *mjerenja zastupljenosti tema* provodi se mjerenje zastupljenost tema medijske agende od interesa za analizu koje su izlaz prethodnih koraka otkrivanja i definicije tema. Ovaj korak se provodi ukoliko je mjerenje zastupljenosti tema potrebno za analizu agende, primjerice u istraživanjima postavljanja agende u kojima se analizira korelacija zastupljenosti tema na medijskoj agendi i javne percepcije važnosti tema. *Zastupljenost tema* definira se kao broj novinskih članaka koji govore o toj temi a *mjerenje zastupljenosti* provodi se označavanjem članaka s pripadnim temama – pristup koji se često koristi i u računalnim analizama medijske agende i u klasičnim studijama u kojima se ljudskim kodiranjem dokumenti svrstavaju u tematske kategorije. Predlaže se pristup mjerenju u kojem se za svaku mjerenu temu izrađuje lista visokodiskriminativnih riječi na temelju kojih se zatim uči tematski model čije teme odgovaraju mjerenim temama što se postiže korištenjem apriornih vjerojatnosnih razdioba izvedenih iz diskriminativnih riječi.

Dosadašnji pristupi računalnoj analizi medijske agende mjere zastupljenost tema pomoću

broja članaka ili broja riječi pridruženih temi i mjerenje se provodi izravno na temelju podataka iz tematskog modela – težina tema za članke ili varijabli koje pridružuju teme pojedinim riječima. Pri tome se mjerenje ne izdvaja kao zaseban korak i ne provodi se kvantitativno vrednovanje mjerenja, već se mjerenje zastupljenosti tema vrednuje posredno, promatranjem korelacije zastupljenosti tema s događajima i s drugim varijablama od interesa. Takva vrednovanja u dosadašnjim istraživanjima ne ukazuju na nedostatke u opisanom postupku mjerenja, međutim izostaje precizno kvantitativno vrednovanje koje je važno provesti, posebno u kontekstu primjena na znanstvena istraživanja teksta [41]. Bitna razlika predložene metode je definicija novih tema od interesa za analitičara i te teme ne moraju odgovarati temama modela iz koraka otkrivanja tema te se stoga ne mogu uvijek mjeriti dosadašnjim pristupom, izravno iz podataka već izgrađenih tematskih modela. Iz navedenih razloga, mjerenju se pristupa kao zasebnom koraku, problem mjerenja se definira kao problem klasifikacije dokumenata s višestrukim oznakama, pri čemu oznake odgovaraju mjerenim temama, što omogućava precizno kvantitativno vrednovanje mjerenja. Predlaže se i metoda mjerenja novodefiniranih tema, odnosno označavanja dokumenata temama.

Mjerenje definiramo kao instancu problema *klasifikacije s višestrukim oznakama* (engl. *multilabel classification*) [70], pri čemu je zadatak označiti novinske članke s jednom ili više tema preuzetih iz koraka definicije tema. Drugim riječima, u kontekstu klasifikacijskog problema teme odgovaraju klasama, a za klasifikaciju koristimo i nazive *označavanje dokumenata s temama* te *pridruživanje tema dokumentima*. Samo *mjerenje zastupljenosti tema* provodi se na način da se provede klasifikacija nakon čega se zastupljenost teme izračuna kao broj dokumenata označenih tom temom. Višestruke oznake, odnosno mogućnost da dokumenti pripadaju većem broju klasa, odražavaju činjenicu da dokumenti mogu govoriti o više od jedne teme, a na temelju članaka označenih s temama može se provesti i analiza supojavljivanja mjerenih tema u dokumentima. Ovaj pristup mjerenju zastupljenosti tema odgovara čestom pristupu u istraživanjima postavljanja agende u kojem se novinski članci ručno kodiraju u kategorije koje opisuju medijsku agendu. Vrednovanje metoda klasifikacije s višestrukim oznakama može se provesti nizom mjera [70].

Razmatraju se dvije metode za zadatak označavanja dokumenata temama, jednostavna i fleksibilna metoda nadziranog učenja koja daje dobre rezultate, te pristup temeljen na slabo nadziranim (engl. *weakly supervised*) tematskim modelima izgrađenima na temelju od ljudi konstruiranih skupova riječi koji odgovaraju mjerenim temama.

Mjerenje tema nadziranim učenjem Razvijene su mnoge metode nadziranog učenja za rješavanje problema klasifikacije s višestrukim oznakama [70] i svaku od njih moguće je primijeniti u opisanom slučaju klasifikacije dokumenata u tematske kategorije. Metoda *binarne relevantnosti* (engl. *binary relevance*) za klasifikaciju s višestrukim oznakama temelji se nekoj

od metoda binarne klasifikacije na način da za svaku od klasa izgradi model odlučivanja za pripadost toj klasi, te zatim primjerima (dokumentima) pridružuje klase na temelju odluka binarnih klasifikatora za svaku pojedinu klasu. Prednost binarne relevantnosti u odnosu na druge metoda je brzina učenja, koje se svodi na učenje binarnih modela, te lakoća proširenja modela s novim klasama – za svaku novu klasu potrebno je samo istrenirati model za novi binarni problem pripadnosti toj klasi [70]. Te prednosti u kontekstu analize agende omogućuju brže i lakše uključivanje novih tema u proces mjerenja. Nedostaci binarne relevantnosti su pretpostavka o neovisnosti klasa odnosno neučenje njihovih međuzavisnosti, te moguća neuravnoteženost skupova za učenje binarnih problema koji mogu sadržavati malo pozitivnih primjera [70]. Međutim, binarna relevantnost u kombinaciji s dobrim binarnim klasifikatorom rezultira dobrim performansama i može parirati složenijim metodama [70, 71, 72].

Kao metoda nadziranog učenja u ovdje opisanom pristupu koristi se metoda binarne relevanosti u kombinacijim sa strojem potpornih vektora (engl. *support vector machine* – SVM) kao binarnim klasifikatorom. Pri tome se koristi varijanta stroja potpornih vektora s linearnom jezgrenom funkcijom [73] koji daje dobre rezultate za problem klasifikacije tekstnih podataka . Pri mjerenju zastupljenosti tema metodama nadziranog učenja, za razliku od nenadziranih pristupa koji koriste tematske modele, potrebno je u izgradnju modela uložiti ljudsko vrijeme potrebno za izgradnju skupa podataka za učenje odnosno za označavanje dokumenata klasama. Za opisanu metodu označavanja dokumenata s temama nadalje se koristi oznaka *BR-SVM*.

Mjerenje tema slabo nadziranim tematskim modelima Drugi predloženi pristup označavanju dokumenata temama temelji se na *slabo nadziranim tematskim modelima* i predstavlja nadogradnju dosadašnjih pristupa mjerenju medijske agende koji za označavanje dokumenata koriste varijable tematskih modela (najčešće težine tema za dokumente). No dok dosadašnji pristupi mjere samo u velikoj mjeri slučajne teme naučene modelima, slabo nadzirani pristup podržava mjerenje proizvoljnih od strane analitičara definiranih tema. Pri slabo nadziranom učenju (engl. *weakly supervised learning*) ljudsko znanje koje definira korelaciju primjera i klasa ne sastoji se od primjera označenih s klasama nego od drugih podataka čija je izrada brža od označavanja primjera, primjerice od riječi označenih s klasama. Predloženi pristup koristi ljudsko znanje u vidu riječi koje opisuju mjerene semantičke teme kako bi se izgradio i za mjerenje koristio tematski model čije teme odgovaraju semantičkim temama.

Predložena metoda srodna je slabo nadziranim metodama za klasifikaciju teksta koje ugrađuju ljudsko znanje u model pomoću klasa definiranih skupovima riječi [74], posebice s nedavno predloženim metodama koje koriste opisani pristup u kombinaciji sa tematskim modelima [75, 76, 77, 78]. Takvi slabo nadzirani modeli ili pretpostavljaju da klase direktno odgovaraju temama modela [75, 76] ili modeliraju klase uvođenjem zasebnih tema-klasa i varijabli koje određuju pripadnost dokumenata klasi [77, 78]. Ljudsko znanje u vidu klasa opisanih sku-

pom riječi uključuje se u izgradnju modela izravno preko strukture modela [77, 78], pomoću apriornih distribucija [75], inicijalizacije tema (težina riječi za teme) [77], ili putem umjetne izgradnje skupa tekstova za učenje modela [76]. Ove metode imaju niže performanse od nadziranih metoda klasifikacije, no za njihovu izgradnju potrebno je znatno manje ljudskog vremena – vrijeme potrebno za izradu skupova riječi koji opisuju klase. Samom procesu izrade skupova riječi nije posvećeno mnogo pažnje – jednostavno se odabiru riječi koje dobro opisuje klasu.

Pristup sa slabo nadziranim tematskim modelima čini se pogodnim za rješavani problem označavanja dokumenata temama iz više razloga. Prvo, dokumenti se označavaju mjerenim semantičkim temama iz koraka definicije koje su dobivene na temelju tema modela izgrađenih u koraku otkrivanja tema što sugerira da su tematski modeli dobar model i za mjerene teme. Drugo, tematski modeli provode klasifikaciju s višestrukim oznakama modeliranjem tema kao cjeline, za razliku od binarne relevantnosti koji gradi zaseban klasifikator za svaku temu i ne modelira međuovisnost tema. Treće, u odnosu na nadzirani pristup zadavanje klasa skupom riječi ima potencijal za velike vremenske uštede u odnosu na označavanje dokumenata klasama [75, 76, 77, 78]. Posljednje, postojeći eksperimenti ukazuju da pri razdvajanju semantički bliskih klasa slabo nadzirani tematski modeli daju bolje rezultate od nadziranih modela [75, 77]. U slučaju mjerenja medijske agende bliske klase mogu biti česta pojava pošto klase odgovaraju temama koje mogu biti definirane s ciljem preciznog mjerenja bliskih aspekata nekog koncepta.

Ovdje predloženi pristup klasifikaciji s višestrukim oznakama temeljen na slabo nadziranim tematskim modelima koristi model LDA i prema [75, 76] pretpostavlja direktnu korespondenciju tema modela s mjerenim temama odnosno klasama. Označivači opisuju klase skupom visoko karakterističnih riječi i to znanje se ugrađuje u model na način da se iz svakog skupa riječi formira apriorna vjerojatnost za temu modela koja odgovara klasi. Specifičnost klasa koje odgovaraju temama medijske agende je ta što su očekivano slične međusobno i slične temama koje se ne mjere izravno ali koje model svejedno mora razdvojiti od mjerenih tema. Zbog toga je postupak odabira riječi koje opisuju klase složeniji od do sada predloženih postupaka. Cilj postupka je odabrati visokospecifične i visokodiskriminativne riječi za mjerenu temu na temelju promatranja uz riječi vezanih dokumenata. Opisani pristup temeljen na apriornim vjerojatnostima primjenjiv je i na one vjerojatnosne modele različite od modela LDA koji koriste apriorne vjerojatnosne razdiobe za modeliranje vjerojatnosti riječi unutar tema. Treba napomenuti da je izbor modela za klasifikaciju neovisan o izboru modela koji se koristi u koraku otkrivanja tema. Glavna razlika predloženog u odnosu na postojeće pristupe je razrađena metoda odabira riječi koje opisuju klase odnosno mjerene teme. S druge strane, predloženi pristup je jednostavan s aspekta strukture tematskog modela i unošenja podataka o riječima u model – koriste se standardni model LDA i apriorne vjerojatnosne razdiobe.

Odabir visokodiskriminativnih riječi Ovdje se opisuje postupak izgradnje liste visokodiskriminativnih riječi koje dobro opisuju mjerenu semantičku temu definiranu tijekom koraka otkrivanja i definicije tema. Podaci na temelju kojih se vrši odabir riječi su podaci o temi zabilježeni tijekom tih koraka – druge semantičke teme bliske mjerenoj temi i teme modela koje odgovaraju tim semantičkim temama. Ideja postupka je preko tema modela povezanih s mjerenom temom doći do liste riječi koje su očekivano vezane uz mjerenu temu te zatim odabrati one riječi za koje povezani dokumenti (povezanost se mjeri tf-idf težinama) zaista pripadaju mjerenoj temi.

Prvi korak je sastavljanje liste semantičkih tema vezanih uz mjerenu temu. Zatim se sastavi, na temelju tablice semantičkih tema konstruirane u koraku otkrivanja tema, lista tema modela koje odgovaraju semantičkim temama. Te teme modela predstavljene su utežanim listama riječi i dokumenata koje se mogu interpretirati kao semantičke teme. Teme modela bliske mjerenoj temi zatim se pregledavaju redom, u poretku prema procijenjenoj semantičkoj sličnosti s mjerenom temom. Za svaku temu modela, odabiru se one riječi za koje se na temelju dokumenata utvrdi visoka povezanost s mjerenom temom. Riječi se rangiraju prema mjeri koja kombinira njihovu povezanost s temom i diskriminativnost riječi na razini zbirke tekstova. Ta mjere definirana je kao umnožak težine riječi za temu i inverzne dokumentne frekvencije riječi (koja se standardno koristi za računanje tf-idf težina). Povezanost s temom očekivano korelira s povezanošću s mjerenom semantičkom temom, dok se mjerom diskriminativnosti nastoji doći do riječi koje su vezane specifično uz mjerenu temu. Riječi se razmatraju prema opisanom poretku i za svaku se riječ pregleda lista dokumenata uređena prema tf-idf težinama, standardnoj mjeri povezanosti riječi i dokumenta na razini zbirke tekstova. Ukoliko među 20 članaka s najvišim tf-idf ocjenama za riječ njih najmanje 80% pripada mjerenoj semantičkoj temi, riječ se dodaje na listu diskriminativnih riječi. Pregled rangiranih riječi teme provodi se dok korisnik ne procijeni da su riječi postale nevezane uz mjerenu temu, nakon čega započinje obrada sljedeće teme. Cijeli postupak pregleda tema i odabira riječi provodi se sve dok se ne obradi ukupno 300 riječi ili dok se ne odabere 15 riječi.

Izgradnja modela Na temelju izgrađenih lista riječi visokodiskriminativnih za mjerene teme definiraju se apriorne Dirichletove distribucije za teme modela koje će odgovarati mjerenim temama. Definicija apriornih distribucija izvodi se pomoću dva parametra: za svaku visokodiskriminativnu riječ iz liste definira se njena apriorna vjerojatnost P_s, dok se za svaku ostalu riječ izravno definira vrijednost pr apriornog vektora koja odgovara toj riječi. Sam vektor koji definira Dirichletovu apriornu distirbuciju zatim se konstruira na način da se za riječi koje nisu u listi vrijednosti vektora postave na pr, dok se za visokodiskriminativne riječi izračuna vrijednost pr_s na način da očekivanje svake te riječi iznosi P_s. Eksperimenti s izgradnjom modela za mjerenje pokazali su da su pr=0.001 and P_s=0.03 dobre početne vrijednosti. Važan parametar

izgradnje modela je i broj tema T. Na temelju postojećih eksperimenata s modeliranjem medijske agende smatramo da je T=100 dobar izbor za zbirke od nekoliko desetaka tisuća tekstova, T<50 dobar izbor za manje zbirke od nekoliko tisuća tekstova a T>500 za velike zbirke od nekoliko stotina tisuća ili više tekstova.

Za provođenje označavanja dokumenata s temama potrebna je, osim tematskog modela, i *metoda označavanja* koja na temelju težina tema za dokument dobivenih iz modela donosi odluku o označavanju odnosno o pripadnosti dokumenta temama. Drugim riječima, klasifikator temeljen na slabo nadziranim tematskim modelima sastoji se od tematskog modela i metode označavanja. Razmatraju se dvije metode označavanja. Prva je metoda *jednostrukog označavanja* koja označava članak jednom temom s najvećom težinom. Druga je metoda *višestrukog označavanja* koja označava članak temom s najvećom težinom te svim ostalim temama čije težine su veće od praga t. Prva metoda provodi klasifikaciju s jednostrukim oznakama odnosnu standardnu klasifikaciju dokumenata no takvo označavanje je samo poseban slučaj višestrukog označavanja i ne predstavlja problem prilikom vrednovanja.

Učenje tematskih modela na kojima se temelji mjerenje definirano je opisanim parametrima apriorne distribucije i brojem tema. Sam klasifikator ovisi još i o odabranoj metodi označavanja te, u slučaju višestrukog označavanja, o pragu t. Razmatraju se dvije metode optimizacije navedenih parametara klasifikatora. Prva metoda iterativne optimizacije parametara sastoji se od učenja modela i prilagođavanja parametara na temelju promatranja tema naučenih modela i ocjene njihovog poklapanja s mjerenim semantičkim temama. Pri tome se parametar T, broj tema modela, odabire prema preporukama za izgradnju modela u koraku otkrivanja tema. Zatim se u nizu koraka prilagođavaju parametri za apriorne vjerojatnosti. Prilagođavanje se provodi dok se ne utvrdi da teme modela dobro odgovaraju mjerenim semantičkim temama, pri čemu se poklapanje teme modela i mjerene semantičke teme ocjenjuje, kao u koraku otkrivanja tema, na temelju riječi i dokumenata. Provedeni eksperimenti pokazuju da ovaj postupak unutar maksimalno 5 koraka dovodi do dobrih modela.

Drugi pristup optimizaciji parametara klasifikatora je iscrpno razmatranje svih kombinacija te odabir one koja rezultira najboljim klasifikacijskim rezultatima na skupu označenih dokumenata. Za svaki od parametara, na temelju provedenih eksperimenata, predložen je skup vrijednosti za optimizaciju. Za *pr* koji definira vrijednost apriornog vektora za obične riječi predlažu se vrijednosti $\{0.0005, 0.001, 0.002, 0.005\}$, dok se za parametar P_s koji definira vjerojatnosti visokodiskriminativnih riječi predlažu vrijednosti $\{1\%, 3\%, 6\%, 12\%\}$. Za broj tema T predlaže se $\{50, 100, 120, 150\}$. Predlaže se razmatranje obje metode označavanja, a u slučaju višestrukog označavanja predlažu se vrijednosti $\{7\%, 10\%, 13\%, 15\%\}$ za prag t. Optimizacija se provodi tako da se za svaku od kombinacija parametara izgradi tematski model te potom vrednuje klasifikator temeljen na modelu i metodi odlučivanja. Sve kombinacije se zatim rangiraju prema mikro F_1 mjeri, nakon čega se između metoda unutar 0.01 od najboljeg rezultata odabere ona s

najvišom vrijednosti makro F_1 mjere. Pristup s optimizacijom očekivano daje bolji klasifikator od iterativne izgradnje, no cijena koja se plaća je vrijeme potrebno za označavanje dokumenata.

Prema opisanom postupku postoje četiri bitno različita tipa klasifikatora temeljenih na slabo nadziranim tematskim modelima koji se razlikuju prema metodi označavanja i metodi optimizacije parametara. Za klasifikatore s jednostrukim označavanjem koriste se oznake *WSTM-JO-OPT* i *WSTM-JO*, ovisno o tome da li je provedena iscrpna optimizacija parametara ili iterativna optimizacija. Analogno, za klasifikatore s višestrukim označavanjem koriste se oznake *WSTM-VO-OPT* i *WSTM-VO*.

3.4 Analiza agende političkih vijesti iz SAD-a

U ovom odjeljku opisuje se analiza agende američkih političkih vijesti [79]. Cilj istraživanja je ispitivanje predložene metode analize medijske agende kroz primjenu, s posebnim naglaskom na kvantitativno vrednovanje metoda mjerenja tema na medijskoj agendi. Provedeno je sakupljanje zbirke tekstova, otkrivanje tema koje čine medijsku agendu, odabir i definicija skupa tema okvirno vezanih uz građanska i ljudska prava, te mjerenje zastupljenosti tih tema i vrednovanje metoda mjerenja.

Zbirka tekstova Istraživanje je provedeno na zbirci novinskih članaka s političkim vijestima iz SAD-a objavljenih na velikim web portalima. Izgradnji zbirke prethodilo je dvotjedno sakupljanje URL adresa web portala praćenjem poveznica s Google News servisa, nakon čega su portali rangirani po ocjenama popularnosti (internet prometu unutar SAD-a) dobivenih Alexa Rank[2] servisom. Nakon toga je odabrano 25 najpopularnijih portala, njihovim indeksiranjem dobivena je lista izvora vijesti (engl. *news feed*) te su ručno odabrani izvori koji sadrže samo političke vijesti. Zbirka sadrži članke sa sljedećih 19 portala, dobivenih is početne liste uklanjanjem portala bez izvora vijesti s političkim vijestima: *Bloomberg Politics*, *CBS News*, *CNBC*, *CNN*, *Daily News*, *Fox News*, *Houston Chronicle*, *International Business Times*, *NBC News*, *Reuters*, *SFGate*, *The Atlantic*, *TheBlaze*, *The Guardian*, *The Huffington Post*, *The New York Times*, *The Wall Street Journal*, *The Washington Post*, and *Time*.

Novinski članci dobiveni su praćenjem izvora vijesti te ekstrakcijom i pohranom tekstova članaka, za što je razvijena programska potpora.[3] Zbirka sadrži članke objavljene između 26.01.2015. i 13.04.2015. Nakon pregleda slučajnog uzorka članaka, odlučeno je da se uklone vrlo kratki tekstovi te tekstovi koji ne odgovaraju vijestima, poput poruka o greškama pri dohvatu stranice te zaglavlja video i foto galerija. Jednostavna heuristička metoda koja se pokazala dobrom za filtriranje tekstova je uklanjanje tekstova koji sadrže manje od 40 pojavnica

[2] `http://www.alexa.com`

[3] `https://github.com/dkorenci/feedsucker`

(engl. *tokens*). Nakon filtriranja i uklanjanja duplikata, zbirka sadrži 24.532 novinska članka i slobodno je dostupna.[4]

3.4.1 Postupak otkrivanja tema

Otkrivanje tema na medijskoj agendi provedeno je primjenom metode predložene u 3.3.1.

Izgradnja tematskih modela Otkrivanje je provedeno na temelju LDA tematskih modela LDA [1] naučenih brzim algoritmom učenja [8], a kao programska potpora korišten je Gensim paket [80]. Tekstovi u zbirci pretprocesirani su uklanjanjem zaustavnih riječi (engl. *stopword*) i svođenjem riječi na korijenski oblik (engl. *stemming*) pomoću alata iz NLTK paketa [81]. Iz riječnika su uklonjene riječi s vrlo visokom i vrlo niskom frekvencijom – riječi koji se pojavljuju u više od 10 posto dokumenata i u manje od 5 dokumenata. Konačni riječnik sadrži 23.155 različitih riječi a cijela zbirka tekstova sastoji se od 3.951.990 riječi.

Izgrađeni su modeli sa T=50 i T=100 tema. Prema [7], odabrane su vrijednosti hiperparametara α=$50/T$ (kontrola razdiobe vjerojatnosti tema u dokumentu) i β=0.01 (kontrola razdiobe riječi za temu). Hiperparametri algoritma učenja optimirani su prema [8], pretraživanjem prostora parametara prema kriteriju zbunjenosti [1] kao mjeri kvalitete modela. Optimalne vrijednosti parametara su S=1000 (broj dokumenata korišten za osvježavanje parametara modela u svakom koraku), τ_0=1000 i κ=0.5 (parametri za kontrolu relativne važnosti naučenih promjena pri svakom osvježavanju).

Otkrivanje tema Otkrivanje tema u zbirci provedeno je na temelju 5 modela LDA: tri modela sa 50 tema i dva modela sa 100 tema. Teme su pregledane od strane dva označivača i svaka tema je interpretirana na temelju liste najbolje rangiranih novinski članaka za temu (naslova i po potrebi cjelovitog teksta) te liste najbolje rangiranih riječi za temu. Oba označivača su prvo pregledala 50 tema nakon čeka su prodiskutirali i prilagodili konvencije označavanja tema. Preostalih 300 tema raspoređeno je između dva označivača koji su ih odvojeno pregledali i interpretirali. Semantičke teme otkrivene tijekom postupka pregleda i interpretacije dodavane su u zajedničku tablicu semantičkih tema [5].

Na temelju označavanja 350 tema modela otkrivene su ukupno 134 semantičke teme. Pri tome 189 tema modela (54%) dobro odgovara točno jednoj semantičkoj temi, 121 tema modela (34.6%) sadrži ili dvije semantičke teme ili odgovara jednoj temi uz dodatni šum, dok preostalih 40 tema modela (11.4%) odgovara šumu (slučajni novinski članci i nesadržajne riječi). Neke semantičke teme pojavile su se u svim modelima, dok su neke detektirane samo u samo jednom ili dva modela. Tablica 3.1 daje primjere tema modela i vezanih semantičkih tema. Pregled

[4] `http://bit.ly/AGENDADATASET`

[5] tablica semantičkih tema dostupna je na http://takelab.fer.hr/agenda

i interpretacija tema modela te izrada tablice semantičkih tema trajali su 35 osoba-sati, što u prosjeku iznosi 6 minuta po temi.

Provedeni postupak otkrivanja tema pokazao je da se modelom LDA te predloženim postupkom pregleda i interpretacije tema postiže dobar tematski pregled medijske agende. Otkrivene su 134 semantičke teme koje odgovaraju nizu događaja, osoba, organizacija te apstraktnih koncepata vezanih uz razna društvena pitanja. Modeli sa 50 tema naučili su u pravilu općenitije tema a modeli sa 100 tema specifičnije teme, što je u skladu s opažanjima iz postojećih istraživanja. Za modele od 50 tema uočena je veća pojava tema koje odgovaraju dvama semantičkim temama, dok je za modele od 50 tema uočena veća količina tema koje odgovaraju šum. Proces otkrivanja tema je vremenski zahtjevan zbog velikog broja tema i potrebno ga je ubrzati metodama za automatsku detekciju loših tema i ponovljenih tema.

3.4.2 Postupak definicije tema

Na temelju promatranja semantičkih tema definiran je manji skup tema za podrobniju analizu. Odabrani skup tema potaknut je istraživanjima postavljanja agende koje se često fokusiraju na pitanja iz domene ljudskih i građanskih prava [37, 82, 83]. Skup sadrži 12 semantičkih tema okvirno povezanih s tom domenom: *pokret za građanska prava* (engl. *civil rights movement*), *LGBT prava*, *policijsko nasilje* (engl. *police brutality*), *Chapel Hill napad*, *reproduktivna prava* (engl. *reproductive rights*), *nasilje nad ženama*, *smrtna kazna*, *nadzor građana* (engl. *surveillance*), *marihuana*, *pravo na oružje* (engl. *gun rights*), *neutralnost interneta* (engl. *net neutrality*), *cijepljenje*. Semantičke teme ovog tipa u engleskom se nazivaju problemima (engl. *issue*) – naziv koji označava teme od šireg društevnog značaja i interesa koje često bude kontroverze.

Neke od ovih dvanaest tema dobro odgovaraju točno jednoj temi modela, dok za ostale teme, primjerice teme *nasilje nad ženama* i *LGBT prava*, ne postoje odgovarajuće tema modela. Takve novodefinirane semantičke teme demonstriraju potrebu za korakom definicije tema. Iako te teme odgovaraju konceptima od potencijalnog interesa za društvenoznanstvene analize, nisu precizno detektirane niti od jedne teme modela iz koraka otkrivanja i njihovu zastupljenost nije moguće mjeriti pomoću tih modela. Te teme semantički su povezane s otkrivenim temama koje služe kao smjernica za njihovu definiciju no za mjerenje tih tema potrebno je izgraditi nove prilagođene modele. Tablica 3.1 prikazuje odnose između dviju novodefiniranih tema, vezanih semantičkih tema te tema modela u kojima se pojavljuju semantičke teme.

3.4.3 Postupak mjerenja tema

Postupak mjerenja definiranog skupa tema proveden je metodom opisanom u 3.3.3 – najprije su konstruirani skupovi visokodiskriminativnih riječi za mjerene teme na temelju kojih su naučeni slabo nadzirani tematski modeli čije teme se poklapaju s mjerenim temama. Naučen je i nad-

Tablica 3.1: Primjeri tema modela označenih *semantičkim temama* i izvedenim **novodefiniranim** semantičkim temama. Svaka tema modela označena je najbolje rangiranim riječima za tu temu. Tema modela M1.T43 (model 1, tema 43) je primjer mješavine dvije semantičke teme.

Novodefinirana Tema / *Semantička tema* / Tema modela
nasilje nad ženama
ženska prava
M1.T43: žrtva, spol, identitet, zlostavljanje, trudna
seksualno nasilje
M0.T29: student, pobačaj, seksualno, žrtva, muškarci
M11.T25: muškarci, seksualno, djevojka, napad, žena
LGBT prava
transrodna pitanja
M1.T43: žrtva, spol, identitet, zlostavljanje, trudna
gay prava
M1.T24: religiozan, gay, brak, Indiana, sloboda
M10.T54: brak, par, gay, istospolni, sudac

zirani BR-SVM model temeljem dokumenata označenih s temama. te se provodi vrednovanje i usporedba svih metoda mjerenja.

Izrada skupa označenih dokumenata Označen je uzorak od 2800 novinskih članaka iz zbirke koji se koristi za izgradnju i vrednovanje modela. Dokumenti su označeni s 12 prethodno definiranih semantičkih tema od strane dva ljudska označivača. Svakom dokumentu pridjeljene su teme o kojima dokument govori – mogućnosti uključuju niti jednu, jednu ili veći broj definiranih tema. Popis svih tema nalazi se u tablici 3.2. Oba označivača prvo su označila kalibracijski uzorak od 200 članaka i prodiskutirali označavanje. Međusobno slaganje označivača na kalibracijskom skupu, mjereno Cohen-ovim kappa koeficijentom, iznosi κ=0.93, što se smatra savršenim slaganjem [84]. Skup preostalih dokumenata zatim je podjeljen na dva jednaka dijela i svaki je označen od strane jednog označivača. Na taj način dobiven je skup od 2600 dokumenata označenih s temama, koji je podjeljen na skup za učenje (engl. *train set*) sa 1600 dokumenata i skup za ispitivanje (engl. *test set*) sa 1000 dokumenata. 14.8% dokumenata je označeno s jednom ili više tema, dok ostali dokumenti nisu označeni s pripadnim temama. Vrijeme potrebno za označavanje svih 2800 dokumenata iznosi 14 sati.

Izrada modela Prije izgradnje slabo nadziranih tematskih modela, prema postupku opisanom u 3.3.3 za svaku od mjerenih tema konstruirana je lista visokodiskriminativnih riječi. Konstrukcija se temelji na pregledu srodnih tema modela iz koraka otkrivanja odnosno pregledu riječi vezanih uz te teme i dokumenata vezanih uz riječi. Tablica 3.2 sadrži korijenovane riječi za de-

Tablica 3.2: Visokodiskriminativne riječi za mjerene semantičke teme.

Semantička tema	Visokodiskriminativne riječi
pokret za građanska prava	Selma, most, obljetnica, Lewis, ropstvo
policijsko nasilje	policija, Ferguson, pucati, Martin, neozlijeđen, Slager
LGBT prava	brak, istospolni, Indiana, Alabama, Arkansas, Pence, RFRA, diskriminacija, LGBT, transrodni, lezbijka, orijentacija, homoseksualac, spol
nasilje nad ženama	seksualni, napad, sestrinstvo, zlostavljanje, silovanje, silovatelj, Phi, Kappa, UVA, kućno, nasilje
reproduktivna prava	pobačaj, trudnoća, klinika, trudna, za_život, fetus, roditeljstvo, nerođen, kontraceptivna, pilula, rođenje, reproduktivni
pravo na oružje	pištolj, nošenje, vatreno_oružje, puška, NRA, skriveno, streljivo
cijepljenje	cijepivo, ospice, bolest, imunitet, autizam, anti-vaxx, dječja_paraliza, necijepljen
neutralnost interneta	internet, FCC, net, neutralan, širokopojasni, Wheeler, Verizon, ISP
marihuana	marihuana, medicinska, Colorado, trava, posjedovati, kanabis
smrtna kazna	smrt, smrtonosna, injekcija, kazna, vod, Lockett, Gissendan
nadzor građana	NSA, nadzor, Snowden, Edward, GCHQ, metapodaci, Kasperski, privatnost
Chapel Hill napad	Chapel, Abu-Salha, Razan, Hick, Yusor

finirane teme. Može se vidjeti da neke od riječi opisuju općenite koncepte, dok druge riječi opisuju specifične koncepte i imenovane entitete povezane s temama u vremenskom i geografskom kontekstu zbirke tekstova. Izrada listi riječi pokazala je da je postupkom izrade predloženim u 3.3.3 moguće efikasno doći do visokodiskriminativnih riječi za teme – procjena diskriminativnosti riječi temeljem vezanih dokumenata pokazala se lako izvedivom i za sve teme uspješno je izgrađen skup od 5–15 riječi, za što je u prosjeku bilo potrebno 20 minuta po temi. Pri izgradnji slabo nadziranih tematskih modela liste riječi pretočene su u apriorne Dirichletove distribucije kako bi se naučile teme modela koje odgovaraju mjerenim temama.

Izgradnja modela za označavanje dokumenata temama provedena je prema 3.3.3, pri čemu su izgrađene sve varijante klasifikatora temeljenih na slabo nadziranim tematskim modelima (WSTM-JO-OPT, WSTM-JO, WSTM-VO-OPT, WSTM-VO) te nadzirani klasifikator BR-SVM.

Optimizacija parametara slabo nadziranih klasifikatora – prag metode označavanja te parametri definicije apriornih distribucija i broj tema – provedena je na oba načina predložena u 3.3.3, iscrpnim vrednovanjem kombinacija parametara te iterativnom optimizacijom parametara na temelju promatranja modela. Hiperparametri za izgradnju samih modela LDA isti su

onima za učenje modela u koraku otkrivanja tema. U procesu vrednovanja kombinacija parametara performanse modela vrednovane su na polovici označenih dokumenata kako bi modeli bili usporedivi s nadziranim modelom. Iterativno optimirani klasifikatori nisu optmirani pomoću označenih dokumenata, već je broj tema postavljen na T=100, dok su parametri za apriorne distribucije optimirani na temelju promatranja dobivenih tema. Iterativna optimizacija parametara prema predloženom postupku pokazala se efikasnom – parametri koji daju kvalitetne modele pronađeni su za manje od 5 iteracija a usporedba tema modela s mjerenim temama nije zahtijevala mnogo vremena.

Naučen je i nadzirani klasifikator BR-SVM– binarni SVM model u kombinaciji s metodom binarne relevantnosti. BR-SVM klasifikator dokumentima pridružuje skup tema na temelju odluka binarnih modela za svaku od tema. SVM modeli naučeni su iz dokumenata predstavljenih standardnim tf-idf težinama – za riječ w i dokument d , $tfidf(w,d) = (1 + \log freq(w,d)) \times \log(N_{doc}/N_{doc}(w))$. Hiperparametri SVM modela su optimirani unakrsnom provjerom s pet preklopa (engl. *five-fold crossvalidation*) na skupu za učenje korištenjem mikro F_1 kao ciljne funkcije, nakon čega je model s najboljim parametrima naučen na cijelom skupu. Korištena je LIBLINEAR implementacija SVM modela [73].

Nadzirana BR-SVM metoda naučena je na temelju dokumenata označenih temama. U slučaju slabo nadziranih klasifikatora svi modeli izgrađeni su na temelju skupa visokodiskriminativnih riječi i tekstova u zbirci no za klasifikatore optimirane iscrpnim pretraživanjem parametara (WSTM-JO-OPT, WSTM-VO-OPT) korišteni su i označeni dokumenti. Nadzirani BR-SVM klasifikator s jedne, te WSTM-JO-OPT i WSTM-VO-OPT klasifikatori s druge strane konstruirani su na način da budu uporedivi po vremenu označavanja (označavanje dokumenata i izrada listi riječi) za izradu modela. Za navedene slabo nadzirane klasifikatore korištene su diskriminativne riječi i 800 označenih dokumenata, dok je za BR-SVM korišteno 1600 označenih dokumenata. Razlog tome je što vrijeme potrebno za označavanje 800 dokumenata odgovara vremenu potrebnom za konstrukciju listi riječi – vrijeme označavanja 2800 dokumenata iznosi 14 sati, dok su za izradu listi riječi bila potrebna 4 sata.

3.4.4 Vrednovanje modela za mjerenje

Vrednujemo svih pet izgrađenih klasifikatora, nadzirani BR-SVM model te iterativno optimirane slabo nadzirane modele WSTM-JO i WSTM-VO te iscrpno optimirane slabo nadzirane modele WSTM-JO-OPT i WSTM-VO-OPT izgrađene na način da vrijeme označavanja podataka potrebnih za konstrukciju modela bude usporedivo. Iscrpno optimirani slabonadzirani modeli koriste istu količinu vremena kao i BR-SVM, približno 8 sati, dok neoptimirani slabonadzirani modeli koriste 4 sata, vrijeme potrebno za izradu listi diskriminativnih riječi. Klasifikatori se vrednuju na skupu za ispitivanje koji sadrži 1000 označenih dokumenata.

Performanse klasifikatora mjerene su mjerama preciznosti (engl. *precision* – P), odziva

(engl. *recall* – R) te F_1 mjerom (engl. F_1 *measure*). Preciznost je definirana kao udio primjera ispravno pridjeljenih klasi od strane modela (udio dokumenata označenih s temom koji zaista pripadaju temi). Odziv je definiran kao udio svih primjera koji pripadaju klasi za koje je model ispravno zaključio da pripadaju klasi (udio svih dokumenata koji pripadaju temi koje je model označio s temom). Mjera F_1 je kombinacija preciznosti i odziva, definirana kao njihova harmonijska sredina. Pošto se vrednovanje provodi za problem klasifikacije s višestrukim oznakama a navedene mjere su definirane za slučaj binarne klasifikacije, koriste se dva načina agregacije odnosno adaptacije tih mjera za slučaj s više klasa. Makro agregacija jednostavno uprosječuje performanse po pojedinim klasama. Mikro agregacija uprosječuje performanse na razini pojedinog primjera (dokumenta) – svaki dokument se, neovisno o klasi, promatra kao ispravno ili neispravno klasificiran te se zatim vrijednosti mjera računaju kao u slučaju binarne klasifikacije.

Tablica 3.3 sadrži rezultate vrednovanja klasifikatora. Iscrpno optimirani slabo nadzirani klasifikator WSTM-VO-OPT ima najbolje performanse, mjereno mikro i makro F_1 mjerama. BR-SVM klasifikator, iako usporediv sa WSTM-VO-OPT klasifikatorom prema mikro F_1 mjeri, ima značajno niže rezultate mjereno makro F_1 mjerom što ukazuje na neuravnotežene performanse po pojedinim klasama. BR-SVM klasifikator postiže visoku preciznost i niži odziv dok WSTM-VO-OPT ima viši odziv i uravnoteženije vrijednosti preciznosti i odziva. Iterativno optimirani WSTM-VO klasifikator s višestrukim označavanjem daje slabije performanse, osobito preciznost, što ukazuje na potrebu za optimizacijom kod ove metode označavanja. Oba klasifikatora s jednostrukim oznakama WSTM-JO-OPT i WSTM-JO imaju usporedive performanse, što ukazuje da optimizacija parametara nema veliki utjecaj kod jednostrukog označavanja. Performanse iterativno optimiranog WSTM-JO modela prema mikro F_1 mjeri ne zaostaju mnogo u usporedbi sa BR-SVM klasifikatorom, a prema makro F_1 mjeri klasifikatori su usporedivi, s tim da je prednost WSTM-JO klasifikatora ta da za je njegovu konstrukciju potrebno značajno manje vremena.

Tablica 3.4 prikazuje performanse po pojedinim temama za najbolji slabo nadzirani klasifikator WSTM-VO-OPT i nadzirani BR-SVM. Za klase s većim brojem primjera (više od 10) performanse su usporedive, no za klase s manjim brojem primjera BR-SVM, iako zadržava visoku preciznost, u pravilu ima nizak odziv što narušava njegove performanse. S druge strane, WSTM-VO-OPT zadržava visoki odziv dok preciznost u pravilu ostaje dobra. Ovi rezultati ukazuju da je unošenje znanja u model pomoću diskriminativnih riječi u slučaju manjih klasa radi bolje od označavanja primjera. Nizak odziv nadziranog modela ukazuje na ograničene mogućnosti generalizacije na temelju manjeg broja primjera, dok tematski modeli, čini se, mogu dobro naučiti koncept klase iz skupa riječi.

Provedeno je vrednovanje BR-SVM modela promatranjem ovisnosti performansi o broju dokumenata upotrijebljenih za učenje modela. Pri tome je za određeni broj dokumenata uprosječena mikro F_1 mjera izmjerena za pet slučajnih uzoraka te veličine. Dobivena krivulja učenja,

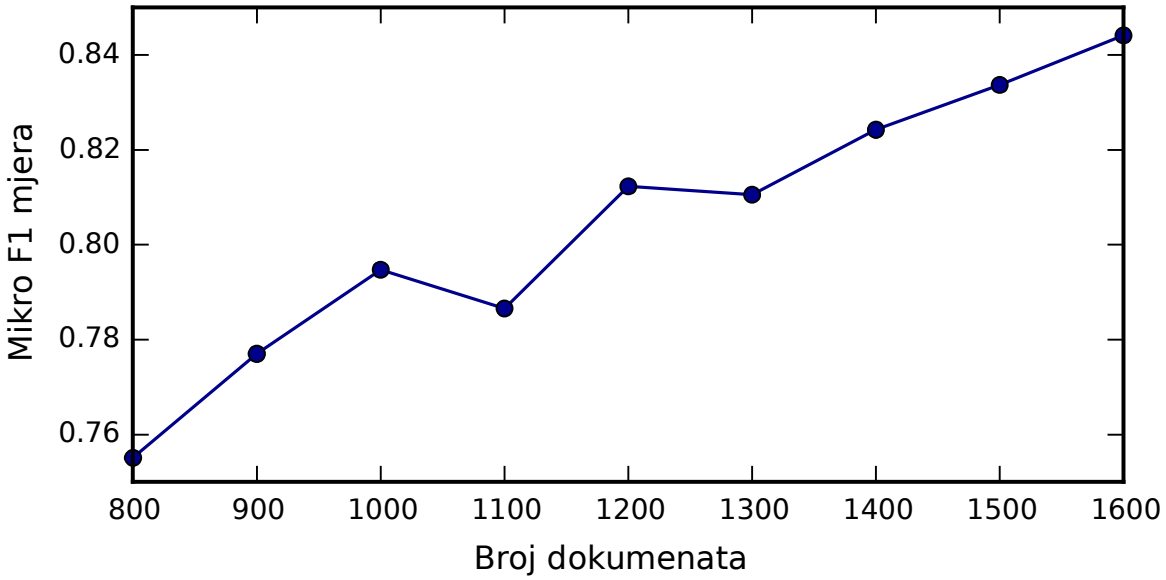
Tablica 3.3: Performanse klasifikatora uprosječene po klasama.

	Mikro			Makro		
Klasifikator	P	R	F1	P	R	F1
WSTM-JO	0.75	0.80	0.77	0.67	0.75	0.69
WSTM-JO-OPT	0.78	0.79	0.79	0.65	0.68	0.65
WSTM-VO	0.67	0.89	0.77	0.63	0.85	0.70
BR-SVM	**0.99**	0.74	0.84	**0.92**	0.59	0.68
WSTM-VO-OPT	0.80	**0.91**	**0.85**	0.74	**0.89**	**0.80**

Tablica 3.4: Performanse najboljih klasifikatora za svaku od klasa odnosno mjerenih tema.

		WSTM-VO-OPT			BR-SVM		
Tema	#	P	R	F1	P	R	F1
LGBT prava	60	1.00	0.93	0.97	0.98	0.95	0.97
cijepljenje	15	0.83	1.00	0.91	1.00	1.00	1.00
policijsko nasilje	15	0.72	0.87	0.79	1.00	0.67	0.80
marihuana	12	0.85	0.92	0.88	1.00	0.92	0.96
pokret za građanska prava	11	0.67	0.91	0.77	1.00	0.45	0.62
nasilje nad ženama	8	0.67	0.75	0.71	1.00	0.25	0.40
reproduktivna prava	7	0.42	0.71	0.53	0.00	0.00	0.00
pravo na oružje	7	0.78	1.00	0.88	1.00	0.29	0.44
neutralnost interneta	6	0.67	1.00	0.80	1.00	0.67	0.80
smrtna kazna	5	0.80	0.80	0.80	1.00	0.40	0.57
nadzor građana	4	0.50	0.75	0.60	1.00	0.50	0.67
Chapel Hill napad	1	1.00	1.00	1.00	1.00	1.00	1.00

prikazana na slici 3.1, ukazuje da se nadzirani BR-SVM klasifikator može poboljšati označavanjem većeg broja dokumenata. Međutim, označavanje većeg broja dokumenata povećava utrošak vremena potreban za izgradnju modela.



Slika 3.1: Krivulja učenja nadziranog BR-SVM klasifikatora.

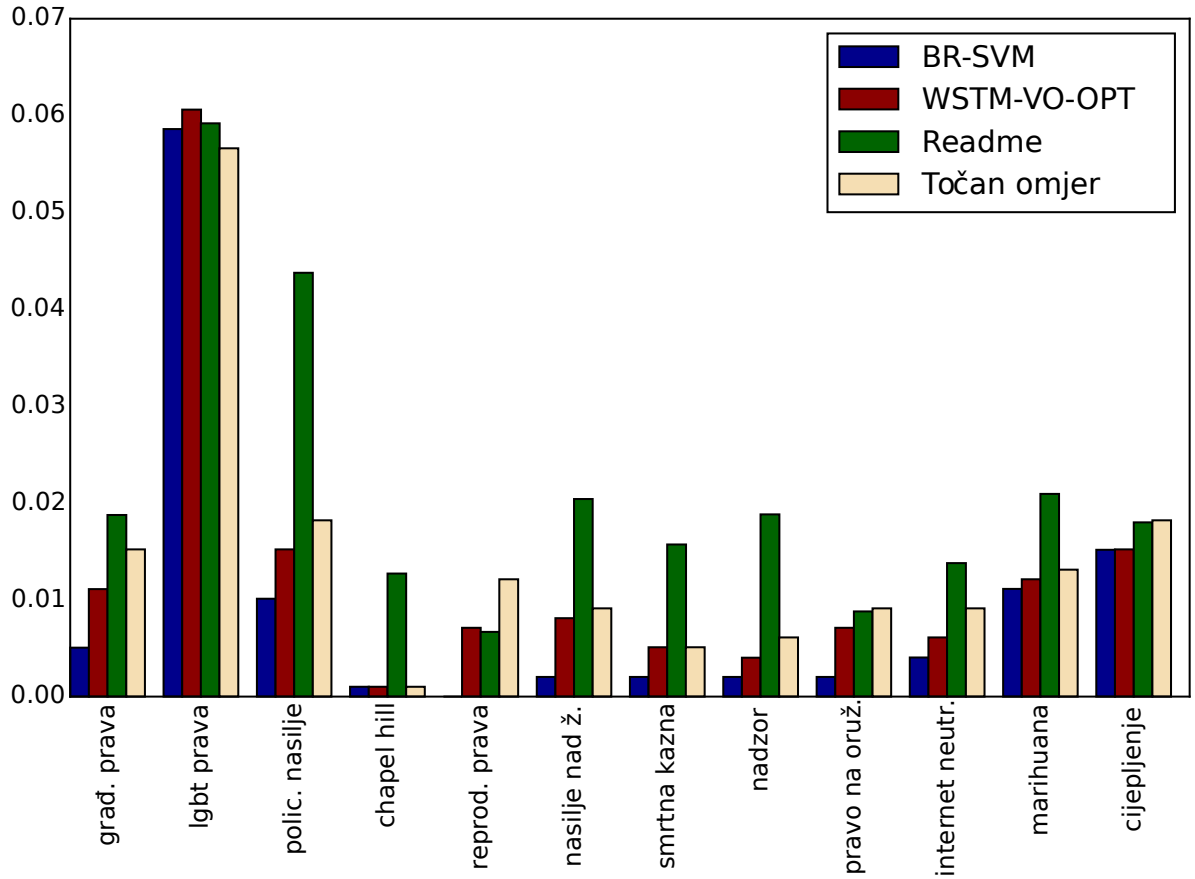
Izravno računanje omjera klasa U društvenoznanstvenim primjenama, često od interesa nije točna kategorizacija pojedinih tekstova već procjena omjera zastupljenosti kategorija u skupovima tekstova [85], primjerice u cijeloj zbirci ili svim tekstovima unutar određenog vremenskog intervala. Navedeno vrijedi i za istraživanja postavljanja agende, u kojima se medijska agenda nekada prikazuje pomoću omjera zastupljenosti tema na agendi [42, 86]. Kako bi se predložene metode mjerenja tema vrednovale u ovom scenariju, uspoređuju se najbolja slabo nadzirana metoda WSTM-VO-OPT i nadzirani BR-SVM s metodom predloženom u [85] (nadalje *Readme*). Readme metoda provodi učenje na označenim tekstovima no procjenjuje omjere klasa izravno, bez kategorizacije pojedinih dokumenata, Eksperimenti u [85] pokazuju da, u slučaju većeg broja kategorija, Readme ima bolje performanse od različitih varijanti SVM klasifikatora. Readme metoda slobodno je dostupna kao softverski paket.[6]

U ovdje razmatranom slučaju klasifikacije dokumenata s višestrukim oznakama, Readme se koristi za računanje, za svaku temu, omjera dokumenata koji pripadaju temi. Model Readme naučen je na skupu za učenje koristeći zadane (engl. *default*) parametre i provedeno je uprosječivanje rezultata pet modela pošto je algoritam učenja stohastički. Vrednovanje procjene omjera kategorija provedeno je primjenom svih metoda na skupu za ispitivanje.

Rezultati usporedbe prikazani su na slici 3.2. Prema [85], kvantitativna usporedba je napravaljena pomoću usrednjene apsolutne greške omjera (engl. *Mean Absolute Proportion Error*

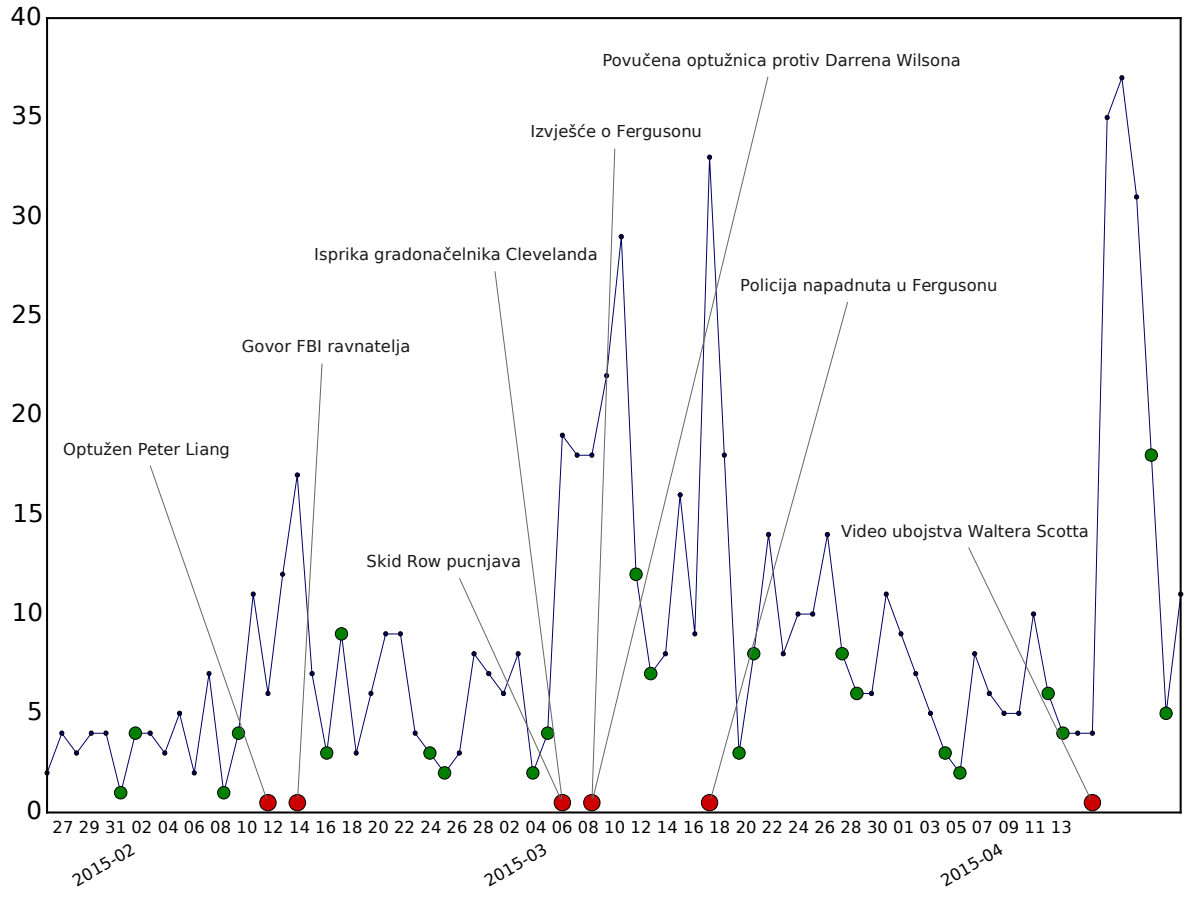
[6] `http://gking.harvard.edu/readme`

– MAPE) i srednje kvadratne greške (engl. *Root Mean Squared Error* – RMSE). MAPE iznosi 0.18 za WSTM-VO-OPT, 0.49 za BR-SVM, i 1.70 za Readme. RMSE iznosi 0.003 za WSTM-VO-OPT, 0.006 za BR-SVM, and 0.01 za Readme. Ovi rezultati pokazuju da pristup sa slabo nadziranim tematskim modelima daje postiže dobre rezultate i u ovom za društvenoznanstvene analize važnom scenariju. Štoviše, ti rezultati znatno su bolji od rezultata nadzirane BR-SVM metode, što se može objasniti prethodno opisanim slabim performansama BR-SVM metode na klasama s malo primjera te neuravnoteženošću preciznosti i odziva te metode.



Slika 3.2: Izračunati i točni omjeri klasa na skupu za ispitivanje.

Kvalitativna validacija Ovdje se primjenjuje često korištena metode prediktivne validacije tematskih modela na validaciju WSTM-VO-OPT modela. Provodi se vizualna analiza korelacije izmjerene zastupljenosti tema s vezanim događajima na primjeru teme *policijsko nasilje*. Validacija se provodi mjerenjem broja članaka na temu policijskog nasilja za svaki dan u vremenskom rasponu koji pokriva zbirku tekstova. Slika 3.3 prikazuje broj uz temu vezanih članaka kroz cijelo razdoblje pokriveno zbirkom tekstova, zajedno za koreliranim događajima. Vidi se da svako povećanje broja članaka korelira s barem jednim većim događajem vezanim uz temu. Ovaj primjer pruža dodatnu potvrdu za kvalitetu mjerenja i ujedno demonstrira način provođenja prediktivne validacije tematskih modela.



Slika 3.3: Broj novinskih članaka automatski označenih temom *policijskog nasilja* i stvarni događaji povezani s temom. Dani vikenda označeni su zelenom bojom.

3.5 Eksperiment na hrvatskim političkim tekstovima

U ovom odjeljku opisuje se analiza predizborne i postizborne medijske agende hrvatskih medija u razdoblju oko parlamentarnih izbora 2015. godine [87]. Istraživanje je provedeno s ciljem ispitivanja predložene metode analize medijske agende kroz primjenu primjenu na istraživanje izborne medijske agende. Naglasak istraživanja je na definiciji relevantnih tema, mjerenju njihove zastupljenosti i analizi agende temeljenoj na mjerenjima. Provedeni su koraci izgradnje zbirke tekstova, otkrivanja tema na medijskoj agendi, te definicije skupa tema vezanih uz izbore, nakon čega je provedeno mjerenje tema i analiza izborne agende.

Zbirka tekstova Istraživanje je provedeno na tekstovima sakupljenim sa sedam vodećih hrvatskih web portala: *Večernji list*, *Jutarnji list*, *Slobodna Dalmacija*, *Glas Slavonije*, *T-portal*, *Novi List*, and *RTL Televizija*. U prvom koraku odabrani su izvori (engl. *news feed*) s vijestima iz zemlje (hrvatske i regionalne vijesti), nakon čega su sakupljeni novinski članci objavljeni u razdoblju službene predizborne kampanje (od 21.10. do 6.11. 2015.), i razdoblju između izbora i konstituirajuće sjednice Sabora (od 8.11. do 28.12. 2015.). Iz zbirke su uklonjeni vrlo

kratki tekstovi (sa manje od 40 alfanumeričkih znakovnih nizova) i ne-tekstovi (poruke o grešci, zaglavlja video i foto galerija, upite za pretplatu itd.). Zatim je provedena deuplikacija grupiranjem gotovo istih tekstova na temelju Levenshteinove mjere udaljenosti (izračunate na razini riječi tekstova) te zadržavanjem samo jednog teksta po portalu iz svake grupe. Nakon filtriranja i deduplikacije zbirka tekstova sadrži 15.394 novinska članka.

3.5.1 Postupak otkrivanja tema

Otkrivanje tema na medijskoj agendi provedeno je primjenom postupka predloženog u 3.3.1.

Izgradnja tematskih modela Za otkrivanje semantičkih tema koristi se LDA tematski model [1] dostupan kao dio Gensim paketa [80]. Pretprocesiranje tekstova provedeno je uklanjanjem stop-riječi i ne-riječi, te korijenovanjem pomoću alata za korjenovanje hrvatskih riječi [88]. Algoritam učenja, njegovi hiperparametri kao i hiperparametri modela preuzeti su iz eksperimenta na američkim političkim tekstovima opisanim u odjeljku 3.4, pošto su obje zbirke istog reda veličine i sadrže tekstove istog tipa – novinske članke. Otkrivanje semantičkih tema provedeno je na temelju tri modela LDA – dva modela sa 50 tema te jednog modela sa 100 tema.

Otkrivanje tema U ovom koraku 200 tema dobivenih učenjem tri tematska modela pregledano je i interpretirano od strane sedam označivača – dvoje autora istraživanja te pet studenata završnih godina novinarstva. Prije početka otkrivanja tema, provedeno je treniranje označivača, pri čemu je postupak otkrivanja objašenjen i demonstriran, nakong čega je uslijedilo probno provođenje otkrivanja tema te rasprava. Označivači su zatim proveli postupak otkrivanja semantičkih tema opisan u 3.3.1, koristeći aplikaciju s grafičkim korisničkim sučeljem za pregled tema. Raspodijeljeni rad većeg broja označivača podržan je mrežnim pristupom aplikaciji instaliranoj na poslužitelju. Svaki od označivača obradio je u prosjeku 30 tema, pri čemu su svakom označivaču dodijeljene teme ravnomjerno odabrane iz svakog od modela. U prosjeku je svaki označivač potrošio 10 minuta na obradu jedne teme (minimalno 5.5 minuta, maksimalno 16.8 minuta). Vrijeme potrebno za obradu jedne teme kretalo se u rasponu od nekoliko minuta za jednostavne do 20-tak minuta za konceptualno složene teme. Ukupno vrijeme utrošeno za obradu svih tema iznosi 33 čovjek-sata.

Revizija semantičkih tema Rezultat postupka otkrivanja tema je skup od 106 semantičkih tema. Dodatnim pregledom tema utvrđene su greške u postupku otkrivanja – neke iste semantičke teme su ponovljene od strane više anotatora, neke su neprecizno označene, dok je u nekim slučajevima veza između semantičkih tema i tema modela upitna. Zaključeno je da se kvaliteta označavanja može poboljšati većim utroškom vremena za obuku anotatora, kao i razradom i provedbom strožeg postupka anotacije. Također su zamijećene teme poput *vremenske prognoze*

i *prometnih izvještaja* koje su nevažne za analizu agende. Iz navedenih razloga proveden je dodatni korak revizije tema tijekom kojeg su pregledane i ispravljene oznake i definicije semantičkih tema te njihove veze s temama modela.

Drugo opažanje o otkrivenim temama je da teme nisu međusobno isključive nego su često u hijerarhijskom odnosu i razlikuju se po razini općenitosti. Stoga su sve teme organizirane u taksonomiju kako bi se olakšao njihov pregled i kako bi se stekli dodatni uvidi u njihovu strukturu.[7] Primjerice, semantička tema *izborne ankete* smještena je pod temu *izborna predviđanja*, koja je zajedno s temom *izborni rezultati* smještena pod *izborni proces*. Taksonomija se pokazala kao vrlo koristan alat za odabir i detaljniju analizu semantičkih tema od interesa pošto omogućuje odabir i razmatranje tema na raznim razinama apstrakcije.

Korak revizije i taksonomizacije tema proveden je zajednički od strane autora istraživanja i za cijeli postupak bila su potrebna tri sata, odnosno 9 čovjek-sati. Po završetku koraka revizije preostala je sedamdeset i jedna semantička tema. Teme su organizirane u taksonomiju koja sadrži dvadeset i jednu kategoriju na najvišoj razini apstrakcije: *kazneni progon javnih osoba*, *pregovori nakon izbora*, *vanjska politika*, *terorizam i izbjeglička kriza*, *Katolička crkva*, *institucija predsjednika*, *oružane snage*, *izborni proces*, *ekologija*, *energetika*, *obrazovanje*, *turizam*, *decentralizacija i reforma lokalne i regionalne uprave*, *zdravstvo*, *mediji i novinari*, *sindikati i radnička prava*, *ekonomija*, *unutarstranački sukobi*, *poljoprivreda*, *odljev mozgova i demografija*, and *nevezani događaji*. Posljednja kategorija odnosi se na događaje iz izborne kampanje koji ne pripadaju u niti jednu od drugih kategorija.

Postupak otkrivanja pokazao je, isto kao i kod analize agende američkih medija, da modeli LDA te predloženi postupak pregleda tema daju dobar tematski pregled medijske agende. Otkrivene semantičke teme, kao što je vidljivo iz taksonomije, pokrivaju široku lepezu društvenih i političkih pitanja, a dodatno su otkriveni i razni događaji te teme vezane uz osobe i organizacije. Razlika u odnosu na prethodno istraživanje je veći broj označivača i veća količina grešaka u otkrivenim temama. Korak revizije tema pokazao se vrlo korisnim i rezultirao kvalitetnijim semantičkim temama. Izgradnja taksnonomije tema pokazala se kao odličan alat koji vodi do boljih uvida u medijsku agendu.

3.5.2 Postupak definicije tema

Nakon što je postupak otkrivanja tema pružio pregled medijske agende daljnja analiza usmjerena je na teme vezane uz parlamentarne izbore, preciznije na teme sadržane u dvije kategorije najviše razine u izgrađenoj taksonomiji: *izborni proces* i *pregovori nakon izbora*.

Razmatranjem semantičkih tema u tim kategorijama otkriveno je da se neke teme pojmovno preklapaju dok definicije nekih tema izostavljaju bitan sadržaj. Zbog toga je definiran novi skup

[7] Postoje tematski modeli koji modeliraju hijerarhijske odnose među temama, primjerice [61].

Tablica 3.5: Popis semantičkih tema i izvedenih novodefiniranih semantičkih tema iz odabranih tematskih kategorija najviše razine.

Kategorije	Novodefinirane teme	Semantičke teme	Opis
Izborni proces	**izborna matematika**	izborna predviđanja, izborne ankete, izborni rezultati	ankete, statistike i nagađanja, predviđanja, izlaznost, rezultati, saborska kombinatorika
	izborni propisi i postupci	izborni propisi i postupci, glasanje izvan prebivališta, izborna pravila i DIP, nepravilnosti	izborni kalendar, kandidature, promatranje izbora, Ivan Turudić, izborno povjerenstvo, sučeljavanja kandidata, etičko povjerenstvo, izborni propisi, nepravilnosti
	izborni programi i kampanja	ekonomski izborni program, medijsko praćenje izbora	izborni program, prepucavanje i komunikacija stranaka i političara
Pregovori nakon izbora	**pregovori**	stranački pregovori, raskol u Mostu	pregovori i zauzimanje pozicija, optužbe i prepucavanja, raskol u Mostu
	supstancijalni pregovori	stranački pregovori	reforma lokalne uprave, poduzetnička zona, ekonomske i fiskalne mjere
	imenovanje mandatara i konstituiranje Sabora	imenovanje mandatara, konzultacije kod predsjednice, konstituirajuća sjednica Sabora	pravne procedure i politički proces imenovanja madatara i konstituiranja Sabora

semantičkih tema koje se međusobno ne preklapaju i koje dobro pokrivaju koncepte od interesa za analizu izborne agende. Skup novodefiniranih tema sastoji se šest tema nastalih kombinacijom prethodno otkrivenih tema, uz iznimku teme *stranački pregovori*, koja je samo razdvojena u dvije novodefinirane teme *pregovori* i *supstancijalni pregovori* koje omogućavaju mjerenje finijih razlika u procesu pregovaranja nakon izbora. Lista novodefiniranih tema i vezanih semantičkih tema iz dvije odabrane kategorije nalazi se u tablici 3.5.

Novodefinirane semantičke teme na kojima se temelji daljnja analiza potvrđuju nužnost koraka definicije tema u društvenoznanstvenim analizama. Ove teme predstavljaju konceptualni alat nužan za analizu međutim niti jedna od njih nije bila precizno pogođena temama modela iz koraka otkrivanja. Štoviše, te teme nisu se dobro poklapale niti s tijekom otkrivanja definiranim semantičkim temama već su proizišle iz daljnje analize tih tema. Dok je kod analize agende američkih medija dio mjerenih tema bio detektiran temama modela, u ovom slučaju bi analiza koja slijedi bila nemoguća bez mogućnosti definicije novih tema.

3.5.3 Postupak mjerenja tema

Nakon što su definirane nove semantičke teme od interesa za analizu, na temelju metode predložene u 3.3.3 provedeno je mjerenje njihove zastupljenosti brojem članaka koji govore o temi. Za mjerenje tema korišten je pristup sa slabo nadziranim tematskim modelima. Prvi korak

Tablica 3.6: Liste visokodiskriminativnih riječi za novodefinirane semantičke teme.

Novodefinirana semantička tema	**Diskriminativne riječi**
izborna matematika	mandat, rezultat, anketa, pobjeda, glas, glasač, izlazni, preferencijalno, izlaznost, prednost, izborna, jedinica
izborni propisi i postupci	odbor, DIP, donacija, izvještaj, potrošiti, donirati, promidžba, GONG, financiranje, zakon, izborna šutnja, povrede, kampanja, debata, pritužba, promatrač
izborni programi i kampanja	ekonomski, program, PDV, obećanje, izborni, ukidanje, Prnjavor, demografski, navodnjavanje, dug
pregovori	Petrov, pregovori, Božo, Prgomet, sastanak, nestranački, nezavisni, Petrina, Drago, ključni, Grmoja, trojni, odgovor, podrška, sastaviti, pritisak
supstancijalni pregovori	reforma, lokalna, samoupravljanje, pojas, devalvacija, inflacija, racionalizacija, model, ukidanje, Lovrinović
imenovanje mandatara i konstituiranje Sabora	mandatar, potpis, konzultacije, osnivanje, Pantovčak, krug, sjednica, Reiner, konstituirajuća, sazivanje, izabran

mjerenja je izrada liste visokodiskriminativnih riječi za svaku semantičku temu. Na temelju tih riječi izgrađen je tematski model s apriornim vjerojatnostima definiranima na način da teme modela odgovaraju semantičkim temama. Tablica 3.6 sadrži diskriminativne riječi za definirane semantičke teme. Predloženi postupak izgradnje lista diskriminativnih riječi i u ovoj primjeni se pokazao efikasnim i izvedivim – za svaku od tema odabrano je 10-15 visokodiskriminativnih riječi za što je u prosjeku bilo potrebno 25 minuta po temi.

Mjerenje zastupljenosti tema, odnosno označavanje dokumenata temama provedeno je na temelju klasifikacijskog modela WSTM-JO pošto su eksperimenti mjerenja u 3.4 pokazali da taj model ima dobre performanse a vrijeme označavanja potrebno za izgradnju modela manje je nego u slučaju ostalih modela. Za WSTM-JO model označavanje se provodi označavanjem dokumenta temom s najvećom težinom za dokument. Broj tema modela fiksiran je na 100 tema a ostali parametri modela vezani uz definiciju apriornih vjerojatnosti mjerenih tema optimirani su iterativno, promatranjem tema dobivenog modela. Iterativna optimizacija se i u ovom se slučaju pokazala vremenski efikasnom i rezultirala je dobrim temama.

Izgrađeni klasifikacijski model nije vrednovan kvantitativno, na temelju skupa dokumenata označenih sa temama, nego je prije analize agende provedena prediktivna validacija koja je potvrdila smislenost modela. Dodatna potvrda kvalitete modela je činjenica da se zaključci analize poklapaju sa zaključcima iz drugih neovisnih društvenoznanstvenih studija.

Nakon izgradnje modela, mjera zastupljenosti svake od definiranih semantičkih tema izra-

čunata je na temelju skupa članaka označenih s tom temom. Korištena je zastupljenost tema definirana kao postotak članaka u zbirci, pošto je ta mjera procjenjena kao informativnija od ukupnog broja članaka koji govori o temi.

3.5.4 Analiza agende

Na temelju prethodno opisanih koraka otkrivanja i mjerenja, koji su rezultirali definicijom šest izbornih tema i mjerom zastupljenosti tih tema, provedena je analiza medijske agende u razdoblju izbora za Hrvatski sabor 2015. godine.

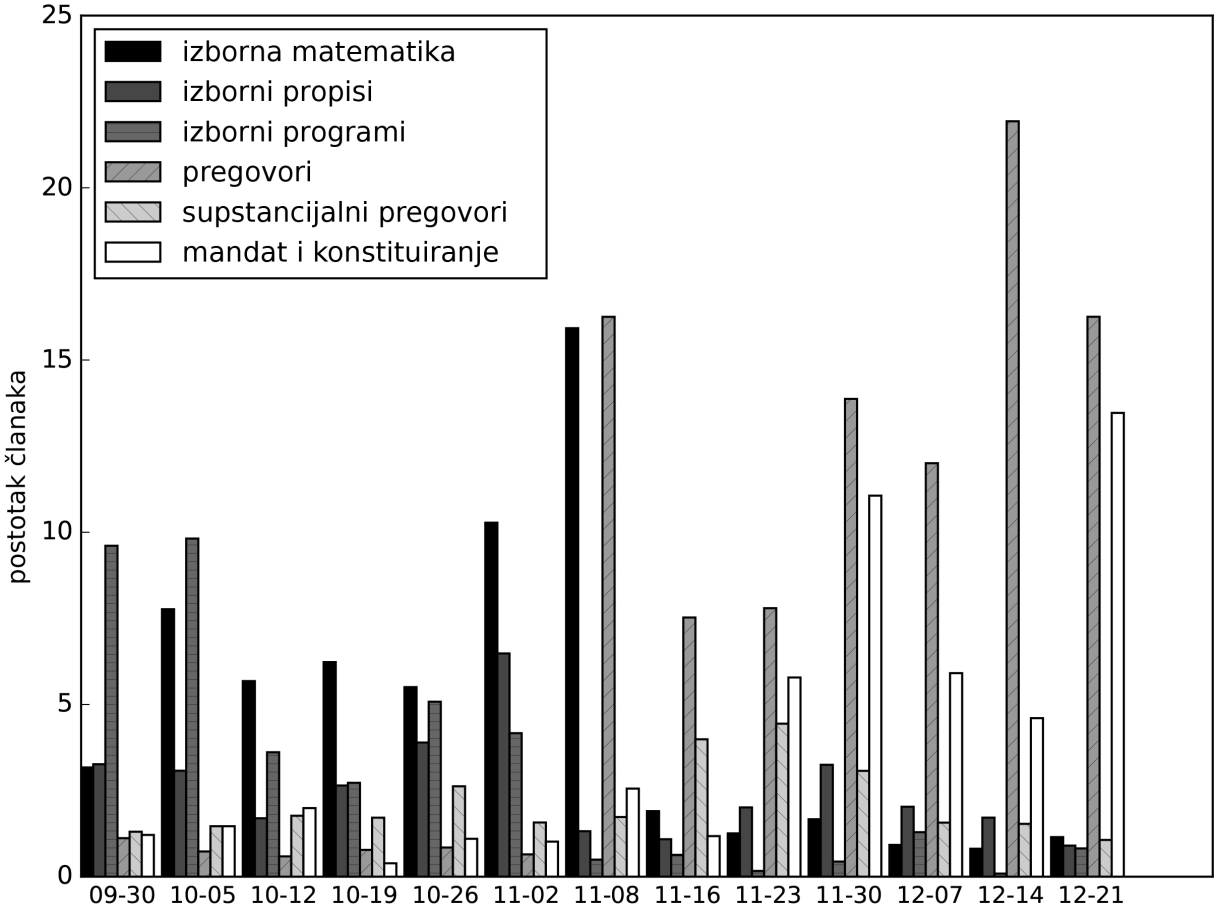
Prediktivna validacija tema Validacija postupka mjerenja tema provedena je standardom metodom prediktivne validacije [39, 41] odnosno analizom korelacije izmjerene zastupljenosti tema s događajima iz predizborne kampanje i poslijeizbornog razdoblja.

Slika 3.4 prikazuje relativnu zastupljenost šest mjerenih izbornih tema. Analizom je utvrđeno da zastupljenost tema (definirana kao postotak članaka u zbirci koji govore o temi) dobro korelira sa stvarnim događajima, što potvrđuje prediktivnu validnost [41] modela.

Primjerice, tema *izborna matematika* (koja se odnosi na rezultate anketa, predviđanja pobjednika i gubitnika, izborne rezultate itd.) dosta je zastupljena u tjednu koji je prethodio izborima, a u tjednu izbora (i na sam dan izbora 08.11.) njena zastupljenost je dodatno porasla.

Dodatna potvrda korelacijske validnosti modela može se naći razmatranjem postizbornih događaja. Pošto niti jedna stranka nije osigurala većinu potrebnu za formiranje Vlade, obje velike stranke – Socijaldemokratska partija Hrvatske (SDP) i Hrvatska demokratska zajednica (HDZ) – pokušale su pridobiti novu stranku Most koja je osvojila značajan broj zastupničkih mjesta. Pregovori među strankama privukli su značajnu medijsku pažnju, što je uspješno detektirano od strane modela. Nadalje, tjedna zastupljenost teme *pregovori* izmjerena modelom korelira s događajima koji su potaknuli interes za tu temu u javnom diskursu (dogovori i nesuglasice između stranaka, potraga za kandidatom za predsjednika Vlade čiji izbor bi osigurao potporu Mosta nekoj od velikih stranaka, itd.) Korelacija s događajima postoji i za temu *imenovanje mandatara i konstituiranje Sabora*, koja je najzastupljenija u danima koji su prethodili konstituirajućoj sjednici Sabora i formiranju Vlade.

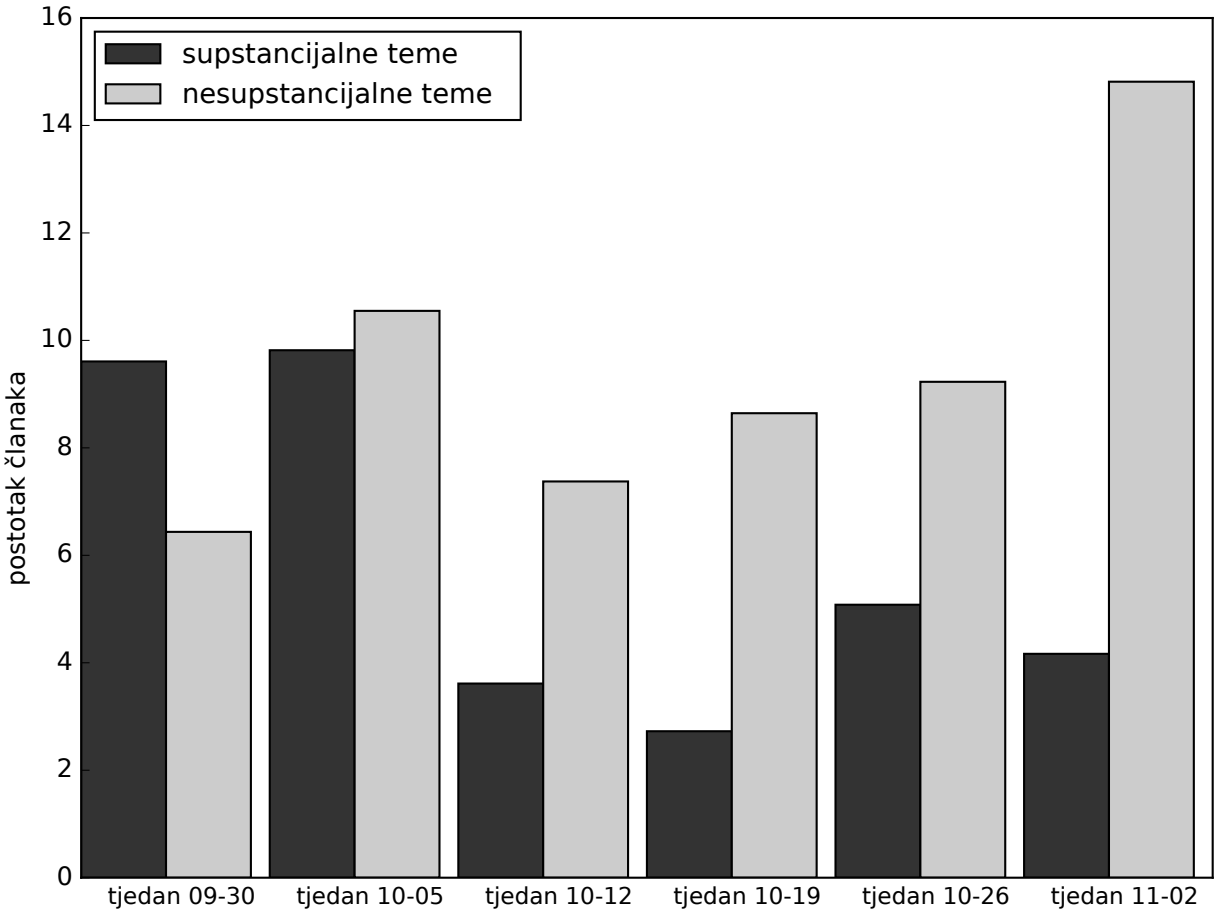
Medijsko izvještavanje Nakon validacije, provedena je analiza podataka iz perspektive političke komunikacije. Na temelju poznate razlike između supstancijalnog i manje supstancijalog izvještavanja o izborima (vidi primjerice [89]), semantičke teme podjeljene su u "supstancijalne" i "nesupstancijalne". U predizbornom razdoblju teme *izborna matematika* i *izborni propisi i postupci* su označene kao nesupstancijalne, dok je tema *izborni programi i kampanja* označena kao supstancijalna. U postizbornom razdoblju, napravljena je razlika između nesupstancijalnih tema vezanih uz pregovore (poput sukoba i dogovora među strankama) te pregovora



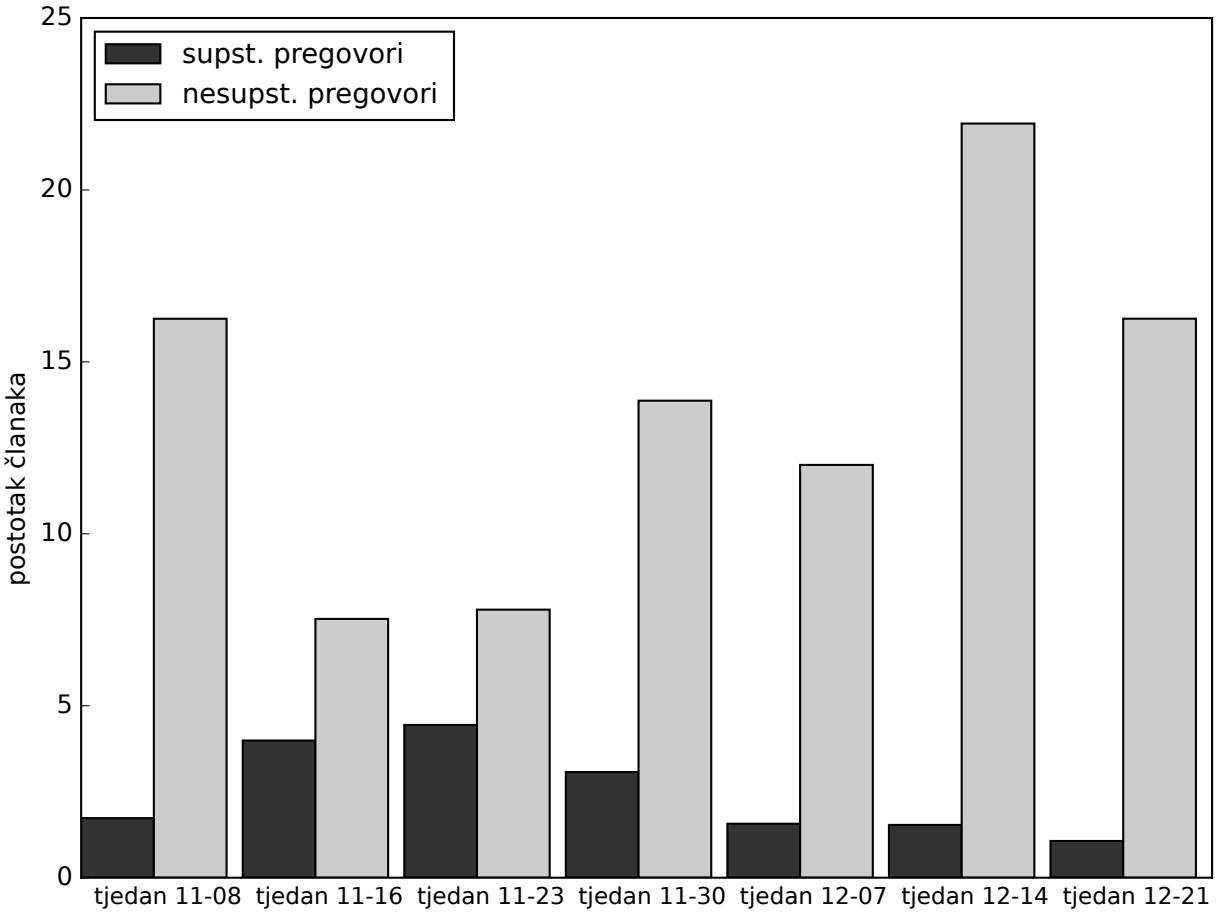
Slika 3.4: Tjedna zastupljenost šest mjerenih izbornih tema tijekom predizborne kampanje i nakon izbora. Svaki tjedan je označen odgovarajućim mjesecom i danom.

koji su se bavili supstancijalnim političkim pitanjima (vidi tablicu 3.5). Slike 3.5 i 3.6 prikazuju tjednu zastupljenost tih tema u predizbornom i postizbornom razdoblju.

Može se jasno vidjeti da u predizbornom razdoblju nesupstancijalni sadržaj prevladava nad supstancijalnim. To se prije svega odnosi na visoku zastupljenost članaka koji se bave pitanjima predizborne "utrke" (npr. ankete, usporedba rejtinga, predviđanje rezultata) ili pokrivaju izbornu kampanju iz perspektive spektakla. Ti članci stoje u suprotnosti sa člancima koji pokrivaju izborne programe. Očekivano, nesrazmjer između supstancijalnog i nesupstancijalnog sadržaja se povećava približavanjem izbora. Zanimljivo je, kao što se može vidjeti na slici 3.6, da je ovaj nesrazmjer još izraženiji u postizbornom razdoblju, što ukazuje da su se tijekom pregovora mediji više bavili političkim pogodbama među strankama nego supstancijalnim sadržajem pregovora. Pitanje da li je ovaj nesrazmjer posljedica medijske usredotočenosti na izborni spektakl ili je rezultat toga što se političari nisu bavili supstancijalnim pitanjima nije u fokusu ovog istraživanja. Može se zaključiti da analiza pokazuje da u medijskom izbornom izvještavanju prevladava tzv. perspektiva igre (engl. *game schema*) a ne supstancijalna politička pitanja, što je pojava koja je zamijećena u nizu zemalja [89, 90, 91].



Slika 3.5: Tjedna zastupljenost supstancijalnih i nesupstancijalnih političkih tema u predizbornom razdoblju. Svaki tjedan je označen odgovarajućim mjesecom i danom.



Slika 3.6: Tjedna zastupljenost supstancijalnih i nesupstancijalnih tema vezanih uz pregovore nakon izbora. Svaki tjedan je označen odgovarajućim mjesecom i danom.

3.6 Uočeni nedostaci metoda vrednovanja

Kroz opisane primjene predložene metode otkrivanja tema na medijskoj agendi na analizu dviju zbirki medijskih tekstova uočeni su nedostaci postojećih metoda automatskog vrednovanja intrinzične semantičke kvalitete tema te potreba za metodama vrednovanja modela s aspekta broja otkrivenih tema i relevantnosti tih tema za provođenje analize. Ovi problemi javljaju se i u ranijim istraživanjima analize agende koja koriste tematske modele. Ovdje se ti problemi razmatraju temeljem iskustava i podataka iz prethodno opisanih analiza te se kao korak prema njihovom rješavanju predlažu daljnji smjerovi istraživanja.

3.6.1 Potreba za mjerom intrinzične semantičke kvalitete tema

Najvažniji aspekt semantičke validacije tema je njihova intrinzična semantička validacija s ciljem procjene interpretabilnosti tema. Glavna metoda procjene kvalitete tema je promatranje tema, dok se automatske mjere koherentnosti koriste u malom broju istraživanja, kao smjernica za odabir broja tema [49, 50]. Uspješna primjena automatskih metoda mogla bi rezultirati automatskim odabirom kvalitetnijih modela, kao i bržim postupkom pregleda tema kao rezultatom filtriranja loših tema modela.

Tijekom dviju provedenih analiza medijske agende tijekom koraka otkrivanja tema pregledano je ukupno 550 tema naučenih tematskih modela. Kvaliteta tema varira od jasno interpretabilnih tema preko tema koje sadrže veću ili manju količinu šuma ili odgovaraju dvama različitim ali prepoznatljivim konceptima do neinterpretabilnih i semantički praznih koje su beskorisne za analizu agende. Iako broj nekvalitetnih tema ovisi i o samoj metodi tematskog modeliranja, takve teme je nemoguće izbjeći zbog aproksimativne prirode i inherentne stohastičnosti modela. Za teme modela naučene na američkim novinskim tekstovima, 189 tema (54%) je interpretabilna, 121 tema (34.6%) je interpretabilna uz veću ili manju količinu šuma, dok 40 tema (11.4%) odgovara šumu ili je na drugi način beskorisno. Na pregled i interpretaciju jedne teme u prosjeku je potrošeno što u prosjeku iznosi 6 minuta po temi, iz čega slijedi da je na neinterpretabilne teme utrošeno 4 čovjek-sata dok je na analizu djelomično interpretabilnih tema (očekivano sporiju od analize kvalitetnih tema) utrošeno 8 čovjek-sati.

Primjena automatskih mjera kvalitete tema mogla bi dovesti do kvalitetnijih modela pri njihovoj izgradnji – učenjem većeg broja (nekoliko desetaka) modela i odabirom najkvalitetnijih. Kod samog pregleda tema, mjere kvalitete mogle bi omogućiti bržu detekciju nekvalitetnih tema. Zadnja primjena mogla bi koristiti upravo za ovdje predloženu metodu otkrivanja koja za otkrivanje agende koristi veći broj tematskih modela, među ostalim kako bi se izbjegao postupka odabira broja tema, što rezultira većim brojem tema koje treba obraditi.

Međutim, početeni eksperimenti s pogodnošću postojećih mjera tematske koherentnosti ukazuju na nepogodnost tih mjera za detekciju interpretabilnosti tema. Razmotrene su mjere

koherentosti izdvojene kao najbolje u evaluaciji niza srodnih mjera na nekoliko skupova podataka [92] – NPMI mjera (engl. *normalized pointwise mutual information*) popularna u eksperimentima evaluacije tematskih modela te C_V mjera predložena u [92]. Preliminarno vrednovanje mjera provedeno je promatranjem korelacije tih mjera s interpretabilnošću tema. Tijekom provodenog postupka otkrivanja tema teme modela označavane su s jednom ili više semantičkih tema te s oznakama šuma. Na temelju ovih podataka svaku temu modela može se svrstati u jednu od četiri klase – interpretabilne teme koje dogovaraju konceptu (jednoj semantičkoj temi) sa ili bez šuma, teme koje su mješavina dva koncepta, te beskorisne teme označene samo sa šumom. Slike 3.7 i 3.8 prikazuje raspon vrijednosti dviju mjera tematske koherentnosti za svaku od prethodno opisanih klasa. Kao što je vidljivo iz slike, vrijednosti mjera koherentnosti slabo se ili nikako razlikuju od klase do klase, što upućuje na nisku korelacije između mjera koherentnosti i semantičke kvalitete.

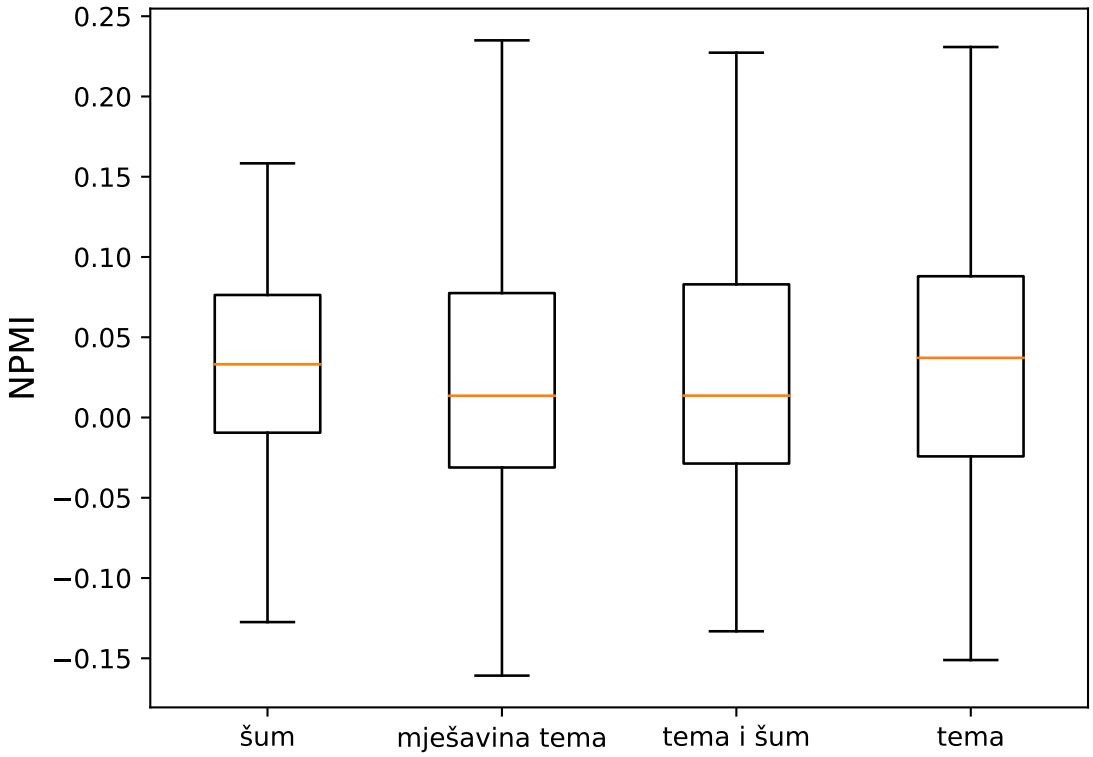
Vjerojatni razlog tome je činjenica da ove mjere ocjenjuju koherentnost tema na temelju uz temu vezanih riječi. No u slučaju medijskih tema, uz teme vezane riječi su u velikom broju slučajeva semantički slabo povezane i neinterpretabilne, što je posljedica velikog broja prolaznih tema koje su blisko vezane uz događaje i priče. Takvi skupovi riječi povezani su u kontekstu određenog događaja, no za razliku od skupova riječi koje opisuju apstraktne koncepte, bez poznavanja tog konteksta djeluju neinterpretabilno. Za razliku od riječi, uz teme vezani dokumenti pogodni su za procjenu kvalitete tema i njihovu interpretaciju. Naime, medijske teme koje odgovaraju semantičkim temama karakterizirane su skupom srodnih dokumenata o istim događajima, pričama ili osobama, koji su u pravilu međusobno slični, tematski fokusirani i moguće ih je brzo interpretirati.

Na temelju navedenih statistika i iskustava iz provedenih postupaka otkrivanja tema, u 4 se predlaže se nova klasa mjera koherentosti koje računaju koherentnost tema na temelju uz temu vezanih dokumenata. Provodi se detaljna kvantitativna validacija novopredloženih i postojećih mjera koherentnosti koja potvrđuje kvalitetu predloženih mjera. Nadalje, u poglavlju 6 predlažu se metode primjene mjera koherentosti tema na ubrzanje koraka otkrivanja medijske agende.

3.6.2 Potreba za mjerom pokrivenosti tema

Glavna primjena tematskih modela na analizu medijske agende je otkrivanje tema koje se provodi interpretacijom tema modela. Teme modela očekivano odgovaraju semantičkim temama – konceptima koji analitičaru pružaju relevantne informacije o tematskoj strukturi zbirke tekstova. Ova eksploratorna primjena tematskih modela, koja se može kretati od brze pregledne analize zbirke do podrobne kvalitativne analize tema, važna je neovisno o domeni primjene.

Prirodno i važno pitanje koje se nameće u kontekstu te primjene je koliki udio od ukupnog broja tema model ili skup modela može pokriti. Neovisno o primjeni, poželjni su modeli koji mogu otkriti, odnosno pokriti veći udio ukupnog broja tema. Unatoč tome što neki eksperi-

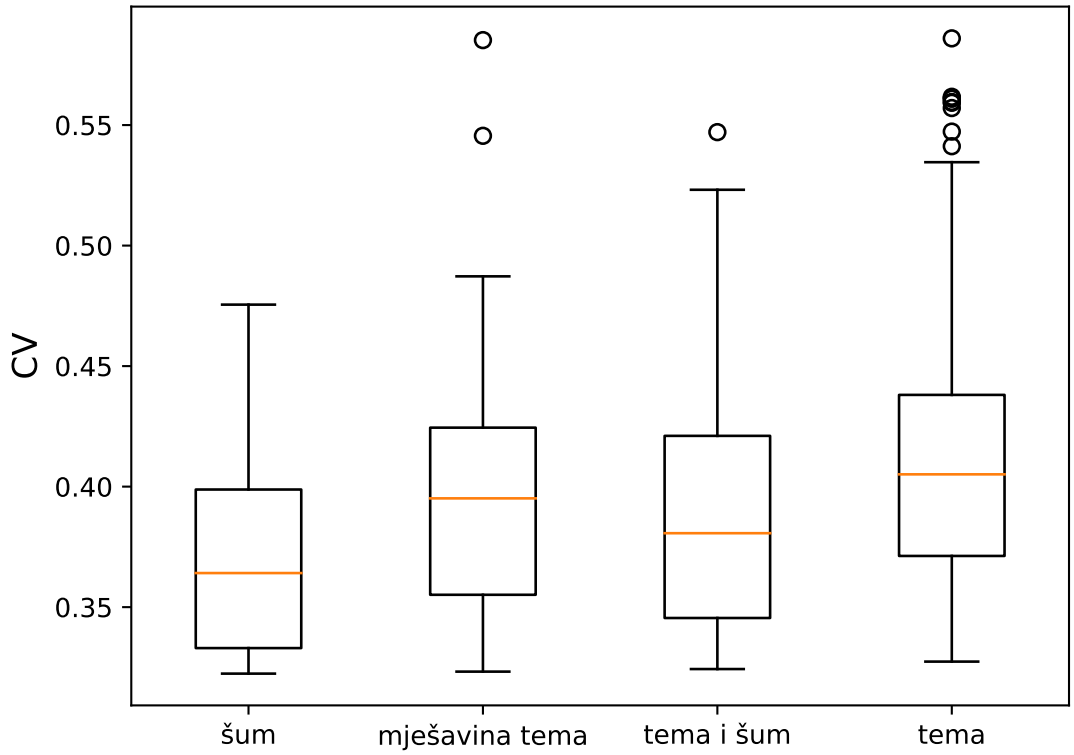


Slika 3.7: Razdioba vrijednosti mjere NPMI za svaku od četiri klase interpretabilnosti tema.

menti s tematskim modelima dotiču ove problematike te ukazuju na to da neki modeli pokrivaju samo manji broj tema [29, 64] ovaj problem nije sustavno obrađen. Vezani problem, važan u kontekstu društvenoznanstvenih analiza teksta, odnosi se na relevantnost tema za istraživanje odnosno pitanje mogu li modeli pokriti teme od interesa za istraživača [40, 41, 48].

U kontekstu metoda za otkrivanje tema na medijskoj agendi, kada bi se pokrivenost modela mogla mjeriti, to bi omogućilo da se izdvoje i primjene modeli koji daju visoku pokrivenost tema, što bi očekivano dovelo do boljeg pregleda medijske agende.

Prethodno provedeni eksperimenti s otkrivanjem tema omogućavaju preliminarnu analizu problematike pokrivenosti tema. Rezultat predloženog postupka otkrivanja je tablica semantičkih tema koje su mapirane na odgovarajuće teme modela. Na temelju tih podataka moguće je izračunati broj semantičkih tema otkrivenih pregledom pojedinih modela ili skupova modela. Slike 3.9 i 3.10 prikazuju rast broja otkrivenih semantičkih tema u ovisnosti o pregledanim modelima. Slike 3.9 prikazuje scenarij pregleda prvo modela sa 50 a zatim modela sa 100 tema. Može se vidjeti da pregled svakog modela rezultira otkrivanjem novih semantičkih tema, kao i da se pregledom samo jednog modela LDA sa 50 tema (često korišten model i broju tema) pokriva manje od polovice od ukupno 134 semantičke teme otrkivene svim modelima. Slike 3.10 prikazuje scenarij pregleda prvo modela sa 100 a zatim modela sa 50 tema. Usporeda uka-

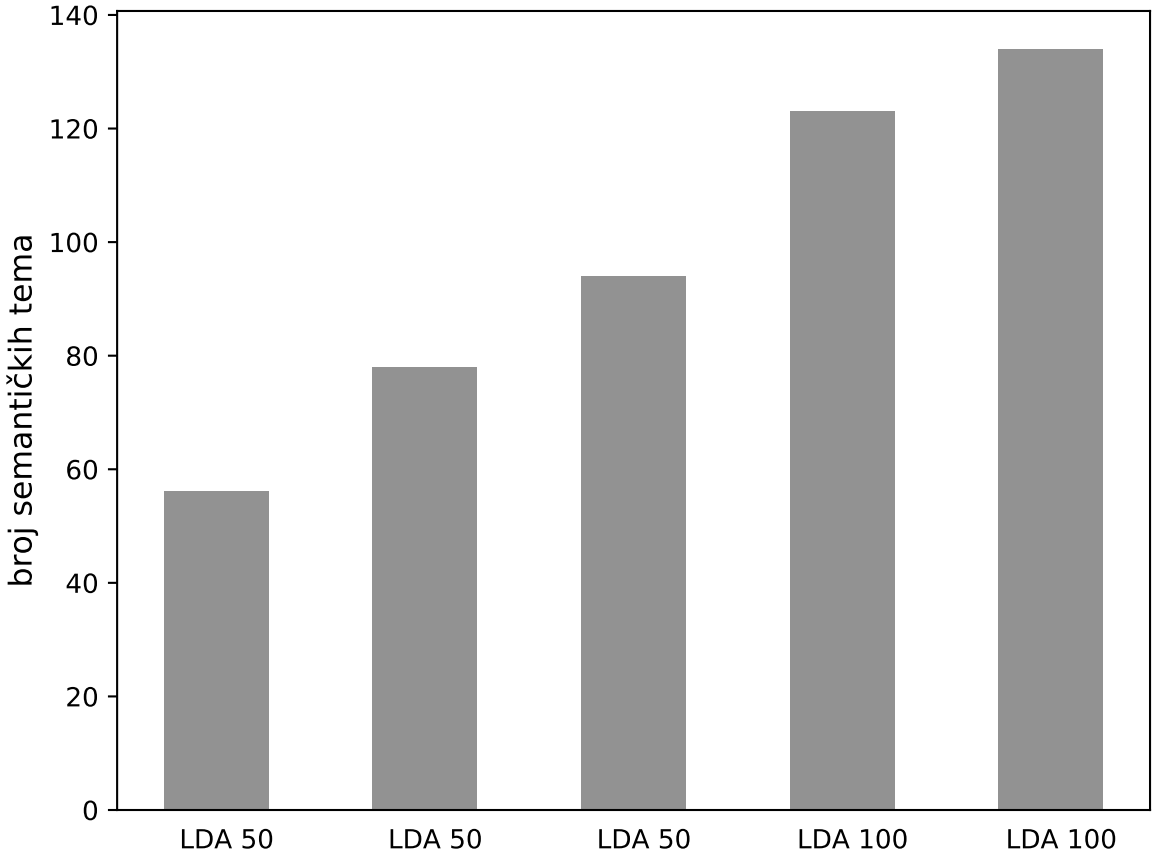


Slika 3.8: Razdioba vrijednosti mjere C_V za svaku od četiri klase interpretabilnosti tema.

zuje da, očekivano, modeli sa 100 tema otkrivaju veći broj semantičkih tema te da je vjerojatno bolje, ukoliko dostupno vrijeme dopušta pregled samo 100 tema, pregledati jedan model sa 100 tema nego dva modela sa 50 tema. Kvalitativna analiza semantičkih tema koje novopregledani modeli dodaju u skup otkrivenih tema otkriva da se radi o temama mogu biti od interesa za analitičara. Primjerice, neke od tema novootkrivenih zadnjim modelom na slici 3.9 su: *financije*, *Chapel Hill napad*, *Medicare zdravstveno osiguranje* i *znanost*.

Opisane statistike kretanja broja otkrivenih tema u odnosu na broj pregledanih modela pokazuju smislenost predloženog pristupa otkrivanju medijske agende na temelju veće broja modela – upotreba samo jednog modela analitičaru ne otkriva značajan broj potencijalno važnih tema.

Problem pokrivenosti tema važan je i neistražen problem a metode vrednovanja pokrivenosti mogle bi dovesti do odabira modela koji omogućuju efikasnije otkrivanje tema na medijskog agendi. Stoga se u poglavlju 5 sustavno pristupa problemu pokrivenosti tema. Predlaže se metoda koja izrade skupa referentih tema čija pokrivenost se mjeri, i metode mjerenja pokrivenosti referentnih tema od strane tematskih modela, te se provodi vrednovanje predloženih metoda i vrednovanje pokrivenosti tematskih modela. U poglavlju 6 opisane su primjene tih novih rezultata na poboljšanje metode za otkrivanje medijske agende.

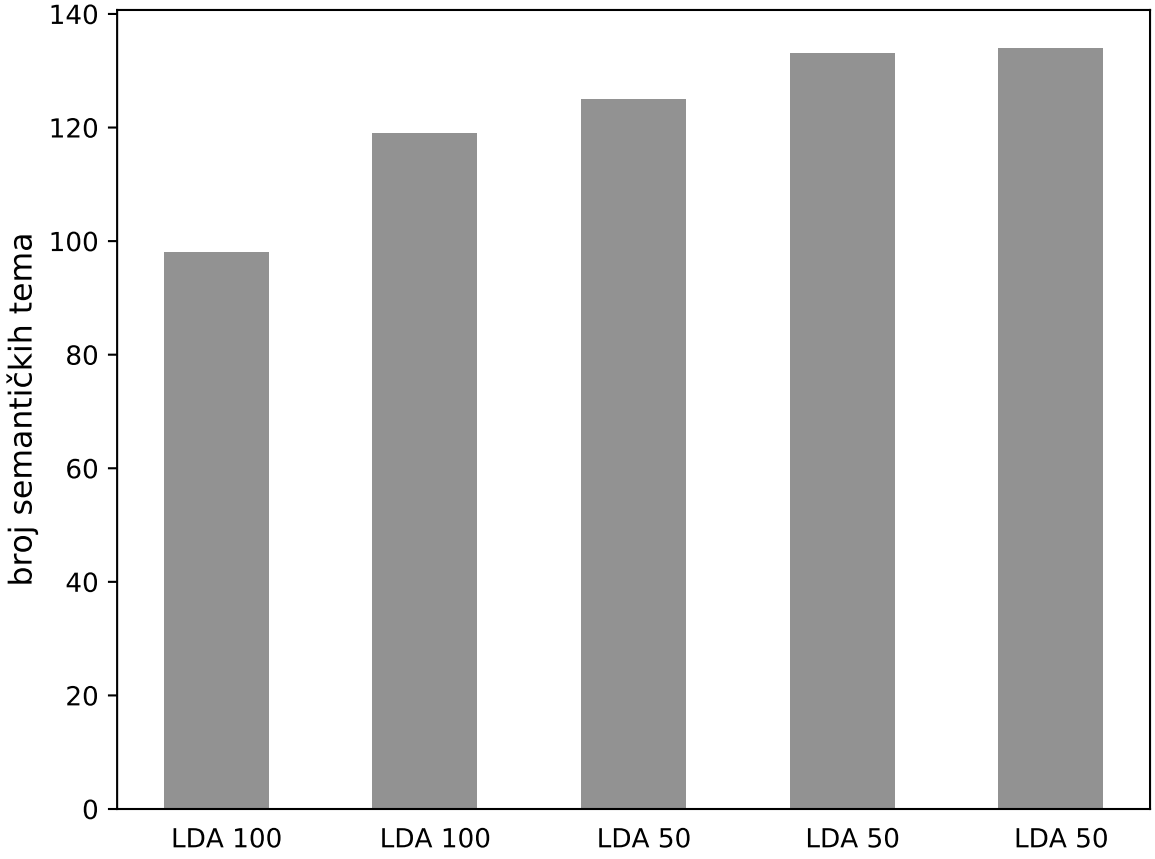


Slika 3.9: Kumulativni broj otkrivenih semantičkih tema u ovisnosti o pregledanim LDA tematskim modelima. Prvo su pregledani modela sa 50 tema, a zatim modeli sa 100 tema.

3.7 Rasprava

U ovom poglavlju analizira se zadatak analize medijske agende koji se sastoji od zadataka otkrivanja tema o kojima mediji govore i mjerenja zastupljenosti tih tema. Zatim se predlaže eksperimentalna metoda analize medijske agende motivirana otklanjanjem nedostataka postojećih pristupa te se opisuju primjene i vrednovanja predložene metode.

Popularni modeli strojnog učenja za analizu medijske agende su tematski modeli, klasa nenadziranih modela pogodna za otkrivanje tematske strukture teksta. Uočena su tri nedostatka postojećih pristupa analizi agende temeljenih na tematskim modelima – upotreba samo jednog modela za otkrivanje tema, nemogućnost definicije i analize novih tema od interesa za analitičara te izostanak preciznog vrednovanja mjerenja zastupljenosti tema. Predlaže se eksperimentalna metoda analize medijske agende koja adresira uočene nedostatke upotrebom većeg broja modela u koraku otkrivanja, omogućavanjem definicije i mjerenja novih tema, te preciznom formulacijom mjerenja tema kao zadatka označavanja dokumenata s mjerenim temama. Osim bavljenja opisanim nedostacima predložena metoda ujedno sistematizira i detaljno pristupa svim koracima postupka analize agende. Predložena metoda ispituje se kroz primjene u



Slika 3.10: Kumulativni broj otkrivenih semantičkih tema u ovisnosti o pregledanim LDA tematskim modelima. Prvo su pregledani modela sa 100 tema, a zatim modeli sa 50 tema.

dvije studije slučaja istraživanja medijske agende.

Predloženi postupak otkrivanja tema koristi veći broj tematskih modela, a pregledom i interpretacijom tema modela definira se skup semantičkih tema za koje se vodi precizna evidencija povezanosti s temama modela. Provedeni postupci otkrivanja tema potvrđuju da predloženi pristup daje dobar pregled tema na medijskoj agendi i otkriva teme od interesa. Analiza broja otkrivenih semantičkih tema u odnosu na pregledane teme modela jasno pokazuje da je upotreba većeg broja modela dovodi do otkrivanja većeg broja smislenih semantičkih tema. S druge strane, korak otkrivanja tema vremenski je zahtjevan pošto je potrebno pregledati velik broj tema dobivenih većim brojem modela. Primjena na analizu hrvatske izborne medijske agende pokazuje potrebu za pažljivom obukom ljudskih označivača koji provode postupak otkrivanja.

Korak definicije tema uveden da bi se analitičaru omogućila definicija i analiza tema od interesa pokazao se nužnim pošto su u obadvije primjene uočene i definirane korisne teme za koje ne postoje odgovarajuće automatski naučene tema modela.

Mjerenje tema provodi se označavanjem dokumenata s temama. Označavanje dokumenata temama definira se kao problem klasifikacije (dokumenata) s višestrukim oznakama (temama), formulacija koja omogućuje preciznu kvantitativnu validaciju mjerenja. Predlaže se metoda

mjerenja temeljena na slabo nadziranim tematskim modelima i konstrukciji skupa visokodiskriminativnih riječi za svaku mjerenu temu. Razmatra se i metoda mjerenja temeljena na nadziranoj metodi binarne relevantnosti u kombinaciji s binarnim SVM klasifikatorom. Vrednovanja mjerenja provedena za teme s američke medijske agende ukazuju na pogodnost slabonadziranih tematskih modela za mjerenje tema na medijskoj agendi. Slabonadzirani modeli rade usporedivo ili bolje od nadziranog rješenja, pri čemu se bolje performanse postižu uz usporedivu količinu ljudskog vremena utrošenu u izgradnju modela, dok se usporedive performanse postižu slabonadziranim modelom izgrađenim na temelju značajno manje utrošenog vremena.

Vrednovanje performansi za pojedinačne klase (teme) ukazuje na znatno bolje performanse slabonadziranih tematskih modela na manje zastupljenim temama. Najbolji slabonadzirani model također bolje od nadziranog procjenjuje udjele zastupljenosti tema. I kvalitativna validacija mjerenja zastupljenosti slabonadziranim modelima potvrđuje njihove performanse – u obje studije slučaja zastupljenost na očekivani način korelira sa stvarnim događajima dok su u analizi hrvatske izborne agende zaključci izvedeni iz zastupljenosti tema usklađeni za zaključcima srodnih istraživanja. Na temelju provedenih eksperimenata, varijanta slabonadziranih modela s jednostrukim označavanjem i iterativnom optimizacijom parametara predlaže se kao najbolji izbor za mjerenje tema.

Iako tematski modeli omogućuju efikasnu analizu velikih količina teksta, nužno je validirati modelima dobivene podatke. Validacija tematskih modela je važan i nedovoljno istražen problem, posebno naglašen u znanstvenim analizama teksta gdje je izražena potreba za kvalitetnim modelima. Na temelju provedenih studija slučaja i analize podataka o otkrivenim temama modela izdvojena su dva uočena problema vezana uz validaciju modela. Prvi je problem automatskog mjerenja intrinzične semantičke kvalitete (interpretabilnosti) tema modela. S jedne strane analize i iskustva s promatranjem tema ukazuju na neadekvatnost postojećih mjera, a s druge strane ukazuju na potrebu primjene ovakve mjere na poboljšanje postupka otkrivanja tema.

Drugi problem je problem pokrivenosti tema, odnosno analize načina na koji tematski modeli pokrivaju semantičke teme, za što unatoč eksperimentima koji ukazuju na ovaj problem nije razvijena nikakava metoda. Metode mjerenja pokrivenosti mogle bi dovesti do efikasnijeg otkrivanja tema, odnosno boljih modela koji uče veći broj potencijalno korisnih semantičkih tema. Problemi pokrivenost i mjerenja interpretabilnosti, iako motivirani primjenom tematskih modela na analizu medijsku agendu, nadilaze jednu tekstnu domenu primjene i predstavljaju općenite probleme semantičke validacije tematskih modela.

Glavni zaključak vrednovanja predloženih metoda mjerenja medijske agende jest pogodnost slabo nadziranih tematskih modela za tu primjenu – takvi modeli mogu smanjiti potrebno vrijeme označavanja i dati dobre performanse, osobito za slabozastupljene teme. Međutim predložena metoda, iako daje dokaz koncepta, je rudimentarna u odnosu na najsofisticiranije do sada predložene pristupe – koristi generički model LDA umjesto modela s problemu prilagođenom

strukturom a proces izgradnje modela zahtjeva ugađanje više parametara. Stoga postoji dosta prostora za poboljšanje performansi mjerenja i ubrzanje postupka izgradnje modela oblikovanjem modela s prilagođenom strukturom ili preuzimanjem neke do postojećih arhitektura.

S druge strane, postojeći eksperimenti pokazuju da predloženi pristup radi usporedivo ili bolje u odnosu na naziranu metodu, dok ostali slabo nadzirani tematski modeli zaostaju za nadziranim metodama . Razlog tome vjerojatno leži u najvažnijoj razlici između predloženog i ranijih pristupa – predloženi pristup koristi elaboriranu metodu odabira visokodiskriminativnih riječi koje dovode do boljih performansi. Ovi rezultati ukazuju da bi kod slabonadziranih pristupa trebalo poboljšati izgradnju skupa riječi. Smatramo da se predloženi postupak izrade visokodiskriminativnih riječi može dodatno ubrzati pomoću automatskih mjera kvalitete riječi i vizualizacijskih alata.

Trebalo bi razmotriti i rješavanje problema pridruživanja tema primjenom drugih nadziranih metoda te osobito metodama polunadziranog i aktivnog učenja. Aktivno učenje bi se moglo pokazati izglednim pristupom koji bi mogao riješiti problem slabih performansi nadziranih pristupa na slabo zastupljenim temama ciljanim dohvatom dokumenata koji predstavljaju te teme. Još jedan potencijalno izgledan pristup je kombiniranje slabonadziranih tematskih modela i nazdiranih pristupa uključivanjem oznaka klasa u strukturu modela, te korištenje aktivnog učenja u kombinaciji sa slabonadziranim modelima.

Poglavlje 4

Mjere koherentnosti tema temeljene na dokumentima

4.1 Uvod i motivacija

Primjena tematskih modela na otkrivanje tema koje čine medijsku agendu, opisana u poglavlju 3, rezultirala je velikim skupom naučenih tema modela koje su pregledane, interpretirane i označene s odgovarajućim semantičkim temama tj. konceptima. Ove analize medijske agende ukazale su na potrebu za automatskim mjerama kvalitete tema pogodnim za filtriranje loših tema radi ubrzanja postupka otkrivanja. Dobri kandidati su mjere koherentnosti [35, 92], klasa mjera koja ocjenjuje semantičku koherentnost i interpretabilnost tema. Međutim, analiza tema modela opisana u 3.6.1 ukazala je na slabu korelaciju između ocjena tematske koherentnosti i kvalitete mjerene poklapanjem tema modela sa semantičkim temama. Na temelju ove analize i opažanja iz postupka pregleda i interpretacije tema modela, ovdje se predlažu nove mjere semantičke kvalitete pogodne za teme iz medijske domene. Predložene mjere ocjenjuju koherentnost tema na temelju vezanih dokumenata umjesto riječi [93].

Postojeće mjere koherentnosti tema, popularna klasa mjera za intrizničnu semantičku validaciju tema, računaju koherentnost na temelju skupa najbolje rangiranih riječi za temu. No u slučaju medijskih tekstova, javlja se velik broj uz događaje vezanih prolaznih tema koje su karakterizirane semantički nepovezanim i neinterpretabilnim skupovima riječi. Ova pojava je opažena tijekom pregleda i interpretacija većeg broja tema naučenih iz zbirki medijskih tekstova. Međutim, pregled tema također je pokazao da je medijske teme u pravilu lako interpretirati na temelju vezanih tekstova. Ovo ukazuje na mogućnost procjene koherentnosti tema na temelju tekstova s vijestima koji su tematski fokusirani, što znači da su tekstovi vezani uz semantički koherentnu temu očekivano međusobno slični.

Predložene mjere dokumentne koherentosti [93] motivirane su eksperimentima otkrivanja tema koje čine medijsku agendu i potrebom za ubrzanjem otkrivanja odabirom kvalitetnijh tema

i modela. Međutim, ove mjere mogu se iskoristiti i u nizu drugih primjena tematskih modela na važnu domenu medijskih tekstova. Nedavno istraživanje je pokazalo da vijesti distribuirane putem interneta, odnosno putem medijskih web sjedišta, agregatora vijesti, web tražilica, društvenih mreža i aplikacija za razmjenu poruka – predstavljaju većinu konzumiranih vijesti [94]. Štoviše, veći broj korisnika dolazi do vijesti putem algoritama nego preko novinara i urednika [94]. Trend povećanja konzumacije tekstnih vijesti putem interneta popraćen je rastućim interesom za primjenom metoda strojnog učenja i obrade prirodnog jezika na automatsku analizu medijskih tekstova. Ove tehnologije omogućavaju korisnicima ekstrakciju korisnih informacija iz velikih količina medijskih tekstova i nalaze razne primjene poput personalizacije vijesti za krajnje konzumente [95, 96, 97] te podrške proizvodnji i širenju vijesti [98, 99]. Tematski modeli, popularna klasa modela strojnog učenja s mnogim primjenama, mogu se na mnogo načina iskoristiti u analizi medijskog teksta. Pregled medijskih primjena tematskih modela dan je u odjeljku 4.2. Metode dokumentne koherentnosti, oblikovane upravo za vrednovanje automatski naučenih medijskih tema, mogle bi omogućiti odabir kvalitetnijih tematskih modela u svakoj od tih primjena.

Kao što je ranije sažeto opisano, metode dokumentne koherentnosti motivirane su prirodom medijskih tema i prirodom skupova riječi koji ih karakteriziraju. Iako je pretpostavka o visokoj korelaciji između semantičke koherentnosti teme i koherentnosti skupa najbolje rangiranih riječi za temu razumna a i točna u dosta slučajeva, pokazuje se da ipak predstavlja samo dio koncepta tematske koherentnosti. Konkretno, težine riječi za teme samo su dio informacija o temama koje model sadrži, a postoje teme koje je teško interpretirati samo na temelju riječi. Primjeri koji to ilustriraju su teme 3 i 4 iz Tablice 4.1. Za razliku od "Ekonomije" i "Sporta", dviju općenitih i apstraktnih tema, teme 3 i 4 su *prolazne*[1], što je svojstvo tipično za teme u zbirci medijskih tekstova. Na razini riječi ove se teme čine nekoherentnima, pošto su riječi semantički slabo povezane. No važna i u dosadašnjim pristupima zanemarena informacija o temama su uz teme vezani dokumenti. Primjerice, ako osoba koja analizira teme pregleda dokumente (novinske tekstove) vezane uz teme 3 i 4 iz Tablice 4.1, vrlo vjerojatno će uočiti da su teme koherentne i interpretirati ih kao "Zatvaranje DHS-a"[2] i "ISIL ratna autorizacija".[3] Ovi primjeri pokazuju da za prolazne teme koje odgovaraju novinskim pričama i događajima, riječi vezane uz temu ne pružaju dovoljno informacija za procjenu tematske koherentnosti.

U takvim slučajevima koherentnost tema se često može lako procijeniti na temelju uz temu vezanih dokumenata. Naime, u slučaju medijskih tema, uz teme vezani dokumenti u pravilu su tematski fokusirani pošto se radi o člancima koji najčešće opisuju jedan događaj, priču, osobu

[1]Termin "prolazna tema" nadalje se koristi za ovakve i slične teme, koje ili izravno odgovaraju događajima ili se njihova definicija temelji na događajima, što primjerice može biti slučaj i kod tema koje odgovaraju osobama i organizacijama.

[2]DHS označava Ministarstvo domovinske sigurnosti SAD-a.

[3]Radi se o debati o davanju ovlasti za rat protiv ISIL-a predsjedniku SAD-a.

Tablica 4.1: Primjeri tema naučenih na temelju zbirke američkih političkih vijesti iz odjeljka 3.4 koja sadrži približno 24.000 članaka s početka 2015. godine. Svaka tema je opisana s deset najbolje rangiranih riječi za temu. Teme su označene na temelju pregleda najbolje rangiranih dokumenata za temu.

	Oznaka teme	10 najbolje rangiranih riječi
1.	Ekonomija	stopa, ekonomija, rast, FED, pad, nisko, tržište, rezerva, cijena, nezaposlenost
2.	Sport	momčad, igra, igrači, sezona, sportovi, liga, navijači, nogomet, pehar, izabrati
3.	Zatvaranje DHS-a	Boehner, domovinska, blokirati, DHS, McConnell, proći, ilegalni, govornik, granica, deportacija
4.	ISIL ratna autorizacija	razlog, veto, rezolucija, Corker, latino, Bob, nacrt, recenzija, Capitol, proći
5.	*(šum)*	članak, George, životinje, Richard, pas, obećanje, era, ništa, ponekad, suradnja

ili organizaciju. To ne znači da se u člancima ne pojavljuju druge teme, no u pravilu dominira jedna tema dok je broj ostalih značajnije zastupljenih tema očekivano malen. Kod drugih vrsta teksta, primjerice u znanstvenim članicma, očekivano se pojavljuje mnogo veći broj ravnopravnije zastupljenih tema. Posljedica ove pojave je da su interpretabilne teme, koje odgovaraju semantičkim temama, karakterizirane skupom visoko sličnih dokumenata. Nekvalitetne teme koje sadrže mnogo šuma ili spajaju veći broj semantičkih tema očekivano su povezane sa skupom međusobno manje sličnih dokumenata. Mjere dokumentne koherentnosti temelje se na pretpostavci da je moguće iskoristiti sličnost uz temu vezanih dokumenata za karakterizaciju kvalitetnih tema.

Na temelju prethodnih razmatranja i opažanja, u ovom poglavlju se predlažu *mjere koherentnosti tema temeljene na dokumentima* koja predstavljaju alternativu dosadašnjim mjerama koherentnosti temeljenim na riječima. Pretpostavka iza ovih mjera je da dokumentna koherentnost može bolje aproksimirati semantičku interpretabilnost tema u slučaju medijskih tema. Predložena metoda za računanje dokumentne koherentnosti sastoji se od tri koraka: (1) koraka odabira uz temu vezanih dokumenata, (2) koraka vektorizacije odabranih dokumenata, i (3) koraka računanja ocjene koherentnosti iz vektora dokumenata pomoću neke od metoda temeljenih na udaljenosti, grafovima ili vjerojatnosnoj gustoći.

U nastavku poglavlja prvo je dan pregled primjena tematskih modela na analizu medijskog teksta te pregled literature o mjerama koherentnosti temeljenim na riječima. Zatim je opisana struktura predloženih mjera dokumentne koherentnosti nakon čega slijede eksperimenti vredno-

vanja tih mjera na dva skupa tema naučenih iz američkih i hrvatskih medijskih tekstova. Potom su provedene kvantitativne i kvalitativne analize odnosa mjera temeljenih na dokumentima i mjera temeljenih na riječima. Poglavlje završava raspravom i prijedlozima za daljnje smjerove istraživanja.

4.2 Tematsko modeliranje vijesti

Mjere dokumentne koherentnosti tema, namijenjene automatskom vrednovanju medijskih tema, mogu se primijeniti, osim na metode analize medijske agende, i na razne druge zadatke tematskog modeliranja medijskog teksta. Ovdje se daje pregled primjena tematskih modela na medijske tekstove, od eksploratorne analize do komercijalnih primjena u sustavima za preporučivanje, sažimanje i dohvat vijesti. Primjene tematskih modela u znanstvenim istraživanjima medijskog teksta, poput postavljanja agende [42] i uokvirivanja [100], opisane su u poglavlju 3. Treba istaknuti da je vrednovanje tematskih modela posebno važno u društvenoznanstvenim primjenama [41], te da mjere predložene u ovom poglavlju predstavljaju korak prema pouzdanim i prihvaćenim metodama automatskog vrednovanja.

Eksplorativna analiza Metode eksploratorne analize koriste tematske modele za izradu statičnih i interaktivnih vizualizacija, kao i sučelja za pregled tekstova koja korisnicima pružaju uvide u velike tekstne zbirke. Jedan na temama temeljeni pristup eksploratornoj analizi vijesti je prikazivanje objekata poput izvora vijesti [54] ili priča [101] kao skupova utežanih tema. Na temelju takvih reprezentacija mogu se izraditi informativni tematski opisi objekata od interesa ili se objekti mogu pretraživati po tematskoj sličnosti. Težine tema za dokumente mogu se iskoristiti za vizualizaciju vremenske zastupljenosti tema. Takve vizualizacije mogu se iskoristiti za analizu zbirke vijesti [102] ili za otkrivanje i vizualizaciju događaja [103]. U slučaju da se koriste vjerojatnosni tematski modeli, iz strukture modela mogu se izračunati vjerojatnosne veze među objektima koji se modeliraju. Primjerice, [102] reprezentira imenovane entitete (osobe, organizacije, itd.) kao riječi te predlaže vizualizacije povezanosti entiteta te entiteta i tema, pri čemu se povezanost računa pomoću uvjetnih vjerojatnosti. Nedavno pregledno istraživanje metoda vizualizacije teksta pokazalo je postojanje velikog interesa za metode temeljene na tematskim modelima [104].

Ostale primjene Tematski modeli primjenjuju se za razne druge zadatke analize medijskog teksta, ili korištenjem težina riječi i dokumenata za teme kako bi se izgradile i obogatile značajke modeliranih objekata, ili korištenjem strukture modela za izvođenje vjerojatnosnih veza između objekata. Primjerice, tematski modeli mogu se koristiti za preporučivanje vijesti na temelju tematskih značajki koje opisuju i interese čitatelja i novinske tekstove [105, 106]. [107]

predlaže metodu sažimanja događaja koja koristi tematski model novinskih članaka i Twitter objava te izgrađuje sažetke rangiranjem objava i rečenica članaka na temelju uvjetnih vjerojatnosti dobivenih iz modela. [108] predlaže sustav za interaktivno otkrivanje priča u vijestima, pri čemu se priče modeliraju kao vremenski uređeni nizovi novinskih članaka koji povezuju početni i završni članak, a novinski članci se modeliraju na temelju značajki dobivenih iz tematskog modela. [109] predlaže poboljšanja metoda dohvata vijesti temeljem upotrebe tematskih modela za obogaćivanje upita i izgradnju jezičnih modela tekstnih dokumenata. Opisana istraživanja pokazuju kako pristupi koji koriste tematske modele daju bolje rezultate od najboljih do tada poznatih metoda ili rezultate usporedive s tim metodama.

4.3 Koherentnost tema temeljena na riječima

Mjere dokumentne koherentnosti predložene u ovom poglavlju nastavljaju se na istraživanja na riječima temeljenih mjera koherentnosti tema (engl. *topic coherence*). Istraživanje tematske koherentnosti motivirano je istraživanjem opisanim u [31], gdje je predloženo mjerenje kvalitete tema modela na temelju njihove interpretabilnosti. [31] su pokazali da modeli koji imaju bolje vrijednosti mjere zbunjenosti (engl. *perplexity*) često imaju manje interpretabilne teme, što ukazuje na to da bi se vrednovanje tema trebalo temeljiti na mjerenju njihove intrinzične semantičke kvalitete i interpretabilnosti.

Iz ove ideje uskoro je izrasla nova familija metoda koje vrednuju semantičku interpretabilnost na način da mjere koherentnost tema [35]. Te metode su temeljene na riječima: nabolje rangirane riječi za temu s jedne predstavljaju ulaz za metode koje računaju koherentnost a s druge strane se koriste za ocjenjivanje tema od strane ljudi kako bi se dobile referentne ocjene koherentnosti. Većina ovih metoda računa koherentnost tema uprosječivanjem semantičke sličnosti najbolje rangiranih riječi za temu (najčešće 5 ili 10 riječi) ili podskupova skupa najboljih riječi [92]. Vrednovanje mjera koherentnosti provodi se pomoću mjera rangiranja ili korelacije na način da se računa slaganje između automatski izračunatih i ljudskih ocjena koherentnosti. Za vrednovanje se tipično koriste AUC mjera (opisana u 4.5.2), te standardni korelacijski koeficijenti (Kendall τ, Spearman ρ i Pearson r). Ljudske ocjene koherentnosti su pri tome ili binarne ocjene (koherentna/nekohoretna tema) ili ordinalne ocjene na Likertovoj skali.

Glavna odluka pri oblikovanju mjera koherentnosti koje uprosječuju sličnosti najbolje rangiranih riječi je izbor mjere sličnosti riječi. Postojeće metode koriste sličnost riječi temeljenu na WordNet ili Wikipedia bazama [35], međusobnoj informaciji po točkama (eng. pointwise mutual information) [32, 35, 110], uvjetnoj vjerojatnosti riječi [111], distribucijskim vektorima (engl. *distributional vectors*) [110], tf-idf težinama [112], te vektorskim reprezentacijama riječi (engl. *word embeddings*) [18, 113]. Kao alternativnu računanju sličnosti parova riječi, [114] su predložili particioniranje skupa najbolje rangiranih riječi u podskupove i usprosječi-

vanje sličnosti parova podskupova. [92] su poopćili dotadašnje pristupe te predložili pristup oblikovanju mjera tematske koherentosti temeljen na agregaciji sličnosti ili parova riječi ili parova podskupova. Pretraživanjem novodefiniranog prostora mjera, [92] su izveli nekoliko novih i kvalitetnih mjera tematske koherenosti. Dvije mjere koherentnosti nisu obuhvaćene prethodno opisanim mjerama: mjera iz [115], temeljena na grupiranju vektorskih reprezentacija najbolje rangiranih riječi i aproksimaciji koherentnosti veličinom najveće grupe, te skup mjera koje računaju koherentnost na temelju rezultata dobivenih pomoću web-tražilice korištenjem upita oblikovanih od najbolje rangiranih riječi teme [35].

Metode predložene u ovom poglavlju spadaju u kategoriju metoda tematske koherentnosti no razlikuju se od svih prethodno opisanih mjera po tome što mjere koherentnost teme koristeći s temom povezane dokumente umjesto s temom povezanih riječi. Drugim riječima, predložene mjere koriste težine dokumenata za temu umjesto težina riječi za temu. Kao što je opisano u uvodu, pretpostavka iza ovog pristupa je da računanje koherentnosti tema na temelju vezanih dokumenata bolje modelira semantičku koherentnost u slučaju medijskih tema često karakteriziranih neinterpretabilnim skupovima riječi. Stoga se, za razliku od prethodnih pristupa tematskoj koherentnosti, predložene mjere temeljene na dokumentima vrednuju na temama označenim ljudskim ocjenama dokumentne koherentosti – ocjenama koherentnosti dobivenim pregledom uz teme vezanih dokumenata umjesto riječi.

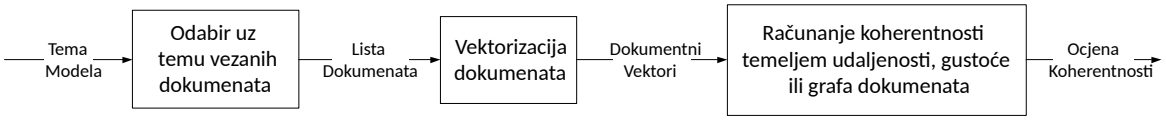
Pristup tehnički najsličniji ovdje predloženom je onaj iz [30], gdje se predlažu mjere kvalitete tema temeljene na vjerojatnostima dokumenata za temu koje se s mjerama temeljenima na riječima kombiniraju u konačnu ocjenu kvalitete tema. Za razliku pristupa iz [30], ovdje je provedeno kvantitativno vrednovanje mjera temeljnih na dokumentima pri čemu se dokumentna mjeru iz [30] koristi kao bazna mjera (engl. *baseline*).

Ovdje korištena metoda vrednovanja ima sličnosti s pristupom opisanim u [28] pošto oba pristupa koriste težine dokumenata za teme. No dok se u [28] vrednovanje provodi na cjelovitim tematskim modelima korištenjem od ljudi označenih dokumenata, ovdje se vrednuju ocjene koherentnosti pojedinih tema na temelju ljudskih oznaka koheretnosti tema.

4.4 Mjere dokumentne koherentnosti tema

U ovom odjeljku opisuju se predložene mjere koherentnosti tema temeljene na dokumentima. Računanje ovih mjera sastoji se od tri koraka: (1) odabira uz temu vezanih dokumenata (2) vektorizacije odabranih dokumenata, i (3) računanja ocjene koherentnosti iz dokumentnih vektora.

Prvi korak kao ulaz prima temu i matricu θ s težinama dokumenata za teme te vraća listu odabranih dokumenata. Odabire se fiksni broj dokumenata s najvećim težinama za temu. Drugi korak vektorizacije dokumenata kao ulaz prima listu tekstnih dokumenata a vraća listu vektora. Vektorizacija tekstova provodi se ili korištenjem standardnih metoda reprezentacije dokume-



Slika 4.1: Tri koraka računanja dokumentne koherentnosti tema.

nata vrećom riječi (engl. *bag-of-words*) i tf-idf težinama (engl. *term frequency–inverse document frequency*) ili agregacijom *vektorskih reprezentacija riječi* (engl. *word embeddings*).[4] U trećem koraku se vektorizirani dokumenti daju na ulaz jednoj od metoda za računanje ocjene koherentnosti koje se mogu podijeliti na metode temeljene na udaljenosti vektora, metode temeljene na vjerojatnosnoj gustoći, i na metode temeljene na grafovima. Metode temeljene da udaljenosti agregiraju udaljenosti parova dokumenata, a vjerojatnosne metode ocjenjuju međusobnu blizinu vektora dokumenata koristeći funkciju vjerojatnosne gustoće. Na grafovima temeljene metode grade graf dokumenata na temelju mjere udaljenosti te računaju koherentnost koristeći jednu od mjera strukturnih svojstava grafa. U nastavku poglavlja opisuju se detalji ova tri koraka računanja mjera koherentnosti.

4.4.1 Odabir uz temu vezanih dokumenata

Svrha ovog koraka je izgradnja liste dokumenata reprezentativnih za temu, odnosno dokumenata povezanih s temom u kontekstu tematskog modela. Ako se odabere prevelik broj dokumenata (u krajnjem slučaju, svi dokumenti) lista dokumenata biti će nekoherentna. Obrnuto, ako se odabere premalo dokumenata (u krajnjem slučaju, jedan dokument s najvećom težinom za temu) lista dokumenata će vrlo vjerojatno uvijek biti visoko koherentna.

Koristi se jednostavna strategija neovisna o tipu tematskog modela: za temu se odabire *BrojDok* dokumenata s najvećim težinama (među svim dokumentima u zbirci) za tu temu. *BrojDok* je parametar koraka odabira dokumenata i u eksperimentima vrednovanja razmatra se nekoliko mogućih vrijednosti. Strategija analogna opisanoj pokazala se dobrom u slučaju mjera koherentnosti temeljenih na riječima, gdje se pokazalo da odabir 10 riječi s najvećim težinama za temu daje dobre rezultate.

Opisana strategija odabira može se formalno opisati pomoću oznaka iz poglavlja 2.1 – težine dokumenata za temu sadržane su u matrici θ, pri čemu je θ_{ij} težina teme j za i-ti dokument (za vjerojatnosne modele θ_{ij} je vjerojatnost pojavljivanja teme j u dokumentu i). Ove težine predstavljaju mjeru povezanosti između tema i dokumenata. Za temu j odabire se prvih *BrojDok* dokumenata iz liste svih dokumenata $D_{i_1}, \ldots, D_{i_N}$ padajuće uređenih prema težini za temu j

[4]Nadalje se za engleski pojam "word embedding" koristi naziv "vektorska reprezentacija riječi". Ovaj pojam označava niskodimenzionalne reprezentacije izgrađene modelima strojnog učenja s ciljem vektorskog modeliranja semantičkih i sintaksnih svojstava riječi.

$(\theta_{i_1 j} \geq \theta_{i_2 j} \geq \ldots \geq \theta_{i_N j})$.

4.4.2 Vektorizacija dokumenata

Svrha koraka vektorizacije je transformacija informacija sadržanih u tekstu dokumenata u vektore koji će biti dani na ulaz metodama za računanje ocjene koherentnosti. Ovi vektori trebaju pružiti metodama računanja koherentnosti dovoljno informacija da ocjene u kojoj mjeri skup dokumenata govori o istoj temi. Očekivano je da će vektorske reprezentacije korištene za grupiranje dokumenata, klasifikaciju dokumenata u tematske kategorije te dohvat dokumenata po tematskim upitima raditi dobro i za računanje tematske koherentnosti skupa dokumenata.

Prva grupa razmatranih metoda vektorizacije sadrži dvije standardne metode često korištene pri klasifikaciji i dohvatu dokumenata. Jedna metoda je prikaz dokumenata vektorom vjerojatnosti riječi u dokumentu, odnosno metoda normalizirane vreće riječi (engl. *bag-of-words*). Druga metoda, vektorizacija tf-idf težinama, koristi podatke o frekvenciji i inverznoj frekvenciji riječi u dokumentima (engl. *term frequency – inverse document frequency*) [116]. Vjerojatnosti riječi i tf-idf vektori računaju se na temelju zbirke novinskih tekstova korištene za izgradnju tematskih modela čije teme se vrednuju. Zbog toga su ove metode domenski specifične odnosno temeljene na informacijama sadržanim u zbirci tekstova određenog tipa.

Druga grupa razmatranih metoda vektorizacije umjesto domenski specifičnih gradi generičke višedomenske vektore dobivene agregacijom, na razini dokumenta, vektorskih reprezentacija (engl. *word embeddings*) riječi u dokumentu. Razmatraju se dvije vrste vektorskih reprezentacija naučenih na temelju velikih zbirki višedomenskih tekstova – CBOW [117] i GloVe [118] reprezentacije. Vektorske reprezentacije riječi CBOW i GloVe su široko korištene u obradi prirodnog jezika, a vektorizacija dokumenata agregacijom vektora riječi pokazala se korisnom za mnoge zadatke, poput grupiranja dokumenata u tematske kategorije [119] te dohvata dokumenata na temelju tematskih upita [120]. Ovi zadaci srodni su s ovdje rješavanim zadatkom procjene tematske koherentnosti skupa dokumenata.

Vektorizacija temeljem frekvencije riječi u dokumentima Ove metode vektorizacije dokumenata temelje se na broju pojavljivanja riječi u dokumentima. Brojanju riječi prethodi pretprocesiranje, koje podrazumijeva barem opojavničenje (engl. *tokenization*), no može uključiti i morfološku normalizaciju riječi, primjerice korjenovanje (engl. *stemming*), te uklanjanje zaustavnih riječi.

Sa N je označen broj dokumenata u zbirci, sa c_{ij} broj pojavljivanja riječi j u i-tom dokumentu, sa d_i veličina i-tog dokumenta, a sa dc_j broj dokumenata u kojima se pojavljuje riječ j. Vektor vjerojatnosti riječi $prob_i$ i-tog dokumenta je vektor empirijskih vjerojatnosti pojavljivanja riječi u dokumentu, i računa se procjenom najveće izglednosti (engl. *maximum likelihood estimation*), $prob_{i,j} = c_{ij}/d_i$. Tf-idf vektorizacija [121] kombinira vjerojatnosti ri-

ječi u dokumentu sa frekvencijama pojavljivanja riječi u drugim dokumentima iz zbirke tekstova. Koristimo tf-idf varijantu kod koje je tf-idf vektor $tfidf_i$ i-tog dokumenta definiran kao $tfidf_{i,j} = tf_{i,j} \times idf_j$, pri čemu je $tf_{i,j} = \log(c_{ij}) + 1$ and $idf_j = \log((N+1)/(dc_j+1)) + 1$. Dodatno, tf-idf vektori dokumenata normaliziraju se na jediničnu euklidsku normu.

Vektorizacija agregacijom vektorskih reprezentacija riječi Ova metoda vektorizacije koristi prethodno izgrađene vektorske reprezentacije riječi [122] – niskodimenzionalne vektore koji modeliraju značenja i sintaksna svojstva riječi a naučeni su na temelju supojavljivanja riječi u velikoj zbirci tekstova. Ovdje se razmatraju dva najčešće korištena tipa vektorskih reprezentacija – CBOW [117] and GloVe [118].

CBOW reprezentacije dobivaju se optimizacijom log-linearnog modela (engl. *log-linear model*) koji predviđa riječ na temelju riječi koje je okružuju. Za eksperimente na engleskom skupu podataka, koriste se 300-dimenzionalne CBOW reprezentacije, naučene na "Google News" zbirci veličine 100 milijardi riječi. Za hrvatski skup podataka, pomoću word2vec alata naučene su 300-dimenzionalne CBOW reprezentacije na hrWaC zbirci hrvatskih web-tekstova [123] veličine 2.8 milijarde riječi. [5]

GloVe reprezentacije dobivaju se učenjem regresijskog modela za aproksimaciju vjerojatnosti supojavljivanja riječi u zbirci tekstova. Na engleskom skupu podataka, koriste se 300-dimenzionalne GloVe reprezentacije naučene na Wikipedia i Gigaword zbirkama tekstova. Na hrvatskom skupu podataka, pomoću dostupnog GloVe alata naučene su 300-dimenzionalne GloVe reprezentacije na hrWaC zbirci.[6]

Vektorska reprezentacija tekstnog dokumenta računa se zbrajanjem vektorskih reprezentacija riječi u dokumentu (ne uključujući zaustavne riječi). Razmatra se i uprosječivanje ovog vektora kako se dobila reprezentacija manje ovisna o duljini dokumenta.

4.4.3 Računanje ocjene koherentnosti

Nakon što su za temu reprezentativni dokumenti odabrani i predstavljeni vektorima, lista vektora dokumenata daje se na ulaz metodi za računanje ocjene koherentnosti. Razmatraju se tri tipa metoda ocjenjivanja koherentnosti: (1) metode temeljene na udaljenostima, koje agregiraju udaljenosti između vektora dokumenata, (2) metode temeljene na vjerojatnosnoj gustoći, koje modeliraju vektore dokumenata pomoću multivarijatne normalne razdiobe, i (3) metode temeljene na grafovima, koje iz vektora dokumenata grade graf i računaju koherentnost koristeći mjere strukturnih svojstava grafa. Ukupno se predlaže devet metoda ocjenjivanja: dvije temeljene na udaljenostima, dvije temeljene na vjerojatnosnoj gustoći, i pet temeljenih na grafovima.

[5] CBOW vektori za engleski i word2vec alat dostupni su na
`https://code.google.com/archive/p/word2vec/`

[6] GloVe vektori za engleski te alat za učenje reprezentacija dostupni su na
`https://nlp.stanford.edu/projects/glove/`

Predložene metode predstavljaju tri različita pristupa koji pretpostavljaju da se koherentnost skupa dokumentnih vektora može dobro opisati pomoću međusobne blizine (metode temeljene na udaljenostima), kompaktnosti (metode temeljene na vjerojatnosnoj gustoći), ili povezanosti (metode temeljene na grafovima).

Koherentnost temeljena na udaljenostima

Metode temeljene na udaljenostima polaze od mjere udaljenosti vektora, odnosno funkcije $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ koja paru vektora pridjeljuje pozitivan realni broj. Mjera udaljenost ne mora biti metrika u matematičkom smislu, već je dovoljno da, poput kosinusne udaljenosti, predstavlja korisnu definiciju udaljenosti.

Razmatraju se dvije jednostavne metode temeljene na udaljenostima: (1) prosječna udaljenost, koja računa prosjek udaljenosti svih parova dokumentnih vektora, i (2) varijanca udaljenosti, koja računa prosječnu udaljenost vektora dokumenata od centralnog (uprosječenog) vektora. U oba slučaja, konačna ocjena koherentnosti računa se negacijom gornjih ocjena kako bi se mjera rasapa pretvorila u mjeru koherentnosti.

Koherentnost temeljena na vjerojatnosnoj gustoći

Metode temeljene na gustoći prvo metodom statističkog zaključivanja procijene parametre multivarijatne normalne razdiobe koja najbolje opisuje skup dokumentnih vektora i zatim računaju koherentnost kao prosjek iz razdiobe dobivenih log-gustoća (engl. *log-density*) vektora. Ova metoda je vođena intuicijom da veća gustoća odgovara boljem grupiranju vektora oko maksimuma funkcije vjerojatnosne gustoće, koje pak odgovara višoj koherentnosti skupa dokumentnih vektora.

Parametri multivarijatne funkcije gustoće vjerojatnosti su vektor srednje vrijednosti $\mu \in \mathbb{R}^n$ i kovarijacijska matrica $\Sigma \in \mathbb{R}^{n \times n}$. Kako bi se smanjio broj parametara i spriječila prenaučenost (engl. *overfitting*), uvodi se pretpostavka da je matrica kovarijacije ili dijagonalna marica ($\Sigma = \text{diag}(\sigma_i^2)$), pri čemu je σ_i^2 varijanca i-te vektorske komponente, ili izotropna matrica ($\Sigma = \sigma^2 \mathbf{I}$).

Za procjenu parametara funkcije gustoće vjerojatnosti koristi se procjenitelj najveće izglednosti (engl. *maximum likelihood*). Prije procjene parametara odnosno učenja modela opcionalno se provodi i smanjenje dimenzionalnosti vektora dokumenata korištenjem metode glavnih komponenti (engl. *principal component analysis*).

Koherentnost temeljena na grafovima

Metode temeljene na grafovima prvo grade graf odabranih uz temu vezanih dokumenata a zatim računaju ocjenu koherentnosti koristeći neku od mjera strukturnih svojstava grafa. Graf je neusmjeren, vrhovi grafa odgovaraju dokumentima a bridovi grafa ovise o udaljenosti vektora

dokumenata. Razmatrane mjere modeliraju ideju kompaktnosti ili povezanosti grafa: centralnost blizine (engl. *closeness centrality*), centralnost podgrafova (engl. *subgraph centrality*), koeficijent grupiranja (engl. *clustering coefficient*), broj povezanih komponenti (engl. *number of connected components*) i veličina minimalnog razapinjućeg stabla (engl. *minimum spanning tree*)[7]

Izgradnja grafa Razmatraju se dvije metode izgradnje bridova grafa koji odgovaraju parovima dokumenata. Prva metoda gradi potpuno povezan težinski graf (engl. *fully connected weighted graph*) u kojemu težine bridova odgovaraju udaljenostima između vektora dokumenata. Druga metoda koristi prag udaljenosti i gradi graf koji sadrži samo bridove između onih parova dokumenata čija međusobna udaljenost je manja od praga. Pri tome se ili mogu zadržati informacije o težini bridova i dobiva se težinski graf, ili se informacije o težini odbacuju i dobiva se neutežani graf.

Centralnost blizine Prva razmatrana graf mjera je centralnost blizine (engl. *closeness centrality*) [125], koja je za vrh v definirana kao inverz prosječne udaljenosti najkraćeg puta od vrha v do svih ostalih vrhova do kojih postoji put iz v:

$$cc(v) = \frac{|C(v)| - 1}{\sum_{w \in C(v)} d(v,w)} \tag{4.1}$$

pri čemu je $C(v)$ skup svih vrhova koji se mogu doseći iz vrha v (skup vrhova u povezanoj komponenti koja sadrži vrh v). Centralnost blizine izoliranog vrha ($C(v) = v$) iznosi 0.

Kako bi se izbjegla visoka vrijednost centralnosti za vrhove fragmentiranog grafa sa mnogo malih povezanih komponenti, centralnost blizine normalizira se relativnom veličinom povezane komponente vrha:

$$cc_{\text{norm}}(v) = \frac{|C(v)| - 1}{N - 1} \frac{|C(v)| - 1}{\sum_{w \in C(v)} d(v,w)} \tag{4.2}$$

pri čemu je N broj svih vrhova u grafu.

Ocjena koherentnosti grafa dokumenata računa se kao normalizirana centralnost blizine uprosječena po svim vrhovima grafa:

$$CC(G) = \frac{1}{N} \sum_{v \in G} cc_{\text{norm}}(v) \tag{4.3}$$

Centralnost podgrafova Centralnost podgrafova (engl. *subgraph centrality*) [126] je mjera centralnosti vrha korelirana s brojem zatvorenih šetnji (ciklusa s mogućim ponavljanjem vr-

[7] Sve razmatrane mjere dostupne su kao dio NetworkX [124] knjižnice dostupne na `http://networkx.readthedocs.io`.

hova) koje započinju u vrhu. Neka je $\mu_k(v)$ broj zatvorenih šetnji duljine k koje započinju u vrhu v. Centralnost podgrafova vrha v definira se kao:

$$sc(v) = \sum_{k=1}^{\infty} \frac{\mu_k(v)}{k!} \tag{4.4}$$

Broj zatvorenih šetnji $\mu_k(v)$ skalira se faktorom $k!$ kako bi se osigurala konvergencija reda. Centralnost podgrafova može se efikasno izračunati pomoću spektralne dekompozicije matrice povezanosti grafa. Težine bridova ne utječu na centralnost podgrafova pošto ona ovisi o broju šetnji a ne o njihovim duljinama. Centralnost podgrafova ne primjenjuje se u slučaju potpunog dokumentnog grafa, pošto svi potpuni grafovi s jednakim brojem vrhova imaju istu ocjenu centralnosti.

Ocjena koherentnosti računa se uprosječivanjem centralnosti podgrafova za sve vrhove grafa:

$$SC(G) = \frac{1}{N} \sum_{v \in G} sc(v) \tag{4.5}$$

Koeficijent grupiranja Trokut koji sadrži vrh v odgovara skupu od tri različita vrha v, u_1, i u_2 takva da postoje bridovi vu_1, u_1u_2, i u_2v. Koeficijent grupiranja (engl. *clustering coefficient*) vrha v definiran je kao udio broja postojećih trokuta koji sadrže v ($T(v)$) u ukupnom broju svih mogućih trokuta koji uključuju v odnosno trokuta koji bi postojali kada bi graf sadražavao sve potrebne bridove među susjedima od v. Koeficijent grupiranja definira se na sljedeći način:

$$cc(v) = \frac{T(v)}{\frac{deg(v)(deg(v)-1)}{2}} = \frac{2T(v)}{deg(v)\big(deg(v)-1\big)} \tag{4.6}$$

Verzija koeficijenta koja uključuje težine bridova i koja se primjenjuje u slučaju težinskog dokumentog grafa, definirana je sa:

$$cc(v) = \frac{1}{deg(v)\big(deg(v)-1\big)} \sum_{u_1,u_2} \big(w'(v,u_1)w'(u_1,u_2)w'(u_2,v)\big)^{1/3} \tag{4.7}$$

Pri tome se zbraja po svim parovima vrhova koji zatvaraju trokut s vrhom v, a $w'(u,v)$ je težina brida između vrhova u i v podijeljena s težinom najtežeg brida u grafu.

Ocjena koherentnosti računa se uprosječivanjem koeficijenata grupiranja svih vrhova u grafu:

$$CC(G) = \frac{1}{N} \sum_{v \in G} cc(v) \tag{4.8}$$

Povezane komponente i razapinjuće stablo Razmatraju se još dvije mjere temeljene na strukturi povezanosti grafa. Prva mjera je inverz broja povezanih komponenti u grafu (engl. *number of connected components*), koja se primjenjuje samo u slučaju nepotpunog dokument-

nog grafa pošto svi potpuni grafovi imaju jednu povezanu komponentu. Druga mjera je negativna težina minimalnog razapinjućeg stabla grafa (engl. *minimum spanning tree*), koja se primjenjuju samo u slučaju potpunog dokumentnog grafa.

4.5 Odabir i vrednovanje mjera dokumentne koherentnosti

U ovoj odjeljku odabiru se i vrednuju prethodno predložene mjere dokumentne koherentnosti tema.[8] Vrednovanje i odabir provode se na temelju dva skupa podataka koji sadrže teme označene ljudskim ocjenama dokumentne koherentnosti. Prvo se definira skup mjera koje odgovaraju razumnim vrijednostima parametara, nakon čega se odabiru mjere s najboljim performansama na razvojnom skupu podataka (engl. *development set*) koje se zatim vrednuju na dva skupa za ispitivanje (engl. *test set*). Za ocjenu kvalitete mjera koristi se površina ispod ROC-krivulje (engl. *area under the ROC curve* – AUC). Na kraju se provodi analiza strukture najboljih mjera. Odjeljak započinje opisom izgradnje skupova podataka.

4.5.1 Skupovi podataka

Uobičajeni način vrednovanja mjera koherentnosti je usporedba izmjerenih koherentnosti tema s ocjenama ljudskih označivača. Pošto se ovdje vrednuju mjere temeljene na dokumentima ljudske ocjene bi također trebale biti temeljene na dokumentima, a ne riječima, vezanima uz temu. U nedostatku skupa podataka za vrednovanje koji odgovara ovom kriteriju, izgrađeni su novi skupovi podataka koji sadrže teme iz zbirki novinskih tekstova koje su od strane ljudi označene s ocjenama koherentnosti temeljenim na dokumentima. Ocjene koherentnosti moguće je dobiti izravnim ocjenjivanjem koherentnosti tema od strane označivača ili neizravno, izvođenjem ocjena koherentnosti tema iz oznaka temeljenih na semantičkoj interpretaciji tema. Ovdje se koristi potonji pristup koji je omogućio da se iskoriste postojeći skupova podataka iz eksperimenata s medijskom agendom koji sadrže pregledane i interpretirane teme modela označene pripadnim semantičkim temama.

Početna točka za izradu skupa podataka je skup podataka iz 3.4, koji sadrži 350 tema 5 modela LDA naučenih na zbirci od 24.000 članaka s američkih web portala. Vrednovanje mjera koherentnosti provedeno je, slično pristupu iz [32], na temama različitih modela kako bi se za vrednovanje koristio raznolikiji skup tema.

Teme modela su pregledane i interpretirane od strane dva označivača koji su proveli postupak otkrivanja tema opisan u 3.3.1. Svaka tema je interpretirana na temelju liste vezanih novinskih članaka i riječi i označena s jednom ili više semantičkih tema – osoba, organizacija, događaja, priča i apstraktnih koncepata. Pri tome je bilo moguće označiti temu modela

[8] Izvorni kod svih mjera i eksperimenata dostupan je na `https://rebrand.ly/doc-coh-code`.

s niti jednom, jednom ili više semantičkih tema, te s oznakom šuma koja označava neinterpretabilne teme sa slučajnim i nepovezanim člancima i riječima. Označivači su interpretirali teme oslanjajući se na uz temu vezane dokumente (novinske članke), pošto su se riječi pokazale nepovezanima, nejasnima ili preopćenitima dok su dobro oblikovani naslovi dokumenata pružali kvalitetne i specifične informacije. Posljedično, odluke o semantičkim temama i njihovoj povezanosti sa temama modela, te odluke o postojanju šuma donesene su na temelju uz temu vezanih dokumenata, dok su riječi u najboljem slučaju služile kao potvrda ovih odluka. Nakon označavanja, 52% tema modela označeno je s jednom semantičkom temom, 15% s jednom semantičkom temom i oznakom šuma, 17% tema označeno je s dvije semantičke teme, 4% s dvije semantičke teme i oznakom šuma, dok je 12% tema označeno kao šum.

Na temelju opisanih oznaka tema, kao *koherentne* su definirane one teme modela označene jednom semantičkom temom (sa ili bez oznake šuma), dok su sve ostale teme definirane kao *nekoherentne*. Odabran je pristup s binarnim ocjenama koherentnosti tema korišten u [112, 113] iz razloga što se takve ocjene mogu dobiti na prirodan način iz opisanih oznaka tema. Drugi pristupi vrednovanju mjera koherentnosti koriste ordinalne ocjene odnosno skalu koherentnosti tema no ne postoji prirodan način mapiranja takvih ocjena na postojeće oznake. Prema opisanom načinu binarnog ocjenjivanja, teme su koherentne ako je označivač prepoznao da tema odgovara jednoj semantičkoj temi odnosno konceptu, što je u skladu s definicijama tematske koherentnosti iz [35] i [111]. Teme su nekoherentne ako odgovaraju bilo šumu bilo mješavini dviju različitih semantičkih tema odnosno koncepata. Opisanim postupkom je 235 tema (67%) ocijenjeno koherentnima dok je 115 tema (33%) ocijenjeno nekoherentnima. Na uzorku od 50 tema označenih od strane oba označivača, označivači su bili suglasni za 88% tema, a koeficijent slaganja označivača kappa [84], koji u obzir uzima i mogućnost slučajnog slaganja, iznosi 0.674.

Opisani skup od 350 tema podijeljen je na dva podskupa: *razvojni skup* sa 120 tema, te *skup za ispitivanje* sa 230 tema. Razvojni skup koristi se za odabir mjera koherentnosti odnosno optimizaciju njihovih parametara a skup za ispitivanje koristi se za vrednovanje odabranih mjera. Kako bi se osigurala reprezentativnost oba skupa tema podjela je izvršena stratificiranim uzorkovanjem (engl. *stratified sampling*). Preciznije, za svaki od skupova očuvan je omjer sljedećih pet originalnih oznaka tema: jedna semantička tema, semantička tema i šum, dvije semantičke teme, dvije semantičke teme i šum, te šum.

Uz opisani skup tema naučenih iz američkih novinskih članaka, za vrednovanje se koristi i dodatni skup za ispitivanje čija svrha je procjena robusnosti rezultata, prije svega s obzirom na jezik teksta. Ovaj skup za ispitivanje sastoji se od tema naučenih iz zbirke tekstova s hrvatskih web portala koja se koristi u istraživanju opisanom u 3.5. Skup se sastoji od ukupno 250 tema dobivenih učenjem četiri modela LDA – tri modela sa 50 tema i jednog modela sa 100 tema.[9]

[9]U originalnom eksperimentu naučeno je ukupno 200 tema, no kako bi veličina skupa test-hr bila usporediva s

Ocjene koherentnosti tema dobivene su provođenjem istog postupka koji se koristio za američke teme – pretakanjem oznaka semantičkih tema i šuma u binarne ocjene koherentnosti. Od ukupno 250 tema, 166 tema (66%) je ocjenjeno koherentnima, dok su 84 teme (34%) ocjenjene nekoherentima.

Opisani skupovi podataka s američkim i hrvatskim novinskim temama slobodno su dostupni.[10] Nadalje se za dva skupa za ispitivanje, američki i hrvatski, koriste oznake test-us i test-hr.

4.5.2 Metoda vrednovanja mjera koherentnosti

Vrednovanje kvalitete mjera koherentnosti provodi se pomoću mjere površine ispod ROC-krivulje (engl. *Area Under the ROC Curve* – AUC) [127]. AUC je mjera kojom se mogu vrednovati metode klasifikacije i rangiranja. Kao mjera rangiranja AUC je primjenjena na vrednovanje mjera koherentnosti temeljenih na riječima [112, 113] pri čemu se uspoređuju binarne oznake tema (koheretna ili nekoherentna tema) i numeričke ocjene koherentosti. AUC mjera koristi se za vrednovanje predloženih mjera dokumentne koherentnosti zbog njene pogodnosti za usporedbu binarnih ocjena koherentnosti tema, dobivenih iz ljudskih oznaka prethodno predloženom metodom, sa numeričkim ocjenama koherentnosti.

Općenito, za model M koji računa ocjene pouzdanosti za elemente $x \in D$ označene binarnim oznakama klasa, AUC mjera jednaka je vjerojatnosti da, za dva elementa x i x' takva da x pripada pozitivnoj a x' negativnoj klasi, pozitivni element dobiva veću ocjene, odnosno da vrijedi $M(x) > M(x')$ [112]. U slučaju mjera koherentnosti i tema označenih oznakom koherentnosti (pozitivna klasa) ili nekoherentnosti (negativna klasa), AUC za mjeru koherentnosti Coh je vjerojatnost da, za koherentnu temu t i nekoherentnu temu t', koherentna tema dobije veću ocjenu koherentnosti, odnosno da vrijedi $Coh(t) > Coh(t')$. AUC mjera poprima vrijednosti iz intervala $[0,1]$, pri čemu je 0 najgora a 1 najbolja ocjena, a 0.5 je očekivana ocjena neinformativne slučajne mjere.

Druga interpretacija AUC mjere temelji se na ideji da je model M koji računa ocjene pouzdanosti moguće pretvoriti u binarni klasifikator korištenjem praga pouzdanosti (ocjena iznad praga odgovara pozitivnoj klasi i obratno). Stopa lažno pozitivnih primjera (engl. *false positive rate* ili *fall-out*) i stopa stvarno pozitivnih primjera (engl. *true positive rate* ili *recall*) ovog klasifikatora za različite vrijednosti praga definiraju krivulju (koordinate točaka na krivulji). Ova krivulja naziva se ROC-krivulja (engl. *receiver operating characteristics curve* ili *ROC curve*). Savršeni klasifikator odgovara točki $(0,1)$ (nema lažno pozitivnih primjera i svi stvarno pozitivni primjeri klasificirani su kao pozitivni). S druge strane, performanse slučajnog klasifikatora za različite vrijednosti praga odgovaraju točkama na dužini koja spaja točke $(0,0)$ i $(1,1)$. AUC

veličinom skupa za ispitavanje, naučen je i označen još jedan dodatni model sa 50 tema.

[10] `https://rebrand.ly/doc-coh-dataset`

mjera definirana je kao površina ispod ROC-krivulje. U slučaju binarnog klasifikatora tema temeljenog na pragu i mjeri koherentnosti, stopa stvarno pozitivnih primjera ili odziv (engl. *recall*) odgovara udjelu koherentnih tema koje je klasifikator prepoznao kao koherentne, dok stopa lažno pozitivnih primjera odgovara udjelu nekoheretnih tema koje je klasifikator prepoznao kao koherentne.

4.5.3 Bazna metoda dokumentne koherentnosti

Kao bazna metoda (engl. *baseline method*) pri vrednovanju predloženih metoda dokumentne koherentnosti koristi se na dokumentima temeljena mjera "tematske značajnosti" (engl. *topic significance*) predložena u [30]. Koliko je autoru rada poznato, ovo je jedina na dokumentima temeljena mjera kvalitete tema. Ova mjera reprezentira temu kao vjerojatnosnu razdiobu nad dokumentima u zbirci koja se dobiva normalizacijom vjerojatnosti tema za dokumente. Zatim se ocjena značajnosti teme računa kao udaljenost između ove razdiobe i neinformativne uniformne razdiobe, pri čemu se kao mjera udaljenosti koristi kosinusna udaljenost ili KL-divergencija. Ovdje se koristi varijanta temeljena na kosinusnoj udaljenosti jer daje bolje performanse na svim skupovima podataka. U [30] ova mjera se ne vrednuje zasebno nego se kombinira, sa slično definiranim mjerama temeljenim na riječima, u konačnu mjeru kvalitete tema koja se zatim kvalitativno vrednuje pregledom tema s visokim i niskim ocjenama.

4.5.4 Metoda odabira mjera

Ovdje se definiraju parametri koji opisuju strukturu predloženih mjera koherentnosti iz odjeljka 4.4. Zatim se definiraju razumne vrijednosti tih parametara koje definiraju skup razumnih varijanti mjera koherentnosti. Na tim skupom mjera provodi se daljni odabir najboljih mjera koje se potom vrednuju na skupovima za ispitivanje. Cilj postupka je pronalazak kvalitetnih na dokumentima temeljenih mjera tematske koherentosti – mjera koje imaju visoku korelaciju (mjerenu AUC mjerom) s ljudskim ocjenama koherentnosti tema.

Kategorizacija mjera koherentnosti Radi preglednije analize, predložene mjere koherentnosti grupirane su u šest strukturnih kategorija prikazanih u tablici 4.2. Svaka kategorija definirana je s dva svojstva mjera: metodom računanja ocjene koherentnosti i metodom vektorizacije dokumenata. Ova dva svojstva najznačajnija su svojstva predloženih mjera koherentnosti i njihova kombinacija u bitnome određuje strukturu i vrstu mjere.

Prvo svojstvo je metoda računanja ocjene koherentnosti, koja određuje način na koji se dokumenti modeliraju (kao točke vektorskog prostora ili vrhovi grafa) i način računanja koherentnosti iz reprezentacija dokumenata (pristupi su opisani u odjeljku 4.4.3). Metode temeljene na udaljenostima (UDALJENOST) agregiraju neku od mjera udaljenosti dokumentnih vektora,

Tablica 4.2: Šest kategorija mjera dokumentne koherentnosti, svaka od kojih odgovara kombinaciji metode računanja koherentnosti i metode vektorizacije dokumenata. Za svaku kategoriju naveden je broj mjera unutar kategorije.

Računanje koherentnosti	Vektorizacija	Broj mjera
UDALJENOST	BROJRIJECI	48
UDALJENOST	VEKTRIJECI	80
GUSTOĆA	BROJRIJECI	96
GUSTOĆA	VEKTRIJECI	128
GRAF	BROJRIJECI	936
GRAF	VEKTRIJECI	1560

dok metode temeljene na vjerojatnosnoj gustoći računaju mjeru rasapa dokumentnih vektora oko središta vjerojatnosne razdiobe. S druge strane, metode temeljene na grafovima (GRAF) izgradnjom grafa dodaju strukturu skupu dokumenata i zatim računaju ocjenu koherentnosti pomoću mjera strukturnih svojstava grafova.

Drugo svojstvo, metoda vektorizacije dokumenata (detalji u odjeljku 4.4.2), definira reprezentacije najbolje rangiranih dokumenata za temu koje se daju na ulaz metodama ocjenjivanja koherentnosti. Metode vektorizacije podjeljene su u dvije klase, BROJRIJECI i VEKTRIJECI. Metode iz prve klase temelje se na frekvenciji riječi odnosno na brojanju pojavljivanja riječi u dokumentima i reprezentiraju dokument vjerojatnostima riječi u dokumentu ili tf-idf težinama. Ove reprezentacije su izvedene iz iste zbirke tekstova koja se koristi za izgradnju tematskih modela čije teme se vrednuju. Pretprocesiranje dokumenata koje prethodi brojanju riječi sastoji se korijenovanja i uklanjanja zaustavnih riječi. S druge strane, VEKTRIJECI metode vektorizacije grade reprezentacije dokumenata agregacijom vektorskih reprezentacija riječi (CBOW i GloVe reprezentacija), koje su pak izgrađene odnosno naučene na temelju velike vanjske zbirke tekstova. Osim razlike u postupku izgradnje vektora dokumenata, važna razlika između ove dvije klase metoda je ta što su BROJRIJECI vektori domenski specifični (u ovom slučaju to su domene američkih i hrvatskih političkih vijesti), dok su VEKTRIJECI vektori generički tj. izgrađeni iz domenski vrlo raznolikih tekstova. Ova razlika može utjecati na računanje ocjene koherentnosti: za razliku od domenski specifičnih vektora, generički vektori će očekivano biti višeznačni i neće nužno odražavati domenski specifična značenja nekih riječi.[11] S druge strane, generički vektori mogu bolje modelirati značenje rijetkih riječi, i mogu biti statistički robustniji

[11] Pošto mnoge višeznačne riječi mogu imati domenski specifična značenja, sužavanje domene teksta korištenog za izgradnju reprezentacija riječi očekivano smanjuje višeznačnost riječi. Ovaj odnos značenja riječi i domene teksta može se iskoristiti za poboljšanje metoda koje rješavaju zadatak razlikovanja značenja riječi [128].

Tablica 4.3: Parametri mjera koherentnosti i vrijednosti parametara, grupirani prema metodi računanja ocjene koherentnosti. Prva tri parametra zajednička su svim metodama.

Računanje koherentnosti	Parametar	Vrijednosti
(sve metode)	*BrojDok*	10, 25, 50, 100
	Vektorizacija	`bow`, `tfidf`, `cbow`, `glove`
	AgregacijaRijeci	`prosjek`, `suma`
UDALJENOST	*Udaljenost*	`l1`, `l2`, `kosinus`
	Agregacija	`prosjek`, `varijanca`
GUSTOĆA	*Matrica*	`izotropna`, `dijagonalna`
	Dimenzionalnost	`bez_smanjenja`, 5, 10, 20, 50, 100
GRAF	*Udaljenost*	`l1`, `l2`, `kosinus`
	GrafMjera	`centralnost-blizine`, `podgraf-centralnost`, `koef-grupiranja`, `broj-komponenti`, `min-stablo`
	Prag	`ne`, 0.02, 0.05, 0.1, 0.25, 0.5, 0.75
	TežinskiGraf	`da`, `ne`

zbog toga što su izgrađeni na temelju mnogo veće zbirke tekstova.

Parametri mjera koherentnosti Radi sistematizacije i odabira na dokumentima temeljenih mjera koherentnosti predloženih u odjeljku 4.4 uvode se parametri tih mjera koji definiraju detalje sva tri koraka računanja mjera: odabira uz temu vezanih dokumenata, vektorizacije dokumenata, i računanja ocjene koherentnosti. Tablica 4.3 prikazuje ove parametre i njihove vrijednosti razmatrane pri odabiru mjera. Parametri su podjeljeni prema metodi računanja ocjene koherentnosti, osim prva tri parametra koji su zajedničkim svim metodama. Zajednički parametri definiraju način konstrukcije skupa dokumentnih vektora koji reprezentiraju temu modela, dok ostali parametri, svojstveni pojedinim metodama ocjenjivanja, definiraju detalje računanja koherentnosti. Opisi parametara mogu se naći u odjeljku 4.4 u kojem se definira struktura mjera.

Treba naglasiti da neke kombinacije vrijednosti parametara nisu smislene. Konkretno, parametar *AgregacijaRijeci* je primjenjiv samo ako je vrijednost od *Vektorizacija* jednaka `cbow` ili `glove`. Za metodu ocjenjivanja GUSTOĆA, ako je *Vektorizacija* jednak `bow` ili `tfidf` (visoko dimenzionalni vektori), parametar *Dimenzionalnost* poprima vrijednosti iz cijelog skupa (5, 10, 20, 50, 100), dok u slučaju da je *Vektorizacija* jednak `cbow` ili `glove` (nisko dimenzionalni vektori), *Dimenzionalnost* poprima vrijednosti 5, 10, i 20. Vrijednost parametra *Dimenzionalnost* jednaka `nemože se kombinirati sa svim metodama vektorizacije, što za metode bow i tfidf` rezultira vektorima čija je dimenzionalnost približno 24.000, a za metode `cbow` i `glove` 300-

dimenzionalnim vektorima (izvorna veličina vektorskih reprezentacija riječi). Neke kombinacije vrijednosti parametara nisu smislene niti za metodu ocjenjivanja GRAF: za potpuno povezani graf dokumenata (*Prag* jednak `ne`) primjenjive su samo mjere `centralnost-blizine`, `koef-grupiranja` i `min-stablo`. Za nepotpune grafove čiji bridovi su filtrirani prema pragu udaljenosti (*Prag* je pozitivan realni broj), `podgraf-centralnost` i `broj-komponenti` se primjenjuju samo u slučaju netežinskog grafa (*TežinskiGraf* jednak `ne`). Tablica 4.2 sadrži broj smislenih kombinacija vrijednosti parametara za svaki od tipova mjera. Sveukupno se razmatra 2.848 različitih mjera koherentnosti.

Vrijednosti parametra *Prag*, koji pripada GRAF metodi ocjenjivanja, definirane su kao udjeli zbog toga što različite metode vektorizacije dokumenata očekivano rezultiraju različitim rasponima udaljenosti vektora. Stoga je za svaku kombinaciju vrijednosti parametara *Udaljenost* i *Vektorizacija* (koja određuje skup međusobnih udaljenosti vektora) procjenjena razdioba udaljenosti vektora na slučajnom uzorku od 100.000 parova dokumentnih vektora dobivenih iz zbirke tekstova. Sama vrijednost parametra *Prag* odgovara percentilu u tako dobivenoj raziobi. Ovako definirani prag udaljenosti može poprimiti vrijednosti 0.02, 0.05, 0.1, 0.25, 0.5, i 0.75 koje odgovaraju drugom, petom, desetom percentilu itd.

4.5.5 Vrednovanje i analiza odabranih mjera

Nakon što su predložene mjere dokumentne koherentnosti opisane skupom parametara, izvršen je odabir mjera vrednovanjem svih smislenih kombinacija vrijednosti ovih parametara. Preciznije, za sve mjere unutar određene kategorije (tablica 4.2) definirane smislenim vrijednostima njihovih parametara (tablica 4.3) izračunat je njihov AUC na razvojnom skupu (odjeljak 4.5.1), nakon čega je odabrana mjera s najvišom vrijednosti AUC mjere. Opisana optimizacija provedena je samo na američkim medijskim tekstovima, a mjere se vrednuju i na američkom (*test-us*) i hrvatskom (*test-hr*) skupu za ispitivanje kako bi se robustnost odabranih mjera ispitala na dva različita skupa podataka. Vrijednosti AUC mjere za svaku od kategorija mjera koherentosti izračunate na *test-us* i *test-hr* skupovima za ispitivanje prikazane su u tablici 4.4. Za svaku od kategorija, tablica sadrži AUC za odabranu mjeru koherentnosti, odnosno mjeru iz te kategorije s najvišom AUC vrijednosti na razvojnom skupu. Sa *doc-dist-cosine* je označena bazna metoda (engl. *baseline*) opisana u odjeljku 4.5.3.

Kao što se može vidjeti iz tablice 4.4, mjera iz kategorije GRAF-BROJRIJECI je najbolja – postiže AUC veći od 0.8 na oba skupa za ispitivanje te nadmašuje ostale mjere za najmanje 0.027 AUC. Za ispitivanje statističke značajnosti razlika između AUC vrijednosti najbolje GRAF-BROJRIJECI mjere i ostalih mjera (uključujući baznu mjeru) koristi se DeLongeov test [129][1]. P-vrijednosti testova prikazane su u tablici 4.4 pored odgovarajućih AUC vrijednosti.

[1]DeLongeov test je osmišljen za usporedbu AUC vrijednosti dviju koreliranih ROC-krivulja (ROC-krivulja dobivenih primjenom različitih mjera na iste podatke). Koristi se implementacija iz pROC R paketa [130].

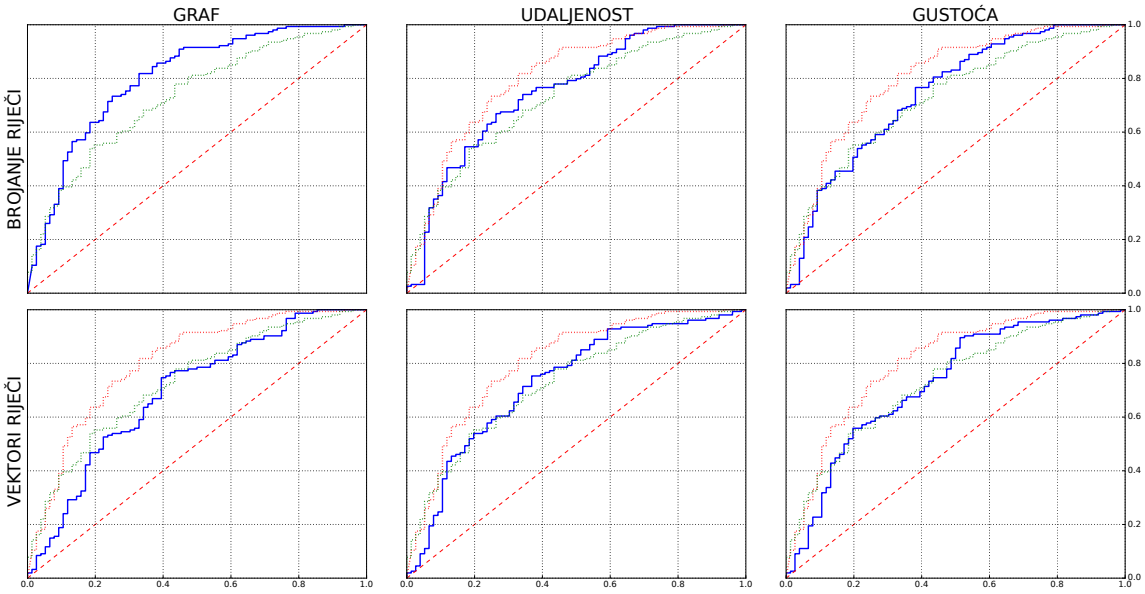
Tablica 4.4: AUC vrijednosti odabranih mjera koherentnosti iz svake od šest kategorija i bazne metode doc-dist-cosine, mjerene na skupovima za ispitivanje. Mjere su uređene prema AUC vrijednostima, a p-vrijednosti su izračunate usporedbom najbolje rangirane mjere i ostalih mjera.

Kategorija mjere		*test-us*		*test-hr*	
Računanje koherentnosti	Vektorizacija	AUC	p-vrijednost	AUC	p-vrijednost
GRAF	BROJRIJECI	**0.804**	–	**0.812**	–
UDALJENOST	BROJRIJECI	0.754	0.001	0.785	0.028
GUSTOĆA	BROJRIJECI	0.745	0.000	0.774	0.009
UDALJENOST	VEKTRIJECI	0.732	0.001	0.746	0.029
doc-dist-cosine	–	0.730	0.006	0.748	0.025
GUSTOĆA	VEKTRIJECI	0.728	0.001	0.725	0.005
GRAF	VEKTRIJECI	0.694	0.003	0.671	0.000

Iz tablice 4.4 se može vidjeti visoka sličnost rangiranja mjera prema AUC vrijednostima na oba skupa podataka, kao i to da na skupu podataka *test-hr* BROJRIJECI mjere i bazna mjera postižu više AUC vrijednosti nego na skupu *test-us*.

Slika 4.2 prikazuje ROC-krivulje najboljih mjera koherentnosti na skupu *test-us*. Kao što je opisano u 4.5.2, ROC-krivulja mjeri klasifikacijske performanse mjera koherentnosti: pojedina točka na krivulji odgovara klasifikatoru temeljenom na mjeri koherentnosti i pragu koherentnosti pomoću kojeg se odlučuje da li je tema koherentna ili nekoherentna. Slika 4.2 uspoređuje najbolje mjere koherentnosti (plava krivulja) s baznom mjerom (zelena krivulja) i s globalno najboljom mjerom GRAF-BROJRIJECI (crvena krivulja). ROC-krivulje pokazuju da sve mjere rade bolje od slučajnog klasifikatora (dijagonalni pravac). Nadalje, ROC-krivulje nadopunjuju podatke u tablici 4.4 prikazom razlika između GRAF-BROJRIJECI mjere i ostalih mjera. Za GRAF-BROJRIJECI mjeru, odziv od 0.8 ili više može se postići uz stopu lažno pozitivnih primjera od najmanje 0.33. Drugim riječima, ako je cilj da GRAF-BROJRIJECI klasifikator otkrije 80% koherentnih tema, potrebno je dopustiti da najmanje 33% nekoherentnih tema bude klasificirano kao koherentno. Ostale mjere mogu postići odziv od 0.8 sa stopom lažno pozitivnih blizu 0.5 (s iznimkom GUSTOĆA-BROJRIJECI mjere sa stopom lažno pozitivnih od 0.43). S druge strane, ako je potrebno postići stopu lažno pozitivnih ispod 0.2, što odgovara pouzdanoj detekciji nekoherentnih tema, GRAF-BROJRIJECI mjera postiže tu stopu uz odziv od 0.64 ili manje, dok druge mjere mogu postići tu stopu uz odziv od 0.56 u najboljem slučaju.

Kao što je vidiljivo iz tablice 4.4 i slike 4.2, bazna mjera opisana u 4.5.3 ima dobre performanse i samo je nešto lošija od ostalih mjera, s iznimkom globalno najbolje GRAF-BROJRIJECI



Slika 4.2: ROC-krivulje najboljih mjera dokumentne koherentnosti iz tablice 4.4. Krivulja bazne mjere prikazana je zelenom bojom. Krivulja najbolje od svih mjera (gore lijevo) prikazana je crvenom bojom uz krivulje ostalih mjera.

mjere koja ima primjetno bolje performanse od bazne mjere.

U prethodnom vrednovanju svaka od kategorija mjera predstavljena je jednom odabranom mjerom – mjerom s najboljim AUC-om na razvojnom skupu. Stoga se postavlja pitanje jesu li odabrane mjere reprezentativne za najbolje mjere iz odgovarajućih kategorija. Kako bi se odgovorilo na to pitanje, za svaku od šest kategorija mjera (tablica 4.2) umjesto samo jedne najbolje mjere odabran je manji broj mjera s najboljim performansama na razvojnom skupu – 5% mjera za velike GRAF kategorije i 10 mjera za ostale kategorije. Te mjere su zatim vrednovane na *test-us* i *test-hr* skupovima za ispitivanje. Analiza tih rezultata pokazuje da najbolje mjere iz tablice 4.4 zaista jesu, uz jednu iznimku, reprezentativne za njihove kategorije odnosno da su njihove performanse usporedive s drugim mjerama unutar iste kategorije koje postižu najbolje performanse na skupovima za ispitivanje *test-us* i *test-hr*. Iznimka je GRAF-VEKTRIJECI kategorija, za koju mjera najbolja na razvojnom skupu ne daje loše rezultate skupovima za ispitivanje, no druge mjere iz te kategorije postižu primjetno bolje rezultate na tim skupovima. Međutim, i najbolje mjere iz te kategorije zaostaju za globalno najboljom GRAF-BROJRIJECI mjerom – na *test-us* skupu usporedive su s odabranim UDALJENOST mjerama, dok su na *test-hr* skupu usporedive s odabranom UDALJENOST-VEKTRIJECI mjerom.

Postavlja se pitanje koje vrijednosti parametara odgovaraju najboljim mjerama, koje bi trebale biti prvi izbor u raznim primjenama. Tablica 4.5 prikazuje vrijednosti parametara dvije najbolje mjera iz svake od kategorija – mjere s najboljim performansama na razvojnom skupu i mjere s najboljim performansama na *test-us* skupu za ispitivanje. Lista svih parametara svake od kategorija mjera nalazi se u tablici 4.3. U nastavku je dan sažetak analize parametara najboljih

Tablica 4.5: Vrijednosti parametara najboljih mjera dokumentne koherentnosti. Za svaku od šest kategorije prikazane su mjera za najboljim rezultatom na razvojnom skupu i mjera s najboljim rezultatom na *test-us* skupu za ispitivanje.

		AUC	
Kategorija	Parametri	Razvojni skup	Skup *test-us*
GRAF-BROJRIJECI	*Vektorizacija*=`tfidf`, *BrojDok*=50, *Udaljenost*=`l2`, *Prag*=0.05, *TežinskiGraf*=`ne`, *GrafMjera*=`podgraf`	0.778	0.812
	Vektorizacija=`tfidf`, *BrojDok*=50, *Udaljenost*=`kosinus`, *Prag*=0.02, *TežinskiGraf*=`ne`, *GrafMjera*=`podgraf`	0.782	0.804
GRAF-VEKTRIJECI	*Vektorizacija*=`cbow`, *BrojDok*=50, *Udaljenost*=`kosinus`, *Prag*=0.25, *TežinskiGraf*=`ne`, *GrafMjera*=`podgraf`	0.730	0.766
	Vektorizacija=`glove`, *BrojDok*=50, *Udaljenost*=`l1`, *AgregacijaRijeci*=`prosjek`, *Prag*=0.25 *TežinskiGraf*=`da`, *GrafMjera*=`koef-grupiranja`	0.792	0.694
UDALJ-BROJRIJECI	*Vektorizacija*=`bow`, *BrojDok*=50, *Udaljenost*=`kosinus`, *Agregacija*=`prosjek`	0.735	0.754
	Vektorizacija=`bow`, *BrojDok*=50, *Udaljenost*=`kosinus`, *Agregacija*=`varijanca`	0.739	0.754
UDALJ-VEKTRIJECI	*Vektorizacija*=`cbow`, *BrojDok*=50, *Udaljenost*=`kosinus`, *Agregacija*=`prosjek`	0.711	0.746
	Vektorizacija=`glove`, *BrojDok*=25, *Udaljenost*=`kosinus`, *Agregacija*=`varijanca`	0.719	0.732
GUSTOĆA-BROJRIJECI	*Vektorizacija*=`tfidf`, *BrojDok*=50, *Dimenzionalnost*=100, *Matrica*=`izotropna`	0.704	0.745
	Vektorizacija=`tfidf`, *BrojDok*=50, *Dimenzionalnost*=`ne`, *Matrica*=`izotropna`	0.704	0.745
GUSTOĆA-VEKTRIJECI	*Vektorizacija*=`cbow`, *BrojDok*=25, *Dimenzionalnost*=5, *AgregacijaRijeci*=`prosjek`, *Matrica*=`izotropna`	0.701	0.734
	Vektorizacija=`cbow`, *BrojDok*=25, *Dimenzionalnost*=10, *AgregacijaRijeci*=`prosjek`, *Matrica*=`izotropna`	0.708	0.728

mjera.

Pokazuje se da sve GRAF-BROJRIJECI mjere s najboljim performansama provode filtriranje bridova grafa na temelju praga težine, a potom odbacuju težine bridova i rade s neutežanim grafom. Od pet graf mjera predloženih za računanje ocjene koherentnosti (odjeljak 4.4.3), samo tri mjere se pojavljuju kao komponente najboljih mjera koherentnosti: centralnost podgrafova, centralnost blizine i koeficijent grupiranja. Sve tri navedene mjere računaju ocjene lokalne povezanosti vrhova grafa i uprosječuju ih kako bi izračunale vrijednost mjere za cijeli graf, za razliku od preostale dvije graf mjere (povezane komponente i min. razapinjuće stablo) koje računaju globalne ocjene povezanosti. Što se tiče praga težine bridova, mjere koje računaju centralnost (centralnost blizine i centralnost podgrafova) koriste manje pragove koji rezultiraju rijetkim grafovima (percentili 0.02, 0.05, i 0.10), dok mjere koje računaju koeficijent grupiranja koriste više pragove (percentili 0.25 i 0.5).

Najbolje GRAF-VEKTRIJECI mjere imaju istu strukturu kao i najbolje GRAF-BROJRIJECI mjere, no ne dostižu njihove performanse. Sve najbolje mjere temeljene na VEKTRIJECI vektorizaciji dokumenata koje koriste `l1` i `l2` mjere udaljenosti provode konstrukciju dokumentnog vektora uprosječivanjem a ne zbrajanjem vektora riječi, što je očekivano pošto su mjere `l1` i `l2`, za razliku od kosinusne udaljenosti, osjetljive na duljinu vektora. Pokazuje se da je kod svih mjera za reprezentaciju dokumenata bolje koristiti BROJRIJECI vektore od VEKTRIJECI vektora, a za graf mjere BROJRIJECI vektori daju globalno najbolje mjere. Ovo pokazuje da je za zadatak aproksimacije dokumentne koherentnosti vektorizacija dokumenata frekvencijom riječi bolja od vektorizacije agregacijom vektorskih reprezentacija riječi.

4.6 Eksperimenti s mjerama koherentnosti riječi

U uvodu ovog poglavlja te u odjeljku 3.6.1 dani su argumenti za tvrdnju da bi u slučaju medijskih tema mjere temeljene na dokumentima trebale biti bolji izbor od mjera temeljenih na riječima. U ovom odjeljku ispituje se odnos predloženih mjera dokumentne koherentnosti i mjera koherentnosti tema temeljenih na riječima.

Prvo se vrednuje aproksimacija ljudskih ocjena dokumentne koherentnosti medijskih tema od strane najboljih mjera koherentnosti temeljenih na riječima. Kako bi se stekli dodatni uvidi u razlike između dvije vrste mjera koherentnosti, provodi se i kvalitativna analiza tema iz četiri različite kategorije definirane vrlo visokim ili vrlo niskim ocjenama mjera koherentnosti temeljenih na riječima i dokumentima.

4.6.1 Odabir mjera koherentnosti temeljenih na riječima

Odabir na riječima temeljenih mjera koherentnosti izvršen je prema istraživanju opisanom u [92], gdje je provedena detaljna analiza i vrednovanje velikog broja mjera na šest skupova podataka. Za daljnje eksperimente odabrano je pet najboljih mjera iz ovog istraživanja. Svaka od tih mjera računa ocjenu koherentnosti teme na temelju najbolje rangiranih riječi za temu: (1) C_{UCI} mjera [35], (2) C_{NPMI} mjera [110], (3) C_A mjera [110], i mjere (4) C_V i (5) C_P, obadvije otkrivene pretraživanjem prostora parametara provedenim u [92].[2]

Pristup iz [92] temelji se na poopćenju strukture mjera koherentnosti riječi. Računanje mjera razloženo je na dva koraka – particioniranje skupa uz temu vezanih riječi na podskupove, te računanje prosječne sličnosti[3] parova podskupova. Ovo poopćenje obuhvaća pristup koji računa prosječnu sličnost parova riječi, koji odgovara slučaju kada su podskupovi definirani kao skupovi koji sadrže pojedine riječi. Sličnost između dva podskupa riječi računa se iz vjerojatnosti supojavljivanja riječi u zbirci tekstova, koja je osim samom zbirkom određena i metodom pretprocesiranja te definicijom jedinice supojavljivanja. Pri tome jedinica supojavljivanja može biti dokument, paragraf ili klizeći prozor (engl. *sliding window*).

Mjere C_{UCI} i C_{NPMI} računaju prosječnu sličnost parova riječi koja se definira ili kao međusobna informacija po točkama (engl. *pointwise mutual information* – PMI) ili kao normalizirana verzija međusobne informacije (engl. *NPMI*). Kod obje mjere supojavljivanje se računa na temelju klizećeg prozora. Mjera C_A također uprosječuje sličnost parova riječi, koja se računa na način da se riječi reprezentiraju vektorima NPMI sličnosti s ostalim najbolje rangiranim riječima teme, nakon čega se sličnost riječi izračuna kao sličnost odgovarajućih vektora. Mjera C_V uprosječuje sličnosti parova koji se sastoje od riječi teme i njoj komplementarnog skupa koji sadrži ostale najbolje rangirane riječi za temu. Slično pristupu u C_A, sličnost riječi i komplementarnog skupa računa se indirektno, temeljem sličnosti vektora. Mjera C_P uprosječuje sličnosti parova koji se sastoje od riječi teme i drugih najbolje rangiranih riječi sa višim rangom, pri čemu se sličnost mjeri pomoću uvjetne vjerojatnosti riječi u odnosu na dani skup riječi [131].

4.6.2 Procjena dokumentne koherentnosti mjerama koherentnosti riječi

Ovdje se ispituje koliko dobro najbolje mjere koherentnosti temeljene na riječima aproksimiraju ljudske ocjene koherentnosti temeljene na dokumentima. Opisane mjere temeljene na riječima namjenjene su predviđanju, na temelju skupa najbolje rangiranih riječi tema, ljudskih ocjena koherentnosti tema temeljenih na promatranju najbolje rangiranih riječi. S druge strane, ocjene koherentnosti tema temeljene na dokumentima pridjeljene su od strane označivača nakon pregleda najbolje rangiranih dokumenata za temu. Mjere temeljene na riječima vrednuju se na

[2]Svih pet mjera dio je softverskog paketa Palmetto, dostupnog na `https://github.com/dice-group/Palmetto`.

[3]U [92] se za mjeru sličnosti dva skupa riječi koristi izraz "mjera potvrde" (engl. *confirmation measure*).

isti način kao i mjere temeljene na dokumentima – na *test-us* i *test-hr* skupovima za ispitivanje (odjeljak 4.5.1), pri čemu se kao mjera kvalitete koristi AUC (odjeljak 4.5.2).

Vrednuju se mjere koherentosti riječi optimirane na zadatku aproksimacije ljudskih ocjena koherentnosti temeljenih na riječima [92]. Sve ove mjere koriste 10 najbolje rangiranih riječi teme i računaju supojavljivanje riječi na temelju Wikipedia zbirke tekstova. Kod primjene mjera na američke novinske teme, supojavljivanje je izvedeno iz engleske Wikipedia zbirke (verzija iz Lipnja 2016.), dok se za hrvatske teme supojavljivanje računa iz hrvatske Wikipedia zbirke (verzija iz Studenog 2017.). U oba slučaja supojavljivanje se računa na temelju iste metode pretprocesiranja koja se koristi za izgradnju tematskih modela.[4] Za svaku od Wikipedia zbirki je prije pretprocesiranja provedeno filtriranje neinformativnih dokumenata odnosno web stranica micanjem stranica za preusmjeravanje, razrješavanje višeznačnosti oznaka (engl. *disambiguation page*), te stranica koje opisuju Wikipedia kategorije i portale.

Tablica 4.6 prikazuje performanse najboljih na riječima temeljenih mjera koherentnosti i performanse bazne mjere *doc-dist-cosine* temeljene na dokumentima, opisane u odjeljku 4.5.3. Navedene p-vrijednosti dobivene su DeLongeovim testom (korištenim i kod vrednovanja u 4.5.5) pri čemu je nulta hipoteza nepostojanje razlike između AUC vrijednosti bazne mjere i mjera temeljenih na riječima. Može se vidjeti da bazna mjera temeljena na dokumentima ima značajno bolje performanse od na riječima temeljenih mjera – najbolja mjera temeljena na riječima ima AUC malo iznad 0.6, dok bazna mjera postiže AUC vrijednosti od najmanje 0.73. Particioniranje skupa najbolje rangiranih riječi teme u parove koji se sastoje od riječi i skupa riječi, korišteno kod mjera C_V i C_P, dovodi do nešto boljih AUC vrijednosti od particioniranja u parove riječi korištenog kod ostalih na riječima temeljenih mjera.

Ovi rezultati su kvantitativna potvrda uočene slabe korelacije između riječi tema i dokumentne koherentosti tema, opisane u odjeljku 3.6.1, koja je motivirala razvoj mjera dokumentne koherentnosti.

4.6.3 Kvalitativna analiza mjera koherentnosti

Prethodno vrednovanje najboljih mjera koherentnosti temeljenih na riječima pokazuje da ove mjere loše aproksimiraju ljudske ocjene dokumentne koherentosti i da su njihove performanse daleko ispod performansi mjera temeljenih na dokumentima. Stoga se postavlja pitanje odnosa između ove dvije vrste mjera. Da li je na dokumentima temeljena koherentnost bolji model za koherentnost tema ili su dokumentna koherentnost i koherentnost riječi dva različita ali korelirana i moguće komplementarna aspekta koherentnosti?

Kako bi se odgovorilo na ovo pitanje, provedena je kvalitativna analiza tema iz skupa američkih medijskih tema. Korišteno je 230 tema iz *test-us* skupa (opisanog u 4.5.1) a analiza je

[4] Za američke teme isprobana su i supojavljivanja iz eksperimenta u [92], dostupna za preuzimanje, no njihova upotreba dovela je do lošijih rezultata za sve mjere.

Tablica 4.6: Performanse mjera koherentnosti riječi na zadatku aproksimacije dokumentne koherentnosti tema, u usporedbi s baznom mjerom temeljenom na dokumentima.

	test-us		*test-hr*	
Mjera	AUC	p-vrijednost	AUC	p-vrijednost
doc-dist-cosine	0.730	–	0.748	–
C_V	**0.607**	0.002	0.508	0.000
C_A	0.579	0.001	0.442	0.000
C_P	0.548	0.000	**0.614**	0.009
C_{NPMI}	0.498	0.000	0.595	0.002
C_{UCI}	0.482	0.000	0.571	0.001

provedena pregledom kategorija tema dobivenih variranjem obje vrste koherentnosti. Preciznije, formirane su četiri kategorije tema od kojih svaka odgovara visokoj ili niskoj dokumentnoj koherentnosti te visokoj ili niskoj koherentnosti riječi.

Kao mjera dokumentne koherentnosti koristi se najbolja mjera iz GRAF-BROJRIJECI kategorije, opisana u odjeljku 4.5.5, dok se kao mjera koherentnosti riječi koristi mjera C_P. Mjera C_P dobro aproksimira ljudske ocjene koherentnosti temeljene na riječima, a loše aproksimira ocjene temeljene na dokumentima (kao što može vidjeti iz tablice 4.6). Za svaku od ove dvije mjere odabrano je 30% tema s najvišim i 30% tema s najnižim ocjenama koherentnosti. Presjekom tih skupova tema dobivene su četiri opisane kategorije tema. Za potrebe ove analize, sve teme koje odgovaraju semantičkim temama kategorizirane su ili kao *konkretne* (teme koje odgovaraju događajima, pričama i entitetima) ili kao *apstraktne* (teme koje odgovaraju raznim vrstama apstraktnih koncepata).

Visoka dokumentna koherentnost, niska koherentnost riječi Od ukupno 23 teme svrstane u ovu kategoriju, većina tema (21) je konkretna. U tablici 4.7 nalazi se pet primjera tema od kojih su sve konkretne osim teme "cijepljenje". Niske ocjene na riječima temeljene koherentnosti su u skladu s obzervacijom da su najbolje rangirane riječi ovih tema semantički nepovezane. Međutim, visoke ocjene dokumentne koherentnosti tema su posljedica toga da su uz teme vezani dokumenti slični novinski članci koji opisuju isti entitet, priču ili događaj. Valja istaknuti da bi ove teme na temelju ocjena najboljih na riječima temeljenih mjera vjerojatno bile odbačene kao nekvalitetne teme, pošto te mjere ne mogu prepoznati njihovu koherentnost. Analogno, ljudski analitičar koji nije dobro upoznat sa zbirkom tekstova vjerojatno bi ocijenio skupove najbolje rangiranih riječi ovih tema nekoherentnima. Stoga su na dokumentima

Tablica 4.7: Teme s visokom dokumentnom koherentnosti i niskom koherentnosti riječi.

Oznaka teme	10 najbolje rangiranih riječi
izbori u Chicagu	gradonačelnik, Chicago, Emanuel, de, Giuliani, Garcia, voli, Blasio, Rudy, drugi_krug
uspostavljanje veza s Kubom	temelj, Kuba, lista, malley, skup, Rubio, kubanski, donacije, Trump, otok
Ted Cruz	Cruz, Ted, sloboda, čaj, zamisliti, evangelički, r-texas, proglasiti, kandidatura, Obamacare
pregovori s Iranom	nuklearni, dogovor, sankcije, iranski, oružje, Kerry, okvir, Tehran, Cotton, Corker
cijepljenje	cjepiva, roditelji, znanost, djeca, izbor, huffpost, Carson, ospice, istraživanje, vjeruje

temeljene mjere koherentnosti koristan alat za detekciju upravo ovakve vrste tema.

Visoka dokumentna koherentnost, visoka koherentost riječi Od ukupno 20 tema u ovoj kategoriji, 14 tema je apstraktno a 6 tema konkretno. Ovaj podatak ukazuje na to da su teme s visokom ocjenom koherentnosti riječi očekivano apstraktne. Najbolje rangirane riječi takvih apstraktnih tema su dobro semantički povezane. U tablici 4.8 nalaze se primjeri pet tema iz ove kategorije. Među tim temama, nalaze se dvije konkretne teme koje odgovaraju entitetima i koje pokazuju da konkretne teme mogu imati visoke ocjene koherentnosti riječi. Ovo se događa u slučaju da riječi teme, koje u kontekstu zbirke tekstova opisuju konkretni pojam, odgovaraju nekom apstraktnom konceptu. S druge strane, činjenica da 14 apstraktnih tema ima visoke ocjene dokumentne koherentnosti pokazuje da na dokumentima temeljene mjere mogu prepoznati ne samo konkretne već i apstraktne teme.

Niska dokumentna koherentnost, visoka koherentost riječi Od 12 tema u ovoj kategoriji tri su nekoherentne dok su sve ostale teme apstraktne. To ide u prilog tvrdnji o korelaciji između apstraktnosti tema i visoke koherentnosti riječi. Ove apstraktne teme mogu se svrstati u dvije skupine. Prva skupina sadrži četiri teme (sudske tužbe, novinarstvo, društvene mreže, radio i televizija) čija je niska dokumentna koherentnost posljedica nezastupljenosti tih tema u dokumentima. Preciznije, ove teme se spominju u relativno malom udjelu teksta u dokumentima kojima dominiraju druge teme. Zbog toga su uz teme vezani dokumenti heterogeni i nekoherentni, što negativno utječe na dokumentnu koherentnost tema. Druga skupina sastoji se od pet tema koje su koherentne ali čije su ocjene dokumentne koherentnosti ili pogrešno izračunate

Tablica 4.8: Teme s visokom dokumentnom koherentnosti i visokom koherentosti riječi.

Oznaka teme	10 najbolje rangiranih riječi
okoliš	klima, energija, globalno, znanost, okolišni, zagrijavanje, gorivo, znanstvenici, emisije, tvornice
proračun	milijarda, domaći, fiskalni, ravnoteža, deficit, Medicare, opoziv, prioriteti, Ryan, trilijun
dužnička kriza	dug, kredit, dolari, ugovor, naknada, otplata, porezni_obveznici, zajmoprimci, riznica, potrošači
Robert Menendez	tužilac, Menendez, odvjetnik, kriminalac, optužbe, podnijeti, kazna, tužitelji, osuđen, zatvor
Yemen	saudijski, napad, meta, al, Yemen, obavještajni, Arabija, Hutiji, Pakistan, Qaeda

ili niske zbog visoke apstraktnosti tema i posljedične povezanosti tema s skupom međusobno semantički manje srodnih dokumenata. Tablica 4.9 sadrži sve teme iz prve skupine i primjer jedne teme iz druge skupine. Teme iz ove kategorije pokazuju da bi kombiniranje na dokumentima i na riječima temeljenih mjera koherentnosti moglo biti korisno. Koherentnost riječi mogla bi se iskoristiti kao pomoćna mjera za slučaj koherentnih tema čija koherentnost nije detektirana ili ne može biti detektirana na temelju mjera dokumentne koherentosti, primjerice opisanih tema koje se javljaju u malom udjelu teksta u dokumentu.

Niska dokumentna koherentnost, niska koherentnost riječi Od 26 tema u ovoj kategoriji 18 tema nije koherentno (odgovara šumu ili mješavini tema), dok je preostalih 8 tema koherentno i odgovara jednoj semantičkoj temi no sadrži stanovitu količinu šuma. Većina koherentnih tema, njih 7 od 8, su konkretne teme, što je svojstvo koje korelira s niskom koherentnosti riječi. Ove teme su specifične pošto ujedno i sadrže šum koji snižava dokumentnu koherentnost i odgovaraju konkretnim konceptima što je svojstvo koje korelira s niskom koherentnosti riječi. Stoga je koherentnost tih 8 tema teško detektirati čak i kombiniranjem mjera koherentnosti dokumenata i mjera koherentnosti riječi.

Teme iz ove kategorije zajedno s temama s niskom koherentnosti dokumenata i visokom koherentnosti riječi ukazaju na koristi od kombiniranja dvije vrste mjera koherentnosti. Naime, među temama s niskom dokumentnom koherentnosti, koherentnost temeljena na riječima korelira sa semantičkom koherentnosti: većina (18 od 26) tema s niskom koherentnosti riječi je inkoherentna, dok je većina (10 od 12) tema s visokom koherentnosti riječi semantički koherentna.

Tablica 4.9: Teme s niskom dokumentnom koherentnosti i visokom koherentosti riječi.

Oznaka teme	10 najbolje rangiranih riječi
sudske tužbe	podnijeti, odbor, tužba, žalba, parnica, prekršiti, Georgia, šteta, nagoditi, optužen
radio i televizija	jutro, domaćin, gledati, noć, radio, televizija, tv, mreža, ažurirati, stanica
novinarstvo	pisanje, objavljen, članak, priča, novinari, tisak, magazin, novine, citiran, urednik
društvene mreže	Fox, Twitter, domaćin, noć, tweet, jutarnji, Facebook, com, gledati, opaska
zločin	zatvor, kriminalac, zločin, kazna, osuđen, tužitelj, tužioci, suđenje, odvjetnik, sudac
šum	video, netko, mislio, vjerojatno, možda, inače, čovjek, bilo_što, svatko, da

4.7 Rasprava

Tematski modeli su popularan alat strojnog učenja za otkrivanje tema u tekstnim zbirkama. Međutim, kvaliteta automatski naučenih tema varira i ta pojava potaknula je razvoj niza metoda za automatsko vrednovanje tema, poput mjera semantičke koherentnosti koje računaju koherentnost tema na temelju vezanih riječi. Analiza tema naučenih u postupku otkrivanja tema na medijskoj agendi, opisanom u poglavlju 3, potvrdila je varijaciju u kvaliteti tema i potrebu za mjerama kvalitete tema. Međutim, ova analiza također je ukazala na neadekvatnost postojećih na riječima temeljenih mjera koherentnosti za procjenu kvalitete medijskih tema koje su često karakterizirane semantički nepovezanim skupovima riječi. Na temelju opažanja da se medijske teme mogu dobro interpretirati na temelju dokumenata predložena je nova klasa mjera koherentnosti koje računaju koherentnost tema koristeći uz temu vezane dokumente. Predložene mjere računaju koherentnost tema u tri koraka: korak odabira uz temu vezanih dokumenata, korak vektorizacije dokumenata, te korak računanja ocjene koherentnosti iz vektora dokumenata. Računanje ocjene koherentnosti provodi se jednom od tri metode: agregacijom udaljenosti dokumentnih vektora, računanjem vjerojatnosne gustoće vektora, te modeliranjem dokumenata grafom. Vrednovanje predloženih mjera koherentnosti provedeno je na dva jezično različita skupa podataka koji sadrže medijske teme označene ljudskim ocjenama koherentnosti. Iako su mjere dokumentne koherentnosti tema motivirane razvojem alata za analizu medijske agende, mogu se primijeniti za vrednovanje i odabir tematskih modela u bilo kojoj od brojnih primjena

ovih modela na medijski tekst.

Vrednovanje na dokumentima temeljenih mjera koherentnosti, provedeno u odjeljku 4.5.5, pokazalo je da su najbolje mjere koje provode vektorizaciju dokumenata na temelju frekvencije riječi u dokumentima, grade graf dokumenata i zatim računaju ocjenu koherentnosti pomoću mjera strukturnih svojstava grafa. Ove mjere grade graf dokumenata uklanjanjem svih parova dokumenata čija međusobna udaljenost prelazi niski prag udaljenosti i računaju koherentnost uprosječivanjem, po svim vrhovima grafa, neke od mjera lokalne povezanosti. Visoke performanse ovih mjera potvrđene su i na američkim i na hrvatskim medijskim temama. Struktura tih mjera pokazuje da se dokumentna koherentnost može dobro aproksimirati na temelju informacija o lokalnoj povezanosti bliskih dokumenata. Vrednovanje mjera dokumentne koherentnosti očekivano pokazuje i jak utjecaj metoda vektorizacije dokumenata na kvalitetu mjera. Računanje koherentnosti na temelju 50 najbolje rangiranih dokumenata za temu pokazalo se najboljim izborom – ova vrijednost rezultira najboljim ili gotovo najboljim rezultatima za sve kategorije mjera.

Računanje koherentnosti tema na temelju dokumenata motivirano je opažanjem da koherentnost temeljena na riječima nije uvijek dobar model koherentnosti, primjerice u slučaju prolaznih tema ovisnih o događajima kakve se često pojavljuju u medijima. Eksperimenti opisani u odjeljku 4.6.2, potvrđuju ovu tvrdnju – nabolje na riječima temeljene mjere, optimirane za aproksimaciju koherentnosti skupa riječi, daju slabe rezultate aproksimacije ljudskih ocjena koherentnosti temeljenih na dokumentima. Pri tome neke na riječima temeljene mjere postižu stanoviti stupanj korelacije s dokumentnom koherentnosti.

Provedena kvalitativna analiza tema ukazuje na to da su na riječima i na dokumentima temeljene mjere koherentnosti dvije komplementarne klase mjera koje bi se mogle kombinirati kako bi se preciznije odredila koherentnost tema. Naime, postoje koherentne teme (teme modela koje odgovaraju semantičkim temama) koje bi bile označene kao nekoheretne samo na temelju jedne klase mjera. Kvalitativna analiza također ukazuje na to da visoka koherentnost riječi korelira s apstraktnim temama dok niska koherentnost riječi korelira s konkretnim temama.

Postoji mnogo izglednih smjerova za buduća istraživanja dokumentne koherentnosti. Jedan smjer je poboljšanje predloženih mjera, što je vjerojatno moguće postići eksperimentiranjem s raznim metodama vektorizacije dokumenata, primjerice neuralnim vektorskim reprezentacijama dokumenata [132] ili sofisticiranijim metodama agregacije reprezentacija riječi [119]. Drugi smjer, motiviran eksperimentima u odjeljku 4.6, je kombiniranje mjera temeljenih na riječima s mjerama temeljenim na dokumentima. Ovaj smjer mogao bi dovesti do potpunijeg modela koherentnosti tema i bolje detekcije kvalitetnih tema. Dokumentne mjere mogu se primijeniti za komparativnu analizu različitih klasa tematskih modela, slično analizama provedenim na temelju mjera koherentnosti riječi [18, 22]. Trebalo bi istražiti primjenu mjera dokumentne koherentnosti na druge tekstne domene. Ove mjere bi očekivano radile dobro u

slučaju kratkih i tematski fokusiranih tekstova poput tekstova na društvenim mrežama, a trebalo bi istražiti njihove performanse i moguće prilagodbe na tematski heterogenijim tekstovima poput znanstvenih članaka.

Poglavlje 5

Pokrivenost tema

Pitanja vezana uz *pokrivenost tema* (engl. *topic coverage*) prirodno se javljaju kod primjena i analiza tematskih modela. Tematski modeli na temelju zbirke tekstova uče teme modela, odnosno konstrukte koji odgovaraju utežanim listama riječi i dokumenata, od kojih se očekuje da odgovaraju semantičkim temama odnosno konceptima koji se javljaju u tekstovima. Problematika pokrivenosti tema razmatra pitanja vezana uz pokrivenost semantičkih tema (koncepata) od strane tema modela, primjerice pitanja koliko semantičkih tema očekivano pokriva jedan model, kakve tipove semantičkih tema može pokriti određena klasa modela, te kakva je priroda poklapanja semantičkih tema i tema modela – precizno poklapanje ili aproksimativna semantička sličnost.

Nekoliko postojećoh istraživanja tematskih modela bavi se ili ukazuje na problematiku pokrivenosti. Pojam pokrivenosti tema uveden je u [29], gdje se provodi prva analiza pokrivenosti skupa "referentnih koncepata" od strane tema modela. [29] predlažu metodu mjerenja poklapanja tema modela i koncepata te nekoliko alata za vizualnu analizu tematskih modela temeljenih na ovoj metodi mjerenja. Analize modela ukazuju da pojedini modeli pokrivaju samo dio koncepata. Metoda vizualne analize i validacije tematskih modela predložena u [64] temelji se na grupiranju bliskih tema većeg broja modela i promatranju kako teme pojedinih modela korespondiraju s ovim grupama. Grupe se mogu promatrati kao koncepti a provedeni eksperimenti pokazuju da pojedine grupe, osobito one s manje zastupljenim temama, bivaju pokrivene samo od djela modela. U analizi primjena tematskih modela u društvenoznanstvenim istraživanjima [40] autori ukazuju na problem pokrivenosti "sheme kodiranja" – skupa koncepata na kojima se temelji analiza podataka.

Istraživanja primjene tematskih modela na analizu medijske agende provedena u poglavlju 3 također ukazuju na prisutnost i važnost problematike pokrivenosti tema. U odjeljku 3.6.2 provodi se analiza skupa semantičkih tema i vezanih tema modela – podataka dobivenih interpretacijom tema modela u postupku eksploratorne analize medijskog teksta. Ova analiza pokazuje da interpretacija novih tematskih modela dovodi do prethodno neotkrivenih i poten-

cijalno zanimljivih semantičkih tema, te da pojedini tematski modeli pokrivaju samo dio svih semantičkih tema.

Unatoč važnosti i prirodnosti problematike pokrivenosti u kontekstu tematskog modeliranja, izostaju sistematični pristupi koji provode kvantitativnu analizu pokrivenosti i nedostaju pouzdane i lako primjenjive metode za mjerenje pokrivenosti. Metode mjerenja pokrivenosti i na njima temeljene analize modela mogle bi omogućiti odabir kvalitetnijih modela koji bi omogućili efikasniju eksplorativnu analizu tekstnih zbirki. U kontekstu primjena tematskih modela na analizu medijske agende, modeli koji daju veću pokrivenost mogli bi omogućiti otkrivanje većeg broja semantičkih tema uz manji broj pregledanih tema modela, što bi dovelo do vremenskih ušteda i otkrivanja većeg broja relevantnih koncepata. Kao što je opisano u odjeljku 3.2, u društvenoznanstvenim analizama postavlja se pitanje relevatnosti tema – pitanje da li teme modela odgovaraju konceptima koji su relevantni za analizu koja se provodi. Ovom važnom problemu moglo bi se pristupiti metodama za mjerenje pokrivenosti na način da se mjeri pokrivenost poznatih skupova koncepata relevantnih za pojedine analize. Ovakav pristup bio bi prvi kvantitativni pristup mjerenju relevantnosti tema.

U ovom poglavlju predlaže se niz metoda koje adresiraju različite aspekte problema pokrivenosti. Ovaj smjer istraživanja motiviran je važnošću problematike pokrivenosti u kontekstu validacije i primjene tematskih modela, te nedostatkom istraživanja o toj problematici. Predložene metode predstavljaju početne korake prema boljem razumijevanju ove zapostavljene problematike. Iako je istraživanje problema pokrivenosti tema motivirano eksperimentima analize medijske agende, problem pokrivenosti je općenitiji problem validacije tematskih modela neovisan o tekstnoj domeni. Iz tog razloga vrednovanje predloženih metoda mjerenja i sama analiza pokrivenosti provode se na skupovima podataka iz dvije različite tekstne domene – domene medijskog i biološkog teksta. Svaki od skupova podataka sadrži zbirku tekstova, referentne teme, naučene tematske modele i parove tema označene ljudskim ocjenama poklapanja.

U ostatku poglavlja razmatraju se dvije metode izrade referentnih tema temeljenih na naučenim temama modela i predlažu se dvije metode mjerenja pokrivenosti temeljene na ocjenjivanju semantičkog poklapanja referentnih tema i tema modela – nadzirana metoda koja aproksimira ljudsko ocjenjivanje pokrivenosti te automatska nenadzirana metoda temeljena na mjeri udaljenosti tema. U eksperimentima se koristi veliki skup tematskih modela iz različitih klasa dobiven variranjem parametara izgradnje. Temeljem razvijenih metoda i skupova podataka iz dvije različite tekstne domene, medijske i biološke, provodi se analiza pokrivenosti referentnih tema od strane tematskih modela.

Osim prijedloga i primjene konkretnih metoda analize pokrivenosti, problemu pokrivenosti pristupa se općenito – daje se općenita definicija problema i razmatraju se alternativni pristupi u kontekstu te definicije. U nastavku poglavlja najprije se razmatra problem pokrivenosti tema a zatim se daje pregled vezanih istraživanja. Poglavlje se zaključuje raspravom o rezultatima i

smjernicama za daljnja istraživanja.

5.1 Problem pokrivenosti tema

Ovdje se definira problem pokrivenosti tema, razmatraju se njegovi aspekti i na kraju se daje pregled metoda predloženih u ostatku poglavlja. *Problem pokrivenosti* sastoji se od mjerenja kako teme tematskog model pokrivaju teme iz skupa referentnih tema. Tri su aspekta ovog problema: *referentne teme*, *mjere pokrivenosti* i *tematski modeli*. Referentne teme su teme čija pokrivenost se mjeri i određuju problem pokrivenosti definiranjem vrste tema koje se nastoji pokriti. Tematski modeli su alati za učenje tematskih koncepata iz tekstova i svrha eksperimenata pokrivenosti je ispitati kako razni tipovi i konfiguracije tematskih modela mogu pokriti određene referentne teme. Mjere pokrivenosti računaju ocjenu koja opisuje kako se teme tematskih modela poklapaju s referetnim temama.

Prva i jedina analiza tematskih modela s aspekta pokrivenosti provedena je u [29]. Ovo istraživanje je važno zbog uvođenja pojma pokrivenosti tema i provođenja eksperimenta čija struktura odgovara prethodnoj definiciji problema pokrivenosti. Međutim, u istraživanju nije provedeno vrednovanje predložene mjere pokrivenosti niti kvantitativna analiza pokrivenosti tema a sam problem pokrivenosti nije izdvojen i istaknut kao zaseban važan problem. Istraživanje opisano u ovom poglavlju nastavlja se na [29], predlaganjem i vrednovanjem novih metoda mjerenja, provođenjem kvantitativnih analiza pokrivenosti, te definiranjem i razmatranjem problema pokrivenosti u općenitosti.

Skup referentih tema Skup referentnih tema je ključan pošto kvaliteta tematskih modela iz perspektive problema pokrivenosti u potpunosti ovisi o mogućnosti modela da otkriju referentne teme. Primjerice, istraživanja koja primjenjuju tematske modele za analizu medijske agende, opisana u odjeljku 3.1.3, pokazuju da modeli s manjim brojem tema uče apstraktnije teme dok modeli s većim brojem tema uče specifičnije teme. Prema tim opažanjima, skup apstraktnih referentnih tema trebao bi biti bolje pokriven od modela s manje tema, i obrnuto. Neovisno o primjeni, referentni skup bi trebao sadržavati teme koje se javljaju u tekstovima koji se koriste za učenje modela, odnosno teme koje je moguće detektirati na temelju informacija u zbirci tekstova.

Moguće je zamisliti mnoge definicije referentnog skupa tema koje odgovaraju različitim scenarijima vrednovanja tematskih modela – više ili manje apstraktne teme, teme reprezentativne za usku klasu koncepata ili teme koje pokrivaju širi semantički prostor. Problemu *relevantnosti tema* tematskih modela u društvenoznanstvenim primjenama opisanom u odjeljku 3.2 moglo bi se pristupiti kao problemu pokrivenosti s referentnim temama koje predstavljaju teme korisne za provođenje određene analize.

Moguće je zamisliti razne načine izrade skupa referentih tema, primjerice definicijom tema od strane stručnjaka na temelju pregleda tekstova [64], definicijom tema na temelju koncepata iz neke baze znanja poput ontologije ili enciklopedije, ili ljudskim odabirom i doradom interpretabilnih tema naučenih modela.

Mjere pokrivenosti Mjera pokrivenosti na temelju skupa referentnih tema i skupa tema modela računa ocjenu koja opisuje u kojoj mjeri i na koji način teme modela pokrivaju referentne teme. Dok skup referentih tema opisuje tip tema koje se želi pokriti temama modela, mjera pokrivenosti definira samu prirodu pokrivanja tema.

Dva su aspekta mjerenja pokrivenosti, mjerenje poklapanja dvije teme – referentne teme i teme modela, te mjerenje same pokrivenosti cijelog skupa referentnih tema od strane svih tema modela. Poklapanje parova tema razumno je definirati kao neku vrstu semantičkog poklapanja tema, odnosno koncepata dobivenih interpretacijom tema. Takva definicija može varirati u rasponu od strogog poklapanja koncepata do neke vrste aproksimativne semantičke sličnosti.

Računanje pokrivenosti na razini skupova tema može se izvesti jednostavnom agregacijom ocjena poklapanja, primjerice računanjem broja ili postotka referentnih tema za koje postoje teme modela koje ih dobro opisuju. S druge strane, pokrivenost skupova mogla bi se definirati na temelju strukture povezanosti referentnih tema s temama modela. Takva mjera, za razliku od jednostavne agregacije, mogla bi formirati ocjenu pokrivenosti na temelju modeliranja skupa tema kao cjeline.

Konačno, mjera pokrivenosti može biti interpretabilna odnosno davati korisnu informaciju o modelu sama po sebi ili može biti neinterpretabilna i služiti samo za rangiranje modela prema pokrivenosti.

Tematski modeli i njihova analiza Tematski modeli su alati za učenje tematskih koncepata i svrha eksperimenata pokrivenosti je ispitati kako razni tipovi i konfiguracije tematskih modela mogu pokriti određene referentne teme. U kontekstu problema pokrivenosti tematski modeli se promatraju općenito, kao modeli koji iz tekstova uče skup tema – utežanih listi riječi i dokumenata. Jednom kada su definirani skup referentnih tema i mjere pokrivenosti, moguće je pristupiti analizi tematskih modela. Takva analiza mogla bi ponuditi odgovore na pitanja kakvu pokrivenost daje određeni tip modela, koliko varira pokrivenost modela istog tipa, kako parametri pretprocesiranja i metoda učenja modela utječu na pokrivenost, te kako postići optimalnu pokrivenost odabirom tipa i konfiguracije modela.

U primjenama tematskih modela mjerenje pokrivenosti moglo bi poslužiti za izbor najboljeg skupa postavki učenja modela ili za izbor najboljeg modela među stohastičkim varijacijama naučenim s istim postavkama. Razmatranjem skupova referentnih tema repreznentativnih za određene kategorije koncepata mogli bi se steći novi uvidi o tipovima tema koje pojedine

klase tematskih modela mogu naučiti. U raznim primjenama gdje je od interesa detekcija određene vrste tema to bi moglo dovesti do izbora najboljeg tipa modela. Primjerice kod analize uokvirivanja (engl. *framing*) medijskog teksta bilo bi korisno odrediti modele čije teme dobro odgovaraju okvirima.

Predložene metode i analiza pokrivenosti U ostatku poglavlja se, nakon pregleda postojećih istraživanja povezanih s problemom pokrivenosti, predlažu metoda izrade skupa referentnih tema i dva pristupa mjerenju pokrivenosti, nakon čega se provodi analiza pokrivenosti četiri različite klase modela. Svi eksperimenti provode se na dva domenski različita skupa podataka – medijskom i biološkom.

Skupovi referentnih tema dobiveni su pregledom i doradom tema modela dobivenih postupcima eksplorativne analize, s idejom provođenja eksperimenata pokrivenosti na temama koje ujedno informativne i korisne i moguće ih je otkriti tematskim modelima. Prvi skup referentnih tema dobiven je temeljem medijskih tekstova i modela korištenih za otkrivanje medijskih tema u sklopu eksperimenta iz odjeljka 3.4, Drugi skup referentnih tema dobiven je temeljem modela korištenih za eksplorativnu analizu bioloških tekstova iz [17]. Dok skup medijskih tema sadrži široki raspon tema koje nisu filtrirane prema tipu i koje se razlikuju po razini apstrakcije, skup bioloških tema sadrži apstraktne teme koje odgovaraju biološkom konceptu fenotipa.

Predlažu se dvije metode mjerenja pokrivenosti. Prva metoda temelji se na računanju poklapanja parova tema nadziranim modelom koji aproksimira ljudske ocjene konceptualne jednakosti tema. Ova metoda računa pokrivenost cijelog skupa referentnih tema od strane tema modela kao udio referentnih tema za koje postoje odgovarajuće teme modela. Druga, nenadzirana metoda temelji se na aproksimaciji semantičkog poklapanja tema mjerom udaljenosti tema i pragom udaljenosti koji definira poklapanje. Ova metoda agregira pokrivenosti dobivene variranjem praga i omogućava vizualni prikaz pokrivenosti referentnih tema koji daje informacije o tome kako tematski model funkcionira kroz spektar definicija pokrivenosti. Ove mjere koriste se za analizu većeg broja tematskih modela koji predstavljaju četiri različite klase modela i različite parametre izgradnje.

5.2 Pregled literature

Ovdje se daje pregled istraživanja iz područja tematskih modela i dubinske analize teksta koja su povezana s problematikom pokrivenosti tema na način da ili izravno adresiraju sam problem ili njegove aspekte, ili razmatraju ideju pokrivenosti u primjeni različitoj od tematskog modeliranja.

Metode vrednovanja tematskih modela

U [29] je opisano prvo i jedino istraživanje koje razmatra problem pokrivenosti tema. [29] predlažu metodu vrednovanja tematskih modela temeljenu na vizualnom prikazu odnosa između tema modela i skupa referentnih koncepata. Referentni koncepti izgrađeni su od strane domenskih stručnjaka na temelju pretraživanja zbirke članaka o vizualizaciji informacija i reprezentirani kao liste riječi i dokumenata. Poklapanje tema i koncepata mjere se temeljem vjerojatnosnog modela koji aproksimira vjerojatnost da čovjek ocijeni temu i koncept jednakima. Ovaj model uči se na temelju ljudskih ocjena poklapanja parova tema i iz modela se izvodi više vrsta odnosa tema i koncepata: puno poklapanje, poklapanje koncepta s dijelom teme koja spaja više koncepata, poklapanje koncepta s više tema, i izostanak poklapanja. U eksperimentima vrednovanja modela promatra se kako odnosi tema i koncepata variraju u ovisnosti o hiperparametrima izgradnje i tipovima modela. Pri tome se pokrivenost jednog koncepta definira kao poklapanje s nekom od tema na bilo koji od opisanih načina. Eksperimenti ukazuju na to da se neki koncepti mogu pokriti samo za određene vrijednosti hiperparametara i tipove modela, te na veliku varijaciju vrste poklapanja u ovisnosti o hiperparametrima.

Ovaj eksperiment dotiče sve aspekte općenitog problema pokrivenosti te uvodi sam pojam pokrivenosti (engl. *coverage*) skupa referentnih koncepata od strane tema modela. Međutim, mjerenje pokrivenosti koristi se kao dio alata za vizualno vrednovanje modela i izostaju kvantitativni eksperimenti pokrivenosti. Glavni problem pristupa predloženog u [29] je metoda mjerenja poklapanja tema i koncepata na kojoj se temelji mjerenje pokrivenosti referentnih koncepata. Naime, ljudsko ocjenjivanje poklapanja parova tema provedeno je od strane ocjenjivača-laika nad temama iz specijalizirane domene (znanost o vizualizaciji informacija), označen je mali i nereprezentativni uzorak parova te nije provedeno vrednovanje metode mjerenja. Zbog toga nedostaju informacije o pouzdanosti provedenih mjerenja pokrivenosti te nije jasno da li je opravdana složenost modela mjerenja koja umanjuje njegovu primjenjivost i reproducibilnost.

Ovdje predloženi pristup problemu pokrivenosti definira problem i razmatra ga u općenitosti, te uvodi nove metode izrade referentnih tema i mjerenja pokrivenosti. Provodi se metodološki korektna izrada i vrednovanje predloženih mjera pokrivenosti, te se temeljem tih mjera provodi vrednovanje tematskih modela s aspketa pokrivenosti. Eksperimenti su provedeni na dva skupa podataka koji predstavljaju različite tekstne domene – medijsku i biološku. Provedeno vrednovanje pokazuje da se korištenjem standarndih diskriminativnih modela i jednostavnih značajki udaljenosti tema može postići visoka točnost koja se približava međusobnom slaganju ljudskih označivača. Predlaže se jednostavna nenadzirana mjera pokrivenosti temeljena na mjeri udaljenosti tema koja odlično korelira s nadziranim mjerama koje aproksimiraju ljudsku točnost a može se brzo primijeniti na nove skupove podataka.

U [64] se predlaže metoda vizualne analize stabilnosti tematskih modela. Metoda se temelji na grupiranju tema većeg broja modela i promatranju, putem interaktivne vizualizacije, da li

su pojedini modeli naučili teme iz pojedinih grupa. Izdvajaju se grupe tema koje odgovaraju konceptima i pokazuje da pojedini modeli često propuštaju pokriti neke, osobito manje zastupljene, koncepte. Iako ne spominje problem pokrivenosti, ovo istraživanje povezano je s njim na nekoliko načina: ukazuje na nepotpunu pokrivenost koncepata od strane samo jednog modela, te ukazuje da bi se grupiranje tema više modela moglo koristiti za konstrukciju referentnih tema i kao metoda agregacije tema kojom bi se postigla veća pokrivenost.

[133] opisuju metodu analize tematskih modela temeljenu na generiranju sintetskog skupa tekstova iz skupa predefiniranih tema te mjerenju poklapanja tema naučenih iz tekstova modela s tim temama. Proces generiranja tekstova omogućava variranje količine šuma te uključuje neke pretpostavke o strukturi prirodnog jezika. Mjera poklapanja naučenih s predefiniranim temama ne modelira pokrivenost tema već se temelji na mjeri međusobne informacije (engl. *mutual information*) između dva skupa tema. Ova mjera računa se na razini temama pridruženih riječi, upravo kako bi se izbjegla potreba za izravnim uparivanjem tema. Međutim, ovaj pristup mogao bi se iskoristiti u analizama pokrivenosti sintetskog ili realnog skupa referentnih tema na temelju zbirke sintetskih tekstova generiranih iz tih tema. Pristup bi omogućio analizu pokrivenosti u idealnim uvjetima kada je tekst dobiven vjerojatnosnim generativnim procesom modela uz mogućnost kontrole količine šuma.

Istraživanje opisano u [65] razmatra problematiku i metode vezane uz odabir, vrednovanje i efikasnu izgradnju tematskih modela. Predlaže se metoda analize varijacije srodnih tema koja pokazuje da modeli istog tipa s istim parametrima mogu uz različitu slučajnu inicijalizaciju naučiti teme koje odgovaraju srodnim ali različitim aspektima istog koncepta. Ova analiza je zanimljiva iz perspektive problema pokrivenosti tema zbog toga što ukazuje na problem definiranja referentnih tema s obzirom na velik broj srodnih varijacija jednog koncepta te na vezani problem ocjenjivanja semantičkog poklapanja tema i koncepata.

Dohvat informacija

Ideja pokrivenosti tema prirodno se javlja u području dohvata informacija (engl. *information retrieval*) kod problema u kojima je cilj dohvatiti ograničeni skup tekstova koji pokrivaju što veći broj tema. Glavna razlika između ovih pristupa i problema pokrivenosti u tematskom modeliranju je što se koristi predefinirani skup tema koje odgovaraju općenitim tematskim kategorije, i te teme se koriste i kao referentni skup i za označavanje (kategorizaciju) dokumenata. Ovaj pristup čini problem mjerenja pokrivenosti trivijalnim – dovoljno je samo agregirati skup tema tj. oznaka svih dohvaćenih dokumenata. U tematskom modeliranju se pak teme uče iz zbirke tekstova, nisu predefinirane, i potrebno ih je upariti s referentnim temama nekom netrivijalnom metodom. Dodatno, sama priroda tema u referentom skupu će očekivano dosta varirati, ovisno o vrsti pokrivenosti koja se mjeri.

U [134] autori razmatraju problem odabira skupa blogova koji najbolje pokrivaju skup tema

u scenariju kontinuiranog pristizanja, odnosno toka podataka (engl. *streaming*). Teme su odabrane općenite tematske kategorije koje odgovaraju oznakama (tags) tekstova objavljenih na blogovima. Problem se rješava poopćenjem matematičkih problema pokrivenosti elemenata skupova – problema maksimalne pokrivenosti (engl. *maximum coverage*) i problema pokrivenosti skupa (engl. *set cover*) za streaming scenarij.

U [135] se riješava problem dohvata ograničenog skupa dokumenata koji pokrivaju što veći broj tema. Za vrednovanje se provodi na skupu dokumenata označenih tematskim kategorijama namjenjenim vrednovanju metoda dohvata informacija. Kao rješenje predložena je primjena strukturnog SVM modela (engl. *structural SVM*) [136]. koja rangira dokumente prema kriteriju što veće pokrivenosti odnosno veličine ukupnog skupa riječi dokumenata.

U [137] autori predlažu metodu za promjenu rangiranja (engl. *reranking*) dokumenata s ciljem postizanja da i skup odabranih dokumenata i individualni dokumenti pokrivaju što više tema. Metoda se temelji na grafu dokumenata izgrađenom iz sličnosti dokumenata na razini riječi a vrednovanje je provedeno na tri skupa dokumenata kategoriziranih u tematske kategorija.

Sažimanje teksta

Kod zadatka sažimanja teksta (engl. *text summarization*) [138] cilj je izgraditi sažeti tekstualni prikaz veće količine teksta. Pri sažimanju teksta poželjno je postići što veću pokrivenost informacija u tekstu informacija sažetkom [138]. Standardni način mjerenja pokrivenosti je mjerenje pokrivenosti na razini riječi. Za razliku od pokrivenosti u tematskom modeliranju, riječi su dobro definirane jedinice pokrivenosti i lako je izgraditi skup svih riječi u zbirci tekstova. S druge strane, slično kao kod tematskog modeliranja a za razliku od fiksnog broja tematskih kategorija kod pokrivenosti u dohvatu informacija, skup riječi je sličan temama modela po tome što nije unaprijed zadan nego ga je potrebno izgraditi. Pri tome izgradnja skupa riječi može uključivati normalizaciju i filtriranje riječi.

U [139] se predlaže algoritam za izvlačenje skupa reprezentativnih rečenica iz skupa dokumenata. Problem sažimanja se formulira kao problem cjelobrojnog linearnog programiranja (engl. *integer linear programming*) u kojemu je cilj odabrati skup rečenica maksimalne sličnosti s izvornim tekstom i ograničene ukupne duljine. Pri tome su riječima pridružene težine srodne td-idf težinama a pokrivenost se modelira putem sličnosti vektora riječi – veća pokrivenost odnosno poklapanje u skupu svih riječi većim poklapanjem odnosno sličnošću vektora.

Kod pristupa opisanog u [140] odabire se skup rečenica reprezentativnih za tekst a problem se modelira kao problem maksimalne pokrivenosti skupa s ograničenjem ranca (engl. *maximum coverage problem with knapsack constraint*). Pri tome je cilj pokriti što veći broj riječi riječima odabranih rečenica, uz dodatni cilj odabira riječi s većim težinama koje modeliraju važnost riječi i uz ograničenje na ukupni broj riječi. Razmatra se nekoliko pohlepnih algoritama te pristup grananja i rezanja (engl. *branch and bound*) za rješavanje ovog NP-teškog problema.

Primjene tematskih modela

Pregled literature završava s nekoliko primjena tematskih modela koje su na neki način povezane s idejom pokrivenosti tema.

U istraživanju opisanom u [141] uspoređuju se novinske teme s temama koje se javljaju na twitteru. Teme dobivene tematskim modelima se kombinacijom ljudskog označavanja i automatskih metoda kategoriziraju u unaprijed zadane apstraktne tematske kategorije. Zatim se promatra pokrivenost kategorija od strane pojedinih medija te se analiza razlika između novinskih tema (dobivenih iz New York Times zbirke tekstova) i twitter tema. Provodi se i finija analiza pojedinih kategorija pri čemu se za općenite kategorija analizira priroda konkretnijih tema modela dobivenih iz određenog medija. Na kraju se promatra tematska divergencija medija – za svaki od medija nalaze se teme modela slabo pokrivene od tema iz drugog medija, odnosno teme bez sličnih tema po kriteriju JS-divergencije. U kontekstu u 5.1 predložene definicije problema pokrivenosti, skup općenitih kategorija može se smatrati skupom referentnih tema, i opisani pristup stoga predstavlja jedan od mogućih pristupa problemu definicije ovog skupa. Za referentni skup tog tipa mjerenje pokrivenosti moguće je provesti putem, poželjno automatske, kategorizacije tema. U nastavku predloženi pristup pokrivenosti koristi referentni skup s temama koje po tipu i razini apstrakcije očekivano odgovaraju temama tematskih. Sličnost s pristupom iz [141] je korištenje mjera udaljenosti tema za rješavanje problema vezanih uz pokrivenost.

U [142] autori razvijaju metodu označavanja tema tematskih modela s konceptima iz DBpedie ontologije [143], s naglaskom na konstrukciji informativnih oznaka za teme modele. Međutim, razvijena metoda koja uparivanja tema modela s DBpedia konceptima mogla bi se iskoristiti za pristup pokrivenosti tema kod kojega referentne teme odgovaraju DBpedia konceptima. Pri tome se javlja problem razlike u razini općenitosti između koncepata i tema modela. Opisana metoda adresira ovaj problem tako da se definiraju dvije vrste poklapanja tema koncepata – precizno i okvirno poklapanje, koji ukazuju na dva moguća pristupa mjerenju pokrivenosti.

U [144] autori razmatraju problem dodjele recenzenata znanstvenim člancima. Jedna od razmatranih metoda temelji se na tematskim modelima – članci i recenzentni (recenzent je predstavljen tekstom s opisom ekspertize) opisuju se utežanim skupom tema tematskih modela. Zatim se za članak odabiru recenzenti čije teme najbolje pokrivaju teme iz članka. Za razliku od ovdje razmatranog problema tematske pokrivenosti, u ovom radu se teme modela promatraju kao fiksni tema koji se nastoji pokriti. Problem pokrivenosti, s druge strane, je motiviran upravo fluidnošću tema koje jedan tematski model može naučiti. No u opisanom pristupu očito je poželjno provesti tematsko modeliranje modelom s visokom pokrivenošću kako bi naučeni skup tema uistinu odgovarao što raznolikijem skupu semantičkih tema tj. koncepata.

Tablica 5.1: Dvije zbirke tekstova na kojima se temelje eksperimenti pokrivenosti.

	Medijska zbirka	Biološka zbirka
broj tekstova	24.532	5.994
veličina riječnika	23.155	6.259

5.3 Zbirke tekstova

Analize pokrivenosti tema i vrednovanja metoda mjerenja pokrivenosti provode se na dvije različite tekstne domene – medijskim i biološkim tekstovima. Kod obje tekstne domene postoji interes za eksplorativnu analizu temeljenu na tematskim modelima, a uvidi dobiveni analizom pokrivenosti mogli bi dovesti do efikasnijih modela koji otkrivaju veći broj tema u kraćem vremenu.

Ovdje se opisuju dvije zbirke tekstova korištene u eksperimentima – zbirka medijskih tekstova označena sa *medijska_zbirka*, i zbirka bioloških tekstova označena sa *bio_zbirka*. Opisuju se i metode pretprocesiranja i filtriranja tekstova koje dovode do finalnih skupova riječi i tekstova koji predstavljaju ulaz za učenje tematskih modela i izgradnju referentnih tema. Sve teme iz iste domene korištene u eksperimentima pokrivenosti – naučene teme modela i referentne teme – izgrađene su na temelju iste zbirke i međusobno su usporedive pošto su prikazane kao utežane liste istih riječi i dokumenata. Podaci o zbirkama tekstova i riječnicima koji se koriste u eksperimentima sažeti su u tablici 5.1.

Zbirka tekstova iz određene domene je temelj za izgradnju skupa podataka potrebnih za provođenje analize pokrivenosti na toj domeni. Ova analiza uključuje izradu mjera pokrivenosti i provođenje mjerenja pokrivenosti referentnih tema od strane modela. Svaki od ova dva skupa – *medijski skup podataka* i *biološki skup podataka* – sastoji se od zbirke tekstova, referentnih tema, naučenih tematskih modela odnosno njihovih tema, te parova tema označenih ocjenama poklapanja koji se koriste za izradu nadziranih mjera pokrivenosti. Svaka od ovdje opisanih zbriki tekstova predstavljaju osnovu odgovarajućeg skupa podataka – iz zbirke se uče i modeli čije teme su osnova referentnih tema (odjeljak 5.4) i tematski modeli korišteni u eksperimentima pokrivenosti (odjeljak 5.5), a nadzirane mjere pokrivenosti uče se na označenim parovima ovih tema (odjeljak 5.6).

Zbirka medijskih tekstova Analiza pokrivenosti medijskih tema provodi se na temelju zbirke američkih političkih vijesti. Tekstovi su sakupljeni praćenjem popularnih web portala u razdoblju od tri mjeseca, nakon čega je provedeno uklanjanje tekstova koji ne sadrže vijesti i deduplikacija. Konačna zbirka sadrži 24.532 tekstna dokumenta. Riječnik za izgradnju modela dobiven je pretprocesiranjem koje uključuje uklanjanje zaustavnih riječi, morfološku normalizaciju ri-

ječi, i uklanjanje nisko- i visokofrekventnih riječi. Konačni riječnik sadrži 23.155 riječi. Detalji izgradnje zbirke i riječnika mogu se naći u odjeljku 3.4.

Zbirka bioloških tekstova Analiza pokrivenosti bioloških tema provodi se na temelju zbirke tekstova o mikroorganizmima (bakterijama i arhejama) temeljem koje je provedena eksplorativna analiza fenotipova u [17]. Zbirka se sastoji od tekstova o 1640 različitih bioloških vrsta dobivenih iz pet izvora: (1) *Wikipedia* web enciklopedije, (2) *MicrobeWiki* wiki sjedišta s informacijama o mikroorganizmima, (3) *HAMAP proteomes* baze proteinskih podataka, (4) *PubMed* baze sažetaka radova te (5) *PubMed Central* baze punih tekstova članaka. Dohvat tekstova za svaku od vrsta proveden je metodama pretraživanja i mapiranja ovisnim o pojedinom izvoru. Svi izvori ne sadrže tekstove o svakoj vrsti, odnosno za neke vrste postoje tekstovi iz samo nekih izvora. Konačna zbirka sastoji se od 5994 dokumenta.

Riječnik je dobiven pretprocesiranjem tekstova koje se sastoji od filtriranja riječi i korijenovanja [17]. Filtriranje uključuje uklanjanje referenci iz tekstova radova te uklanjanje zaustavnih riječi i riječi koje se u tekstovima pojavljuju manje od četiri puta. Dodatno filtriranje riječi provedeno je prema kriteriju nužnosti pojavljivanja riječi u većem broju izvora. Naime, u eksperimentima u [17] korištena su dva riječnika, riječnik riječi koje se pojavljuju u svim izvorima tekstova te riječnik riječi koje se pojavljuju u barem četiri od pet izvora. Razlog tome je usporedivost tematskih modela korištenih za eksplorativnu analizu u [17], svaki od kojih je izgrađen iz tekstova iz samo jednog od izvora. Iako se ovdje korišteni tematski modeli izgrađeni za eksperimente pokrivenosti uče na svim tekstovima ove zbirke, koriste se originalni riječnici radi zadržavanja kompatibilnosti tema modela sa referentnim temama dobivenim iz modela originalnog eksperimenta. Preciznije, za izgradnju modela koristi se riječnik s riječima iz barem četiri od pet izvora koji obuhvaća riječi oba riječnika iz originalnog eksperimenta. Taj riječnik sadrži 6259 riječi.

5.4 Referentne teme

Skup referentnih tema sadrži teme čija pokrivenost se mjeri od strane tematskih modela i time je ključan za vrednovanje modela s aspekta pokrivenosti – modeli koji daju visoku pokrivenost su modeli koji mogu naučiti referentne teme. Kao što je opisano u odjeljku 5.1, postoji više pristupa izgradnji referentnih tema – izgradnja temeljem tema modela, izgradnja od strane ljudskih stručnjaka i izgradnja korištenjem ontologija i drugih baza znanja. Ovdje se opisuju dva srodna pristupa kojima se dobivaju referentne teme za daljnje eksperimente, pri čemu je svaki primjenjen na jednu od dvije različite tekstne zbirke, medijsku i biološku. Važna karakteristika ovih pristupa je da proizvode referentne teme koje su u dosegu standardnih tematskih modela, odnosno teme koje tematski modeli očekivano mogu otkriti. To svojstvo proizlazi iz izgradnje

referentnih tema koja je u oba slučaja provedena pregledom, interpretacijom i odabirom tema većeg broja tematskih modela od strane ljudi te korištenjem odabranih tema modela za formiranje referentnih tema. Referentne teme reprezentirane su jednako kao i teme modela – utežanom listom riječi u riječniku i utežanom listom tekstova u zbirci. To omogućuje da svi postupci računanja s temama, poput računanja udaljenosti tema i izgradnje prikaza tema za označivače, mogu jednako tretirati oba tipa tema.

Korištenje skupova referentnih tema koje modeli mogu otkriti smisleno je kod provođenja ovdje opisanih eksperimenata pokrivenosti. Takve referentne teme naime jamče da se pri analizi pokrivenosti neće dogoditi da modeli daju nisku pokrivenost jednostavno zato što referentne teme ne odgovaraju tipu tema koje modeli mogu naučiti odnosno modelirati. Istraživanje pokrivenosti tema nalazi se u ranoj fazi tako da je cilj ovdje provedenih eksperimenata ujedno i razvoj i ispitivanje metoda mjerenja pokrivenosti, a ne samo analiza pokrivenosti modela. Stoga se čini smislenim olakšati problem pokrivenosti i mjeriti pokrivenost tema u dosegu modela, dok se "teži slučajevi" referentnih tema ostavljaju za daljnja istraživanja usmjerena na otkrivanje modela pogodnih za razne tipove tema. No kao što pokazuje analiza provedena u odjeljku 3.6.2 te vrednovanje pokrivenosti opisano u odjeljku 5.8, korištene referentne teme nije jednostavno pokriti – vrednovani tematski modeli mogu precizno pokriti, u najboljem slučaju, tek oko 60% ukupnog skupa tema. Dodatno, korištene referentne teme reprezentativne su za interpretabilne teme modela poželjne u postupcima eksplorativne analize. Ove teme naime odgovaraju medijskim temama koje su se pokazale korisnima za analizu medijske agende te temama koje su izdvojene zbog poklapanja s konceptima fenotipova i korištene za daljnje analize bioloških tekstova.

Medijske referentne teme Ovaj skup referentnih tema izgrađen je pomoću semantičkih tema otkrivenih u sklopu analize medijske agende američkih političkih vijesti opisane u odjeljku 3.4. Referentne teme kao koncepti direktno odgovaraju semantičkim temama i sam postupak njihove izrade može se promatrati kao proširenje postupka otkrivanja tema predloženog u 3.3.1, odnosno kao dodatni korak tog postupka čiji cilj je poboljšanje definicije i izgradnja reprezentacije semantičkih tema. Referentna tema definira se kao lista utežanih riječi i dokumenata koji dobro opisuju odgovarajuću semantičku temu. Ove liste dobivaju se na temelju tema modela koje odgovaraju semantičkoj temi a zapisane su u tablici semantičkih tema koja je izlaz koraka otkrivanja tema opisanog u odjeljku 3.3.1. Teme modela dobivene su učenjem LDA tematskih modela sa 50 i 100 tema.

Početna točka za konstrukciju referentne teme je semantička tema – njena definicija i lista tema modela koje joj odgovaraju. Za svaku semantičku temu koja je povezana s više od jedne teme modela slučajno se odabiru dvije različite teme modela. Naime, zbog varijacije tema modela pregled različitih tema može ukazati na srodne ali različite aspekte semantičke teme i

dati bolji uvid u njenu definiciju, a zbog ubrzavanja postupka pregled je ograničen na samo dvije teme. Zatim se pregledavaju najbolje rangirane riječi i dokumenti odabranih tema modela te se odabiru one riječi koje dobro opisuju semantičku temu i oni dokumenti u kojima se semantička tema javlja. Bilježi se i podatak da li je tema bolje opisana riječima ili dokumentima, pošto je primijećeno da se apstraktne teme u pravilu se mogu bolje opisati skupom riječi dok se konkretne teme poput osoba i događaja mogu bolje opisati skupom dokumenata. Ovaj postupak proveden je od strane dva označivača, koji su najprije proveli postupak na manjem broju tema te prodiskutirali i uskladili smjernice za odabir riječi i dokumenata – odabiru se riječi specifične za temu te dokumenti u kojima se tema pojavljuje u nezanemarivoj količini teksta. Nakon usklađivanja, semantičke teme su ravnomjerno podjeljene između označivača i obrađene. Ovaj postupak doveo je do skupa od 134 referentne teme.

Rezultat ovog postupka je skup referentnih tema od kojih je svaka opisana listom riječi, listom dokumenata, te podatkom o tome koja od listi bolje opisuje temu. Konačni cilj postupka je reprezentirati svaku referentnu temu vektorom riječi u riječniku i vektorom dokumenata u zbirci, odnosno na isti način na koji su reprezentirane i teme modela. Vektor dokumenata referentne teme pridružuje jednake vjerojatnosti svakom od dokumenata iz liste, dok svi ostali dokumenti imaju vjerojatnost 0. Vektor riječi računa se kao utežana suma vektora dobivenog iz liste riječi i vektora dobivenog iz liste dokumenata. U prvom od ovih vektora sve riječi iz liste imaju jednaku vjerojatnost, dok sve ostale riječi imaju vjerojatnost 0. Drugi vektor je normalizirana suma vektora, svaki od kojih sadrži tf-idf težine riječi jednog od dokumenata iz liste. Pri tome težina ww vektora liste riječi i težina wd vektora liste dokumenata ovise o relativnoj važnosti riječi i dokumenata za opis teme. Ako su riječi važnije postavlja se $ww = 0.8, wd = 0.2$, ako su dokumenti važniji postavlja se $ww = 0.2, wd = 0.8$, a u slučaju da je njihova važnost jednaka postavlja se $ww = 0.5, wd = 0.5$.

Biološke referentne teme Drugi skup referentnih tema sastoji se od tema iz biološke domene pri čemu svaka od tema odgovara konceptu nekog fenotipa. Fenotip je karakteristika organizma, odnosno skup međusobno povezanih svojstava organizma. Ove teme konstruirane su u sklopu razvoja metodologije automatskog otkrivanja fenotipa i mapiranja fenotipa na organizme metodama strojnog učenja [17]. Konstrukcija tema provedena je treniranjem većeg broja tematskih modela na zbirci tekstova o mikroorganizmima, grupiranjem tema i odabirom onih grupa čijim pregledom je ustanovljeno da odgovaraju konceptu nekog fenotipa. Skup se sastoji od 112 referentnih tema.

Konstrukcija originalnih tema modela temelji se na pet različitih tekstnih zbirki sa tekstovima koji opisuju 1640 vrsta bakterija i arahea [17]. Na tekstovima svake od zbirki naučeno je 5 NMF modela sa 50 tema i 3 NMF modela sa 100 tema. Modeli su učeni na dokumentima predstavljenima tf-idf težinama riječi a usporedivost tema iz različitih zbirki postigla se upo-

trebom riječnika s riječima koje se javljaju u svim zbirkama. Teme različitih modela zatim su grupirane na temelju Pearsonovog koeficijenta korelacije nakon čega su zadržane samo grupe koje sadrže teme iz barem tri različite zbrike. Ovaj odabir odgovara intuiciji da bi se fenotipovi trebali konzistentno pojavljivati u različitim zbirkama. Vektori riječi tema unutar grupa su zatim uprosječeni i svaka grupa je reprezentirana utežanom listom od 20 riječi s najvećom težinom u prosječnom vektoru. Grupe su pregledane od strane domenskog stručnjaka nakon čega su zadržane grupe s interpretabilnim skupovima riječi za koje je procijenjeno da odgovaraju konceptu nekog fenotipa. Kako bi se otkrili dodatni fenotipovi prethodni postupak proveden je još jednom, s tim da se ovaj put koristio riječnik s riječima koje se pojavljuju u barem četiri od pet zbirki, a tematsko modeliranje je provedeno izgradnjom 3 NMF modela sa 100 tema. Opisani postupak rezultirao je skupom od 112 koherentnih grupa tema koje odgovaraju fenotipovima.

Skup referentnih tema konstruiran je na način da svaka referentna tema odgovara jednom fenotipu odnosno grupi tema, a izrada vektora riječi i dokumenata referentne teme provedena je na temelju originalnih podataka iz opisanog eksperimenta [17]. Vektor riječi teme konstruiran je izravno iz utežanog skupa od 20 riječi koje opisuju odgovarajuću grupu. Pri tome su korištene originalne težine riječi za grupu dok su težine preostalih riječi postavljene na nulu, nakon čega je vektor riječi normaliziran do vjerojatnosne razdiobe. Vektor težina dokumenata za referentnu temu konstruiran je na način da su za odgovarajuću grupu zbrojeni vektori dokumenata svih tema u grupi, nakon čega je provedena normalizacija na vjerojatnosnu razdiobu. Ovi vektori s dokument-tema težinama dobiveni su iz NMF modela originalnog eksperimenta. Valja napomenuti da se pri izgradnji tematskih modela za daljnje eksperimente pokrivenosti koristi zbirka tekstova nastala ujedinjavanjem svih pet opisanih zbirki iz originalnog eksperimenta. Stoga je pri opisanoj konstrukciji referentnih vektora dokumenata provedeno mapiranje koordinata vektora iz originalnih koordinata unutar pojedinih zbirki na koordinate unutar veće ujedinjene zbirke. Riječnik koji se koristi u eksperimentima pokrivenosti sadrži riječi koje se pojavljuju u 4 od 5 pod-zbirki pošto taj riječnik obuhvaća oba riječnika iz originalnih eksperimenata.

5.5 Tematski modeli za eksperimente pokrivenosti

Ključan dio eksperimenata pokrivenosti opisanih u ovom poglavlju je izgradnja skupa tematskih modela. Ovi modeli koriste se za izgradnju i vrednovanje mjera pokrivenosti – nadzirane mjere (odjeljak 5.6) se uče i vrednuju na označenim parovima tema modela a vrednovanje nenadziranih mjera (odjeljak 5.7) provodi se računanjem korelacije s nadziranim mjerama na cijelom skupu modela. Na kraju poglavlja provodi se vrednovanje ovdje opisanih modela mjerenjem pokrivenosti referentnih tema od strane tema modela. U ovom odjeljku opisuju se klase ovih modela i način njihove izgradnje.

Odabir tipova modela Kao što je opisano u odjeljku 2.1, postoji vrlo velik broj klasa tematskih modela i nužno je, radi izvedivosti i preglednosti eksperimenata, ograničiti njihov broj. Odabrano je nekoliko standardnih tipova modela koji predstavljaju različite pristupe tematskom modeliranju. Svi ovi modeli su generički u smislu da koriste minimalne pretpostavke o tematskoj strukturi teksta srodne pretpostavkama LDA modela (odjeljak 2.1). Drugim riječima, modeliraju samo strukturu povezanosti dokumenata i riječi s temama, za razliku od modela koji modeliraju i druge pojave poput sentimenta, autora, izvora vijesti ili vremena nastanka teksta. Svi korišteni modeli pretpostavljaju i da su modelirani tekstovi mješavina više tema – pretpostavlja se veći maksimalni broj tema a broj tema u pojedinim tekstovima uči se iz podataka. Opisana svojstva čine ove modele općenitima i primjenjivima na veliki broj tekstnih domena, uključujući medijske tekstove. Među korištenim modelima od osnovnog LDA modela najviše se razlikuje NMF model koji aproksimira zbirku tekstova jednostavnom nenegativnom faktorizacijom, bez dodatnih vjerojatnosnih pretpostavki.

Eksperimenti pokrivenosti provode se s ukupno četiri tipa tematskih modela, tri generativna vjerojatnosna modela i faktorizacijskim NMF modelom. Prvi tip korištenih modela je latentna Dirichletova alokacija LDA [1], detaljno opisan u 2.1. LDA je seminalni i široko korišteni tematski model koji se standardno koristi u eksperimentima vrednovanja tematskih modela i često se primjenjuje za analizu medijske agende.

Drugi tip modela je asimetrični LDA model, označen sa aLDA, modifikacija osnovnog LDA modela koja fleksibilnije modelira zastupljenost tema. Kod modela aLDA hiperparametar apriorne distribucije vjerojatnosti tema u dokumentima, u 2.1 označen sa $\vec{\alpha}$, nije unaprijed postavljen već se uči iz podataka. Na ovaj način uče se apriorne vjerojatnosti pojavljivanja pojedinih tema, odnosno modelira se mogućnost pojavljivanja zastupljenijih i manje zastupljenih tema. Ovaj pristup može dovesti do kvalitetnijih tema [145] i moguće do bolje detekcije slabo zastupljenih tema [146]. Koristi se verzija modela implementirana u hca softverskom paketu [147] koja modelira apriorne vjerojatnosti tema u dokumentima pridružujući svakoj temi t gama razdiobu definiranu parametrima a_t i b_t. Pri tome se razdioba tema u pojedinom dokumentu računa uzorkovanjem težina tema iz odgovarajućih gama razdioba i njihovom normalizacijom. Ovaj pristup naziva se normalizirana gama apriorna distribucija (engl. *normalised Gamma prior*). Tijekom učenja modela parametri gama distribucija koji određuju vjerojatnosti pojavljivanja pojedninih tema uče se iz podataka.

Treći vjerojatnosni model je neparametarski tematski model temeljen na Pitman-Yor apriornim distribucijama [147], označen sa PYP. [1] Ovaj model srodan je ranijem HDP tematskom modelu iz [148] koji se temelji na hijerarhijskom Dirichletovom procesu. Oba modela ne fiksiraju unaprijed broj tema već ga uče iz podataka. To je izvedeno uvođenjem pretpostavke o postojanju prebrojivo beskonačnog skupa tema i globalne razdiobe koja definira vjerojatnosti

[1] U [147] za ovaj model se koristi oznaka NP-LDA.

ovih tema na razini zbirke tekstova. Nadalje, generativni proces riječi u dokumentu pretpostavlja prebrojivu razdiobu ovih tema na razini dokumenta, uvjetovanu globalnom razdiobom. Svakoj se novogeneriranoj riječi unutar dokumenata ili pridružuje jedna od tema već pridruženih prethodnim riječima ili se iz dokumentne razdiobe uzorkuje nova tema. Na ovaj način ukupan broj tema ovisi temama uzorkovanim tijekom generiranja dokumenata, koje se optimiraju nekim od algoritama vjerojatnosnog zaključivanja. Drugim riječima, sam broj tema, jednako kao i njihova struktura, uči se iz strukture teksta. Prilikom implementacije ukupan broj tema se ograničava na konačan broj koji je očekivano veći od ukupnog broja tema. Na ovaj način se neke od tema uzorkuju tijekom učenja i njihova se struktura mijenja procesom učenja, dok ostale teme ostaju nenaučene odnosno postavljene na početne vrijednosti. Ovakvi neparametarski modeli fleksibilnije modeliraju ukupan skup tema u zbirci te bi mogli dovesti do bolje detekcije tema. Za razliku od HDP modela iz [148], PYP proširenje iz [147] modelira razdiobu prebrojivog skupa tema pomoću Pitman-Yor procesa [149] koji je poopćenje Dirichletovog procesa koji se koristi u [148]. Pitman-Yor proces definira vjerojatnosne razdiobe na prebrojivom skupu tema, odnosno definira vjerojatnosti pojavljivanja pojedinih tema, i određen je s dva parametra – popustom d (engl. *discount*) i koncentracijom c (engl. *concentration*). Ugrubo, koncentracija pozitivno korelira s brojem tema, dok popust kontrolira razliku između manjeg broja vrlo vjerojatnih tema i većeg broja manje vjerojatnih tema. Ovi parametri snažno utječu na strukturu naučenih tema i stoga ih je dobro i same naučiti tijekom procesa učenja umjesto da se tretiraju kao fiksni hiperparametri. Dodatno, model PYP koristi Pitman-Yor proces i za modeliranje strukutre samih tema odnosno razdiobe riječi unutar teme dok HDP model, kao i osnovni LDA model, u tu svrhu koristi manje fleksibilnu apriornu Dirichletovu razdiobu. Zbog opisanih proširenja model PYP modelira i broj i strukturu tema fleksibilnije od HDP modela. Primjerice, omogućava učenje tema vjerojatnosti čijih riječi odgovaraju Zipfovom zakonu.

Posljednji model je nenegativna matrična faktorizacija NMF, popularna alternativa vjerojatnosnim modelima. Koristi se standardni NMF model bez dodatnih pretpostavki o strukturi tema i tekstova pored temeljne pretpostavke o mogućnosti dobre aproksimacije matrice dokument-riječ nenegativnom matričnom dekompozicijom. Model NMF uči se metodom projiciranih gradijenata [150], na skupu tekstova predstavljenih vektorima tf-idf težina, uz inicijalizaciju početnih matrica nenegativnom SVD dekompozicijom [151].

Za opisane modele nadalje se koriste sljedeće oznake, navedene prema redoslijedu opisa modela: *LDA*, *aLDA*, *PYP* i *NMF*.

U eksperimentima se koristi implementacija opisanih vjerojatnosnih modela dostupna u hca[2] paketu [147]. Učenje modela temelji se na optimiranoj varijanti Gibbsovog uzorkovanja [152] nazvanoj uzorkovanje tabličnih indikatora (engl. *table indicator sampling*). Ovaj algoritam, integriran u hca paket, omogućava brži proces učenja i kvalitetnije modele [152]

[2] `https://mloss.org/software/view/527/`

Tablica 5.2: Pregled skupa modela korištenih u eksperimentima pokrivenosti.

Klasa modela	T	# slučajnih varijanti	# naučenih modela
LDA	50, 100, 200	10	30
aLDA	50, 100, 200	10	30
NMF	50, 100, 200	10	30
PYP	$T \leq 300$	10	10
Ukupan broj modela			100

a moguće ga je primijeniti za učenje nekoliko srodnih klasa parametarskih i neparametarskih modela [147]. Učenje "hiperparametara" odnosno strukturnih parametara modela poput d i c parametara Pitman-Yor procesa kod modela PYP i parametara a_t i b_t apriorne razdiobe tema-dokument težina kod modela aLDA, provodi se pomoću adaptivnog uzorkovanja odbacivanjem (engl. *adaptive rejection sampling* – ARS) [153]. Proces učenja provodi se izmjenom Gibbsovog uzorkovanja i uzorkovanja odbacivanjem – nakon završetka određenog broja Gibbs-iteracija, strukturni parametri se osvježavaju ARS metodom. Implementacija modela NMF korištena u eksperimentima je implementacija dostupna kao dio scikit-learn[3] radnog okvira [154].

Izgradnja skupa modela Izgradnja tematskih modela provedena je na zbirkama tekstova opisanim u odjeljku 5.3 – zbirci američkih medijskih tekstova i zbirci bioloških tekstova o mikroorganizmima. Za svaku od zbirki matrica težina dokument-riječ koja predstavlja ulaz za učenje modela jednaka je za sve modele. Ova matrica izgrađena je na temelju skupa dokumenata u zbirci i metode pretprocesiranja teksta odnosno rezultirajućeg skupa riječi. Isti skup dokumenata i riječnik na kojima se temelje modeli koriste se i za reprezentaciju referentnih tema (odjeljak 5.4). Posljedica je da su, za svaku od zbirki sve teme – naučene teme modela i referentne teme – međusobno usporedive pošto su reprezentirane kao utežane liste istog skupa riječi i dokumenata. Ovo svojstvo tema se koristi kod računanja udaljenosti tema te kod ekstrakcije značajki parova tema.

Skup modela koji se koriste u daljnjim eksperimentima pokrivenosti sastoji se od modela određenih sa tri parametra: klasom modela, brojem tema T, te parametrom incijalizacije modela i procesa učenja. Cijeli skup modela gradi se za svaku od zbirki tekstova.

Klasa modela je jedna od četiri prethodno opisane: LDA, aLDA, PYP i NMF. Za sve klase modela osim neparametarskog modela PYP koji sam uči broj tema, grade se varijante modela s tri različite vrijednosti broja tema T. Parametar T je važan parametar tematskih modela koji utječe na prirodu tema i njegovo variranje povećava raznolikost tema te omogućava mjerenje

[3] `https://scikit-learn.org`

odnosa pokrivenosti i broja tema. Medijski skup podataka sadrži ukupno 134 semantičke teme otkrivene pomoću modela sa $T = 50$ i $T = 100$ tema, dok biološki skup podataka sadrži 112 semantičkih tema otkrivenih modelima sa $T = 50$ i $T = 100$ tema. Za svaku klasu modela grade se modeli sa $T = 50$ tema (broj manji od broja tema u zbirci), $T = 100$ tema (broj približno jednak broju tema u zbirci) i $T = 200$ tema (broj veći od broja tema u zbirci). Iznimka je neparametarski model PYP za koji je potrebno definirati jedino maksimalni broj tema koje model može naučiti – ovaj parametar postavljen je na vrijednost 300. Za svaku kombinaciju klase modela i broja tema, naučeno je 10 modela s različitim inicijalnim parametrima generatora slučajnih brojeva (engl. *random seed*). Na ovaj način se simulira stohastičnost kod izgradnje modela, dobiva se raznolikiji skup modela i tema, i omogućava se robustnija procjena pokrivenosti koju daju pojedini tipovi modela. Tablica 5.2 sadrži pregled skupa naučenih modela.

Ostali parametri modela i algoritama učenja postavljeni su na sljedeći način. Za model LDA hiperparametri $\vec{\alpha}$ i $\vec{\beta}$ koji kontroliraju apriorne Dirichletove distribucije za vjerojatnost tema u dokumentima odnosno vjerojatnost riječi u temama definirani su prema heuristici u [7] kao uniformni vektori koji sadrže vrijednosti $\alpha = 50/T$ odnosno $\beta = 0.01$. Za model aLDA parametar $\vec{\beta}$ postavljen je na prethodno opisani način, dok se apriorne vjerojatnosti tema u dokumentima uče iz podataka a parametari a_t i b_t koji definiraju ove vjerojatnosti postavljeni su na vrijednosti $a_t = 0.5$ i $b_t = 10$ za svaku od tema t. Za model PYP potrebno je postaviti početne vrijednosti parametara koncentracije c i popusta d za četiri Pitman-Yor procesa koji definiraju globalne i lokalne dokument-tema i tema-riječ razdiobe. Za sve ove razdiobe početne vrijednosti postavljene su na $c = 10$ i $d = 0.5$.

Učenje vjerojatnosnih modela provodi se kombinacijom Gibbsovog uzorkovanja i uzorkovanja ARS metodom [147]. Gibbsovo uzorkovanje provodi tijekom većeg broja ciklusa (nekoliko stotina), pri čemu se periodički, nakon završetka manjeg broja ciklusa, ARS metodom računaju nove vrijednosti strukturnih parametara (a_t i b_t za aLDA te svi parametri popusta i koncentracije za PYP). Potrebno je definirati maksimalni broj ciklusa Gibbsovog uzorkovanja C i period P računanja strukturnih parametara ARS metodom, kojeg je moguće definirati odvojeno za svaki od parametara. Maksimalni broj ciklusa C postavljen je na velik broj pošto cilj nije brzina već zagarantirana konvergencija Gibbsovog uzorkovanja odnosno učenje kvalitetnih modela. Za parametarske modele LDA i aLDA koristi se vrijednost $C = 800$, dok se za neparametarski model PYP s većim brojem tema i parametara koristi vrijednost $C = 1500$. Promatranjem konvergencije strukturnih parametara za nekoliko različitih vrijednosti parametra P odabrana je vrijednost $P = 11$ za parametre popusta i parametre a_t, dok je zbog veće brzine njihovog uzorkovanja za parametre koncentracije te za parametre b_t odabrana vrijednost $P = 3$. Promatranjem tema naučenih modela i promjene vrijednosti strukturnih parametara tijekom procesa učenja utvrđeno je da definirani postupak učenja dovodi do modela s interpretabilnim temama i da strukturni parametri modela aLDA i PYP konvergiraju brzo, unutar nekoliko desetaka do stotinu ciklusa.

Učenje modela NMF provedeno je prema [16], eksperimentu eksplorativne analize zbirke govora u Europskom parlamentu. Ulaz za algoritam učenja je zbirka tekstova predstavljena kao dokument-riječ matrica s tf-idf težinama. Model NMF se uči algoritmom koji je brza varijanta optimizacijske metode projiciranih gradijenata (engl. *projected gradient method*) koja je razvijena za nenegativnu matričnu faktorizaciju [150]. Radi ubrzanja učenja i dobivanja kvalitetnijih modela faktorizacijske matrice inicijaliziraju se metodom nenegativne SVD dekompozicije [151]. Sve metode korištene za učenje modela NMF dostupne su kao dio scikit-learn radnog okvira.

Svi modeli izgrađeni su pomoću radnog okvira opisanog u poglavlju 7 koji sadrži alate za izgradnju, pohranu i dohvat većeg broja tematskih modela. Izgradnja je provedena na serverskom stroju upotrebom Docker virtualizacijske platforme[4] – softverska okolina za izgradnju modela s Python ekosustavom i potpornim knjižnicama zapakirana je kao Docker kontejner.

5.6 Mjere pokrivenosti temeljene na nadziranom učenju

Mjere pokrivenosti računaju ocjenu pokrivenosti skupa referentnih tema od strane tema modela. U ovom poglavlju se predlažu dvije metode mjerenja pokrivenosti – pristup temeljen na nadziranim modelima koji aproksimiraju ljudske ocjene poklapanja tema i nenadzirani pristup temeljen na ocjenjivanju poklapanja tema pomoću mjere udaljenosti tema.

Ovdje se opisuje nadzirani pristup mjerenju pokrivenosti koji aproksimira postupak ocjenjivanja pokrivenosti referentnih tema proveden od strane ljudi. Ovaj postupak sastoji se od razmatranja, za svaku referentnu temu, svih tema modela i ocjenjivanja poklapaju li se te teme odnosno predstavljaju li isti koncept. Pri tome ocjenjivanje poklapanja provode ljudi na temelju pregleda i interpretacije tema. Iako ovaj postupak ocjenjivanja daje zlatni standard ocjena pokrivenosti, njegova provedba je vremenski vrlo zahtjevna pošto skup referentnih tema može sadržavati više od sto tema, a modeli očekivano sadrže od nekoliko desetaka do nekoliko stotina tema. Čak i uz filtriranje tema modela izglednih za poklapanje s referentnom temom, primjerice pomoću mjere udaljenosti tema, vremenska zahtjevnost zadatka ostaje velika. Potrebno vrijeme se dodatano povećava kada se u obzir uzme potreba za ocjenjivanjem većeg broja modela – različitih tipova modela te modela koji odgovaraju različitim vrijednostima hiperparametara i parametara izgradnje.

Predloženo rješenje ovog problema je izrada nadziranog modela koji aproksimira ljudsku intuiciju jednakosti tema. Ovaj model uči se na skupu parova tema označenih ljudskim ocjenama poklapanja. Postupak ljudskog ocjenjivanja poklapanja temelji se na definciji poklapanja i definiciji ocjena koje mogu biti binarne, ordinalne ili numeričke. Nadzirani model trebao bi biti općenit u smislu da može točno ocjenjivati parove raznorodnih tema – referentnih tema i

[4] `https://www.docker.com/`

tema dobivenih raznim vrstama tematskih modela. Kada je nadzirani model naučen, mjerenje pokrivenosti skupa referentnih tema od strane tema modela provodi se simulacijom postupka ocjenjivanja od strane ljudi – za svaku referentnu temu i svaku temu modela poklapanje se ocjenjuje nadziranim modelom. Konačna mjera pokrivenosti računa se kao broj ili udio referentnih tema za koje postoji poklapanje s nekom od tema modela.

Predloženi pristup definira poklapanje tema kao jednakost koncepata dobivenih interpretacijom tema. Parovi tema ocjenjuju se prema ordinalnoj skali koja sadrži vrijednosti 0 (izostanak poklapanja), 0.5 (poklapanje uz šum ili malu semantičku varijaciju) i 1 (poklapanje tema, odnosno jednakost koncepata). Svaki par tema ocjenjuje se od strane tri označivača što omogućava uprosječivanje ocjena koje očekivano povećava kvalitetu konačne ocjene. Gradi se nadzirani model binarne klasifikacije naučen na parovima tema označenim binarnim oznakama koje odgovaraju poklapanju tema i nepoklapanju tema. Pri tome se binarne oznake poklapanja formiraju uprosječivanjem ljudskih ocjena poklapanja i primjenom praga koji definira prosječnu ocjenu potrebnu za poklapanje tema. Variranjem ovog praga grade se dva nadzirana modela – prvi koji aproksimira *strogo poklapanje* bez šuma i semantičkih varijacija i drugi koji aproksimira *relaksirano poklapanje* tema. Ova dva modela definiraju dvije mjere pokrivenosti, strogu i relaksiranu, koje zajedno daju više informacija o prirodi pokrivenosti tema.

U nastavku odjeljka prvo se opisuje izrada reprezentativnog skupa parova tema koji sadrži referentne teme i teme različitih tipova modela. Zatim se opisuje koncept i provedba označavanja parova tema ljudskim ocjenama poklapanja nakon čega je opisan postupak izgradnje nadziranih modela.

5.6.1 Izrada skupa parova tema

Skup za učenje nadziranih modela je skup parova tema označenih ocjenama poklapanja tema. Ovdje se opisuje izrada skupa parova, dok se postupak označavanja opisuje u odjeljku 5.6.2. Skup parova sastoji se od tema naučenih tematskim modelima opisanim u odjeljku 5.5 te od referentnih tema opisanih u odjeljku 5.4. Cilj je naučiti općenite modele koji dobro procjenjuju poklapanje raznolikih tema, tako da skup parova sadrži teme tematskih modela iz različitih klasa izgrađene sa različitim parametrima, kao i referentne teme.

Slučajno odabran skup parova tema očekivano je visoko uneravnotežen – velika većina parova iz takvog skupa sastoji se od tema koje se ne poklapaju. Lako je demonstrirati uzrok ove pojave na idealnom scenariju tematskog modeliranja. Ovaj scenarij pretpostavlja da su naučeni tematski modeli sa T tema, da su sve teme modela različite semantičke teme (bez šuma i drugih grešaka) i da cijeli skup semantičkih tema također ima točno T tema. Tada je za slučajno odabrani par tema modela vjerojatnost poklapanja je $2/(T-1)$, odnosno reda veličine $1/T$. U realnoj situaciji kada teme više variraju i sadrže greške, te očekivano postoji više semantičkih tema nego tema modela, ova vjerojatnost je očekivano još manja. Pregled preliminarnog uzorka

parova tema potvrđuje ovu procjenu.

Iz perspektive strojnog učenja to znači da je problem učenja poklapanja tema u domeni tzv. neuravnoteženog učenja (engl. *imbalanced learning*) [155, 156]. Neuravnoteženo učenje je scenarij s malim udjelom pozitivnih primjera za učenje u cijelom skupu podataka, što u slučaju parova tema odgovara malom udjelu parova tema koje se poklapaju. Mali udio pozitivnih primjera otežava učenje dobrih modela pošto slučajni uzorak za učenje očekivano ne sadrži dovoljno podataka da se dobro opiše problem odnosno da se definira pozitivna klasa. Razvijeno je mnogo pristupa s ciljem rješavanja i ublažavanja ovog problema, poput metoda prilagodbe originalnog neuravnoteženog uzorka, algoritama učenja prilagođenih neuravnoteženom scenariju, te aktivnog učenja [156].

Međutim, u ovdje razmatranom slučaju kada su primjeri za učenje parovi tema, moguće je iskoristi neku od mjera udaljenosti tema za izgradnju uravnoteženijeg uzorka parova. Intuicija iza ovog pristupa je da, u slučaju pogodne mjere udaljenosti, mala udaljenost dviju tema odgovara višoj vjerojatnosti semantičkog poklapanja tema. Stoga se pristup temelji na uzorkovanju većeg broja parova međusobno bližih tema. Koristi se kosinusna udaljenost (engl. *cosine distance*), mjera korisna za razne zadatke dubinske analize teksta [157] za koju se pokazalo da dobro korelira s ljudskom intuicijom o sličnosti tema [29].

Korištena metoda uzorkovanja koja koristi kosinusnu udaljenost za rješavanje problema neuravnoteženosti radi na sljedeći način. Interval $[0, 1]$ svih mogućih vrijednosti kosinusne udaljenosti podjeli se na K jednakih djelova a parovi tema se na temelju udaljenosti tema rasporede u ove podintervale. Nakon toga se iz svakog od podintervala uzorkuje jednak broj parova. Preliminarni eksperiment potvrdio je da se na opisani način dobiva uravnoteženi uzorak. Parovi tema modela LDA i NMF sa 50 i 100 tema uzorkovani su prema kosinusnoj udaljenosti i uzorak od 300 parova označen je binarnim ocjenama poklapanja (iste ili različite teme). Izračunato je da 36% parova sadrži jednake teme, što je udio koji predstavlja značajno poboljšanje u odnosu na očekivani udio parova jednakih tema u potpuno slučajnom uzorku, koji prema prethodnim razmatranjima iznosi približno 1%.

Uzorak parova korišten za izgradnju nadziranih modela formiran je na sljedeći način prema opisanom postupku. Kako bi se naučili modeli koji mogu ocjenjivati poklapanje raznolikih tema, formiran je skup tema koji sadrži sve tipove modela sa svim vrijednostima broja tema (sažetak u tablici 5.2), kao i referentne teme. Preciznije, za svaku kombinaciju klase modela i broja tema slučajno je uzorkovan jedan tematski model i sve njegove teme dodane su u skup tema. Zatim su u skup dodane referentne teme, pri čemu je svaka ponovljena tri puta kako bi broj referentnih tema bio približno jednak broju tema svake od klasa modela. Nakon toga je izgrađen skup svih različitih parova opisanih tema koji je prema kosinusnoj udaljenosti tema podijeljen u 10 podintervala jednake širine. Iz svakog podintervala slučajno je uzorkovano 50 parova, što je rezultiralo skupom od 500 parova. Skup od 500 parova podijeljen je, slučajnim

uzorkovanjem, u blokove od 50 parova. Ovi blokovi, svaki od kojih je reprezentativan za cijeli skup parova, korišteni su u procesu označavanja parova. Razlog za podjelu na blokove je što prije početka označavanja nije bilo poznato koliko će se ukupno parova morati označiti zbog mogućnosti opetovanog ponavljanja koraka kalibracije označivača. Radi toga je kreiran veći broj parova od potrebnog a prema potrebi su u proces označavanja uključivani novi blokovi od 50 parova.

5.6.2 Označavanje parova tema

Označavanje parova tema ocjenama poklapanja provedeno je za svaki od skupova podataka, medijski i biološki, prema metodologiji analize sadržaja [38]. Početna točka označavanja je definicija poklapanja tema temeljena na jedakosti interpretiranih koncepata. Označivači su prvo upoznati s definicijom i postupkom označavanja nakon čega je provedena dodatna kalibracija primjene postupka diskusijom zajednički označenih parova. Nakon treniranja i kalibracije, uzorak od 300 parova tema označen je od strane troje označivača.

Definicija poklapanja parova tema korištena za označavanje određuje prirodu poklapanja koju aproksimira nadzirani model. Poklapanje je definirano kao konceptualna jednakost tema – dvije teme smatraju se jednakima ako se mogu interpretirati kao isti koncepti, pri čemu interpretirani koncepti moraju biti precizni odnosno specifični. Ovaj pristup odabran je kako bi se pokrivenost referentnih tema mjerila preciznim poklapanjem. Drugim riječima, modeli s visokim ocjenama bi trebali naučiti teme koje dobro odgovaraju referentnim temama, a ne samo srodne teme poput pod-koncepta, nad-koncepata i drugih semantički sličnih koncepata. Ovaj pristup omogućava mjerenje sposobnosti tematskog modela da precizno otkrije referentne teme. Ovisno o prirodi pokrivenosti koja se želi mjeriti, poklapanje tema moguće je definirati na temelju drugih vrsta semantičkog poklapanja.

Potrebno je definirati jedakost dvaju tema – konstrukata definiranih utežanim listama riječi i dokumenata. Postoje dvije dimenzije razlikovanja tema – semantička dimenzija koja se odnosi na razliku koncepata dobivenih interpretacijom tema, te stohastička dimenzija koja je posljedica slučajnosti u izgradnji tematskih modela koja dovodi do šuma odnosno pojave slučajnih riječi i dokumenata. Sa semantičkog aspekta jednakost se definira kao poklapanje koncepata koji odgovaraju temama i pri tome se zatijeva što specifičnija interpretacija tema. Primjerice, ako jedna tema odgovara događaju “izborna kampanja političara X”, a druga događaju “izborna kampanja političara Y” teme se ne smatraju istima, pošto najspecifičnija interpretacija tema nije “izborna kampanja”. Slično, ako se pojave teme “izborna kampanja političara X” i “skandal političara X”, teme se ne smatraju istima, što bi se moglo tvrditi u slučaju da su interpretirane kao “političar X”. Treba napomenuti da su sve prethodne interpretacije legitimne i mogle bi se iskoristiti za druge definicije poklapanja tema. Stohastičkoj dimenziji razlikovanja tema pristupa se na način da se omogućava označavanje tema “jednakima uz postojanje šuma”. Ova

oznaka se koristi ukoliko kod jedne ili kod obje teme postoji uočljiva količina šuma no teme je moguće interpretirati a interpretirani koncepti se semantički poklapaju na prethodno opisani način. Preliminarno označavanje pokazalo je da se u malom broju slučajeva pojavljuju teme koje odgovaraju semantički vrlo bliskim konceptima odnosno manjim varijacijama tj. srodnim aspektima istog koncepta. U tom slučaju koristi se ista oznaka kao i kod jednakosti uz postojanje šuma.

Na temelju prethodnih razmatranja, svaki par tema označava se jednom od tri oznake: 1 (jednakost tema), 0.5 (jednakost tema uz pojavu šuma, vrlo visoka semantička sličnost tema), i 0 (različite teme). Preliminarni eksperimenti pokazali su da je ovakvo označavanje prirodnije i jednostavnije od označavanja na binarnoj skali (jednakosti i različitost tema), upravo zbog postojanja treće kategorije koja opisuje djelomičnu jednakost.

Označavanje parova tema provedeno je primjenom metoda analize sadržaja [38] razvijenih kako bi se osigurala kvaliteta i konzistentnost označenih podataka – osigurana je upoznatost označivača s tematikom tekstova i tema, izrađene su precizne upute za označavanje, a vještina označivača unaprijeđivana je treniranjem sve dok mjera međusobnog slaganja nije dosegla zadovoljavajuću razinu.

Označavani su parovi tema modela iz dvije tekstne domene – novinske i biološke. Nužan uvjet za kvalitetu označavanja je upoznatost označivača s tekstnom domenom [38]. Novinske teme označene su od strane dviju autora istraživanja agende američkih medija koji su u sklopu te studije proveli postupak otkrivanja tema, te od strane studenta anglistike upoznatog s američkom politikom. Biološke teme označene su od strane biološke znanstvenice i dvoje studenata viših godina biologije.

Formirane su precizne upute za označavanje koje opisuju prethodnu definiciju jednakosti tema te sadrže primjere parova i razna pojašnjenja vezana uz proces označavanja. Označavan je skup parova tema opisan u odjeljku 5.6.1 koji sadrži teme modela i referentne teme. Svaka tema je prikazana listom od 15 najbolje rangiranih riječi za temu te listom od 15 najbolje rangiranih dokumenata za temu. Dokumenti su predstavljeni informativnim sažecima – u slučaju novinskih tekstova sažeci su naslovi, a u slučaju bioloških tekstova sažeci su fragmenti originalnog teksta dobiveni heurističkim metodama poput selekcije početnog dijela teksta. Označivačima su bili dostupni i puni tekstovi dokumenata.

Postupak treniranja označivača, čija svrha je osiguranje kvalitete i konzistentnosti označavanja odnosno dosljednog i preciznog provođenja uputa, proveden je prema uputama iz [38]. U svakom od koraka označavanja, svi parovi tema označeni od strane svo troje označivača. Kvaliteta označavanja mjerena je Krippendorffovim α koeficijentom međusobnog slaganja označivača (nadalje α koeficijent). Ovaj koeficijent računa mjeru poklapanja oznaka pri čemu u obzir uzima mogućnost slučajnog poklapanja [38]. Koriste se dvije verzije koeficijenta, nominalna verzija koja razlikuje jedino poklapanje ili nepoklapanje oznaka te ordinalna verzija koja

Tablica 5.3: Međusobno slaganje troje označivača na skupu od 300 parova tema, mjereno nominalnim i ordinalnim Krippendorfovim koeficijentom α.

	Medijske teme	Biološke teme
α_N	0.689	0.648
α_O	0.865	0.797

uzima u obzir i udaljenost oznaka na ordinalnoj skali (0, 0.5 i 1). Ove dvije verzije nadalje su označene sa α_N i α_O.

Postupak označavanja proveden je na način da su označivačima prvo objašnjene upute, nakon čega je označen manji "orijentacijski" skup od 15 parova tema, i razjašnjene su eventualne nejasnoće vezane uz upute i proces označavanja. U sljedećem koraku označen je kalibracijski skup od 50 parova tema i izračunati su α koeficijenti – za novinske teme dobiveni su α_N =0.568 i α_O =0.831, a za biološke teme α_N =0.576 i α_O =0.712. Potom su označivači zajedno s istraživačima prodiskutirali način primjene uputa na parove za koje je postojalo veće neslaganje. Većina grešaka označavanja uzrokovana je ili nerazumijevanjem postupka interpretacije tema modela ili površnošću u postupku pregleda i interpretacije tema. Nakon diskusije označivači su stekli mnogo bolji osjećaj za problem jednakosti tema i provedbu uputa. Potom je označen testni skup od 50 parova i izračunati su novi α koeficijenti – za novinske teme dobiveni su α_N =0.727 i α_O =0.851, a za biološke teme α_N =0.599 i α_O =0.782. U oba slučaja došlo je do poboljšanja slaganja što je očekivana posljedica razjašnjavanja uputa i postupka njihove primjene tijekom diskusije. Koeficijenti slaganja na testnim skupovima procijenjeni su dovoljno dobrim za nastavak označavanja. Zbog toga je, za svaku domenu, testni skupo od 50 tema uključen u konačni skup označenih parova a dodatno je označeno još 250 parova tema kako bi se dobio konačni skup od 300 označenih parova.

Tablica 5.3 sadrži, za obje domene, α koeficijente slaganja na konačnim skupovima označenih parova. Ovi koeficijenti prema [84] odgovaraju "značajnom" slaganju što potvrđuje iskustva iz postupka označavanja. Naime, označivači su ocijenili definiciju jednakosti tema i postupak njene primjene razumljivima i smislenima. Niže ocjene slaganja za parove iz biološke domene posljedica su toga što biološke teme, za razliku od većine novinskih tema, odgovaraju složenim i apstraktnim pojmovima koje je teže interpretirati na temelju skupa riječi i dokumenata. Koeficijenti slaganja i iskustva iz postupka označavanja potvrđuju smislenost korištene definicije poklapanja tema i kvalitetu oznaka, koje stoga predstavljaju dobar temelj za učenje nadziranih modela.

5.6.3 Nadzirani modeli za poklapanje tema

Ovdje se opisuje postupak izgradnje nadziranih modela za aproksimaciju ljudskih ocjena poklapanja parova tema. Ovi nadzirani modeli su osnova metode mjerenja pokrivenosti koja ih koristi kako bi odredila, za svaku referentnu temu, postoje li teme modela koje se poklapaju s njom. Nazdirani modeli uče se na temelju prethodno opisanog skupa parova tema označenih ocjenama poklapanja. Pri tome je zadatak automatskog ocjenjivanja poklapanja postavljen kao zadatak binarne klasifikacije parova tema u dvije klase koje odgovaraju poklapanju tema i nepoklapanju tema. Ove binarne oznake formiraju se na temelju uprosječenih ljudskih oznaka poklapanja. Pri tome se definiraju dva tipa binarnih oznaka, kod kojih pozitivna klasa (poklapanje tema) odgovara ili strogom poklapanju (viši prosjek ljudskih ocjena) ili relaksiranom poklapanju (niži prosjek ljudskih ocjena). Razmatraju se četiri standardna modela binarne klasifikacije a odabir modela i hiperparametara provodi se, za svaki od dva opisana klasifikacijska problema i za svaki od dva skupa podataka, postupkom unakrsne provjere s pet preklopa (engl. *five fold cross-validation*).

Strogo i relaksirano poklapanje tema Definiraju se dva klasifikacijska problema koji odgovaraju dvama načinima definicije binarnih oznaka poklapanja tema (poklapanje i nepoklapanje) na temelju ljudskih oznaka poklapanja. Svaki par tema označen je od strane troje označivača jednom od tri oznake: 1 (poklapanje koncepata dobivenih interpretacijom tema), 0.5 (poklapanje uz postojanje šuma ili malu semantičku varijaciju koncepata), i 0 (bez poklapanja). Binarne oznake definiraju se uprosječivanjem ljudskih oznaka i primjenom praga odluke koji definira minimalni prosjek potreban za poklapanje. Jedan klasifikacijski problem definira poklapanje tema u slučaju je prosjek oznaka 0.75 i viši, odnosno ako je barem dvoje označivača označilo teme sa 1, oznakom dobrog poklapanja. Ovo poklapanje naziva se *strogo poklapanje* i odgovarajuća mjera pokrivenosti označena je sa *NP-strog*. Drugi klasifikacijski problem definira poklapanje tema u slučaju da je minimalni prosjek oznaka 0.5, što odgovara slučaju kada sve oznake troje označivača 0.5, i slučaju kada su oznake troje označivača 0, 0.5 i 1. Ovo poklapanje naziva se *relaksirano poklapanje* i odgovarajuća mjera pokrivenosti označena je sa *NP-relaks*.

Modeli naučeni za aproksimaciju strogog odnosno relaksiranog poklapanja koriste se za mjere pokrivenosti koje ocjenjuju pokrivenost referentnih tema primjenom dva različita kriterija. Ocjene pokrivenosti dobivene s ove dvije vrste mjera nude preciznije uvide u prirodu pokrivenosti (pokrivenost uz strogo ili relaksirano poklapanje tema) koju daju pojedini modeli.

Klasifikacijski modeli. Za računanje poklapanja tema razmatraju se četiri standardna klasifikacijska modela: logistička regresija (engl. *logistic regression*) [158], stroj potpornih vektora (engl. *support vector machine*) [159], slučajna šuma (engl. *random forest*) [160] i višeslojni perceptron (engl. *multilayer perceptron*) [158]. Koriste se implementacije ovih modela dos-

Tablica 5.4: Nadzirani modeli korišteni za aproksimaciju poklapanja tema i vrijednosti hiperparametara razmatrane kod odabira modela.

Model	Hiperparametar	Vrijednosti
`LogReg`	C	$0.001, 0.01, 0.1, 1.0, 10, 100, 1000$
	regMetrika	`l1`, `l2`
`MLP`	*sirinaSloja*	$3, 5, 10$
	C	$0.00001, 0.0001, 0.001, 0.01, 0.1$
`RandForest`	*brojStabala*	$10, 20, 50, 100$
	brojZnacajki	2, 50%, `sve`
	maxDubina	2, 3, `BezOgranicenja`
`SVM`	C	$0.001, 0.01, 0.1, 1.0, 10, 100, 1000$
	γ	$0.001, 0.01, 0.1, 1.0, 10, 100, 1000, 1/BrojZnacajki$

tupne kao dio radnog okvira scikit-learn[5] [154].

Navedeni modeli primjenjuju se na zadatak binarne klasifikacije parova tema u klase koje označavaju poklapanje i nepoklapanje tema. Za potrebe opisa tih modela koriste se sljedeće oznake. Kod zadatka binarne klasifikacije cilj je naučiti mapiranje s ulaznih podataka (u ovom slučaju, parova tema) predstavljenih vektorom značajki x na oznake klasa $y \in \{0, 1\}$. Vrijednost $y = 1$ označava pozitivnu klasu (poklapanje tema), a $y = 0$ označava negativnu klasu (nepoklapanje tema). Jedan ulazni podatak, vektor x, uobičajeno se naziva i primjerom.

Logistička regresija je jednostavan klasifikacijski model koji modelira $p(y = 1|x)$, vjerojatnost pozitivne klase za primjer x, funkcijom $f(ax + b)$, pri čemu je f standardna logistička funkcija, a vektori a i b definiraju afinu transformaciju primjera odnosno vektora značajki. Parametri a i b mogu se naučiti standardnim optimizacijskim metodama, pri čemu se koristi regularizacija kako bi se izbjegla prenaučenost (engl. *overfitting*). U primjeni na klasifikaciju parova tema, hiperparametri logističkog modela koji se optimiraju su regularizacijska konstanta C i norma *regMetrika* koja se primjenjuje na a i b pri regularizaciji, a koja odgovara ili L_1- ili L_2-normi. Za učenje se koristi optimizacijski algoritam iz LIBLINEAR paketa [73] implementiran u radnom okviru scikit-learn.

Višeslojni perceptron je klasifikacijski model koji modelira vjerojatnosti klasa funkcijskom transformacijom $f(x)$, pri čemu je f kompozicija niza funkcija g koje odgovaraju "slojevima" transformacije. Svaki sloj vrši funkcijsku transformaciju oblika $g(x) = act(Ax + b)$ koja se sastoji od kompozicije afine transformacije ulaza i primjene aktivacijske funkcije *act* na transformirani vektor. Nelinearne aktivacijske funkcije omogućuju višeslojnom perceptronu mo-

[5] `https://scikit-learn.org`

deliranje složenih nelinearnih funkcijskih veza. U primjeni na klasifikaciju parova tema zbog malog broja primjera za učenje (300 parova) koristi se jednoslojni perceptron kako bi se ograničio broj parametara modela. Iz istog razloga, za učenje se koristi optimizacijska metoda LBFSG [161] a ne neka od metoda gradijentnog spusta. Regularizacija modela provodi se L_2 regularizacijom. Hiperparametri koji se optimiraju su regularizacijska konstanta C i širina skrivenog sloja $sirinaSloja$.

Slučajna šuma je model koji gradi veći broj klasifikacijskih stabala odluke (engl. *decision tree*) i glasanjem kombinira njihova predviđanja u konačnu odluku o pripadnosti primjera klasi. Pojedina stabla su "slabi" klasifikatori očekivano niske točnosti, tipično stabla ograničene dubine naučena na skupu primjera dobivenom slučajnim uzorkovanjem s ponavljanjem. Dodatni faktor stohastičnosti ovih stabala je i slučajni odabir skupa značajki koje se koriste za učenje stabala, odnosno za grananje čvora stabla tijekom procesa učenja. Pokazano je, teorijski i eksperimentalno, da se kombiniranjem većeg broja opisanih slabih klasifikatora postižu dobre klasifikacijske performanse [160]. U primjeni na klasifikaciju parova tema optimiraju se sljedeći hiperparametri: broj stabala $brojStabala$, broj značajki $brojZnacajki$ i maksimalna dubina stabla $maxDubina$. Pri izgradnji stabala kao kriterij kvalitete za grananje čvorova koristi se gini-indeks.

Stroj potpornih vektora [159] (engl. *support vector machine* – SVM) je model koji pretpostavlja da se granica između dviju klasa u prostoru značajki može dobro aproksimirati hiperravninom. Učenje modela svodi se na procjenu parametara hiperravnine koja uz točnost klasifikacije zadovoljava i uvjet maksimizacije margine udaljenosti između primjera i hiperravnine. Maksimizacija margine poboljšava generalizacijske performanse modela. Iako osnovni SVM model pretpostavlja linearnu hiperravninu, druge varijante modela pretpostavljaju nelinearne hiperravnine koje omogućuju modeliranje složenijih odnosa između klasa. Nelinearnost se postiže korištenjem tzv. jezgrenog trika – skalarni produkt između vektora primjera zamjenjuje se poopćenom jezgrenom funkcijom sličnosti. U primjeni na klasifikaciju parova tema kao jezgrena funkcija koristi se radijalna bazna funkcija (engl. *radial basis function*), a optimirani hiperparametri su parametar bazne funkcije γ i regularizacijska konstanta C koja definira relativne težine pogreške modela i širine margine pri definiciji funkcije gubitka.

Opisani modeli i hiperparametri te vrijednosti hiperparametara razmatrane u procesu odabira modela prikazani su u tablici 5.4. Navedeni hiperparametri modela klasifikacije procijenjeni su kao najvažniji i stoga se provodi optimizacija njihovih vrijednosti. Međutim, opisani modeli imaju i druge hiperparametre čije vrijednosti su postavljene na zadane vrijednosti (engl. *default value*) definirane implementacijom modela u scikit-learn radnom okviru. Te vrijednosti su definirane tako da rezultiraju kvalitetno naučenim modelima – dobre zadane vrijednosti koje olakšavaju izgradnju modela jedan su od principa oblikovanja scikit-learn radnog okvira [154].

Značajke Nadzirani modeli uče veze između značajki parova tema i oznaka klasa (poklapanje ili nepoklapanje) iz podataka. Značajke bi trebale sadržavati informacije o paru tema koje omogućuju kvalitetnu aproksimaciju poklapanja. Preliminarni eksperimenti pokazali su da značajke dobivene konkatenacijom vektora riječi i/ili dokumenata dviju tema daju relativno niske klasifikacijske performanse – F_1 ocjene između 0.4 i 0.6. Ovo se može pripisati tzv. prokletstvu dimenzionalnosti (engl. *curse of dimensionality*) – degradaciji klasifikacijskih performansi zbog visoke dimenzionalnosti primjera (u ovom slučaju više desetaka tisuća dimenzija) i razmjerno malog ukupnog broja primjera za učenje (u ovom slučaju, nekoliko stotina) [157]. Jedno moguće rješenje za ovaj problem je ekstrakcija značajki [157] (engl. *feature extraction*), odnosno transformacija visokodimenzionalnih reprezentacija u manji skup značajki koje predstavljaju kvalitetne informacije za aproksimaciju klase [157]. Postojeći eksperimenti s tematskim modelima pokazali su da mjere udaljenosti vektora riječi tema mogu dobro aproksimirati semantičku sličnost tema [29, 64, 65, 141]. Preliminarni eksperimenti sa značajkama dobivenim primjenom raznih mjera udaljenosti na parove vektora riječi i vektora dokumenata pokazali su obećavajuće rezultate – F_1 ocjene između 0.7 i 0.8. Iz navedenih razloga, odabran je ovaj pristup ekstrakciji značajki.

Mjere udaljenosti korištene za izradu značajki trebale bi imati dva svojstva. Prvo svojstvo je postojanje korelacije između udaljenosti vektora koji predstavljaju teme i ljudske intuicije o semantičkoj sličnosti tema. Primjerice, mjere udaljenosti trebale bi odražavati činjenicu da ljudi donose sud o jednakosti tema na temelju relativnih odnosa riječi i dokumenata, a ne na temelju iznosa težina riječi i dokumenata za teme. Drugo svojstvo je neosjetljivost mjere udaljenosti na informacije u vektorima koje su posljedica tipa tematskog modela a ne semantičkih svojstava tema. Ovo svojstvo je važno pošto je cilj izgradnja robustnih klasifikacijskih modela primjenjivih na teme modela više različitih klasa. Primjerice, za vjerojatnosne modele vektori tema sadrže vjerojatnosti tj. brojeve unutar intevala $[0, 1]$, dok faktorizacijski NMF model ne nameće druga ograničenja osim nenegativnosti na vrijednosti vektora koje stoga mogu poprimiti značajno veće vrijednosti.

Na temelju opisanih smjernica odabrane su četiri mjere udaljenosti: kosinusna udaljenost, Hellingerova udaljenost [162], L_1 udaljenost i L_2 udaljenost. Ove četiri mjere predstavljaju različite familije mjera: kutnu udaljenost, udaljenost vjerojatnosnih razdioba, te dvije standardne geometrijske koordinatne udaljenosti. Vektori riječi i dokumenata se prije primjene svih mjera normaliziraju na vjerojatnosnu razdiobu. Ovaj korak je nužan radi ispravne primjene vjerojatnosne Hellingerove udaljenosti, a u slučaju L_1 i L_2 mjera na ovaj se način dobiva veća neosjetljivost na klasu modela. Kosinusna, Hellingerova i normalizirana L_1 udaljenost pobliže se analiziraju u sljedećem odjeljku 5.7 i pokazuje se da odlično koreliraju sa semantičkom sličnosti tema te da nisu osjetljive na razlike u klasama tematskih modela. Normalizirana L_2 udaljenost također je neosjetljiva na klasu modela i daje dobre informacije za klasifikaciju parova tema

Tablica 5.5: Performanse nadziranih modela za procjenu poklapanja parova tema.

Medijski skup podataka	Strogo poklapanje		Relaksirano poklapanje	
	F_1	σ	F_1	σ
ljudi	0.854	0.034	0.894	0.005
LogReg	**0.835**	0.079	0.846	0.050
MLP	0.814	0.106	**0.852**	0.039
RandForest	0.804	0.110	0.837	0.021
SVM	0.830	0.077	0.841	0.033
Biološki skup podataka	**Strogo poklapanje**		**Relaksirano poklapanje**	
	F_1	σ	F_1	σ
ljudi	0.817	0.032	0.855	0.028
LogReg	**0.787**	0.080	**0.824**	0.039
MLP	0.723	0.062	0.793	0.064
RandForest	0.776	0.057	0.796	0.078
SVM	0.784	0.080	0.795	0.047

iako je korelacija sa semantičkom sličnosti nešto lošija nego kod prethodne tri mjere.

Konačni skup značajki koji opisuje par tema gradi se primjenom ove četiri mjere udaljenosti na parove normaliziranih vektora riječi i dokumenata tema. Svaki par tema na reprezentiran je s osam značajki – četiri značajke udaljenosti vektora riječi i četiri značajke udaljenosti vektora dokumenata.

Odabir modela Nakon što su odabrani klasifikacijski modeli i njihovi hiperparametri (sažeti u tablici 5.4), te definirane značajke parova tema, provodi se postupak odabira najboljih klasifikacijskih modela, odnosno modela koji najbolje aproksimiraju ljudske oznake jednakosti parova tema. Postupak odabira modela provodi se za obje varijante problema klasifikacije, strogu i relaksiranu, opisane na početku odjeljka. Za svaki od modela provodi se, na cijelom skupu označenih parova, procjena performansi modela dobivenih optimizacijom vrijednosti hiperparametara. Ovaj postupak provodi se ugniježđenom unakrsnom provjerom s pet preklopa (engl. *nested five fold crossvalidation*), pri čemu se kao kriterij kvalitete koristi F_1 mjera. Za oba skupa podataka, performanse svih modela na oba klasifikacijska problema prikazane su u tablici 5.5.

Kod obične unakrsne provjere sa K preklopa, svaka kombinacija hiperaparametara se ocjenjuje učenjem modela na svim odabirima $K-1$ preklopa i vrednovanjem na preostalom prek-

lopu. Kod ugnježđene verzije se za svaku "vanjsku" kombinaciju od $K-1$ preklopa provodi puna optimizacija hiperaparametara unutarnjom unakrsnom provjerom, nakon čega se na vanjskom preklopu vrednuje optimirani model. Drugim riječima, ovakav pristup koristi unakrsnu provjeru za procjenu performansi cijelog postupka optimizacije hiperparametara, a ne samo jedne kombinacije hiperparametara. Ugniježđena unakrsna provjera, iako računski zahtjevnija, dovodi do bolje procjene performansi modela dobivenih optimiranjem hiperparametara [163]. Preklopi su formirani stratificiranim uzorkovanjem (engl. *stratified sampling*) kako bi se osigurala zastupljenost klasa reprezentativna za cijeli skup podataka.

Postupak odabira i izgradnje modela koristi F_1 mjeru za ocjenu klasifikacijskih performansi. Konačni cilj postupka je naći najbolji model za ocjenjivanje pokrivenosti skupa referentnih tema od strane tema modela, koja se računa na temelju odluke klasifikacijskog modela o poklapanju parova referentnih tema i tema modela. Optimizacija F_1 mjere dovodi do minimizacije broja grešaka klasifikacije [164], što je očito poželjno svojstvo za dobivanje što preciznijih ocjena pokrivenosti. Dodatno, F_1 mjera uprosječuje odnosno balansira mjere klasifikacijske preciznosti i odziva, svojstava poželjnih za dobru procjenu pokrivenosti. Visoka preciznost je poželjna kako ne bi došlo do krivo procjenjenih poklapanja koja bi dovela do izmjerene pokrivenosti koja je veća od stvarne. Visoki odziv je poželjan kako bi se sva poklapanja s temama modela detektirala i kako se nebi izmjerila pokrivenost niža od stvarne.

Tablica 5.5 sadrži procjene performansi optimiranih modela za svaki od klasifikacijskih problema i skupova podataka. Za svaki model prikazani su prosjek i standardna devijacija F_1 mjere na pet vanjskih preklopa ugniježđene unakrsne provjere. Za svaki vanjski preklop, F_1 mjeri performanse modela dobivenog optimiranjem hiperparametara u unutarnjoj petlji unakrsne provjere. Performanse modela uspoređene su s međusobnim slaganjem ljudskih označivača mjerenim F_1 mjerom. To slaganje odgovara F_1 ocjeni predviđanja oznaka klasa, dobivenih uprosječivanjem ocjena dvaju označivača, od strane trećeg označivača. Konačna ocjena slaganja uprosječena je za sva tri moguća odabira označivača koji predviđa klasu. Oznake klasa i predviđanja označivača dobiveni su primjenom istog praga sličnosti koji je korišten u odgovarajućoj varijanti klasifikacijskog problema.

Iz tablice 5.5 se vidi da `LogReg` model, iako najjednostavniji od sva četiri razmotrena modela, daje najbolje klasifikacijske rezultate. Model logističke regresije daje najbolje performanse za tri od četiri instance klasifikacijskog problema, dok je na preostalom problemu drugi najbolji model. Ostali modeli, iako se na nekim instancama problema približavaju `LogReg` modelu, daju loše rezultate na barem jednoj instanci. Opisana robustnost `LogReg` modela s obzirom na varijantu problema dodatni je argument za odabir tog modela. Stoga je `LogReg` odabran kao najbolji model i u daljnjim eksperimentima prokrivenosti se koristi za računanje poklapanja tema na oba skupa podataka i za obje vrste poklapanja, strogo i relaksirano. Konačni modeli za mjere pokrivenosti dobivaju se, za svaku instancu problema, tako da se prvo optimiraju hiperpa-

rametri (unakrsnom provjerom s pet preklopa), a zatim se nauči model s najboljim parametrima. Optimiranje hiperparametara i učenje izvode se na cijelom skupu označenih parova.

Rezultati pokazuju da performanse klasifikacijskih modela ne zaostaju mnogo s međusobnim slaganjem označivača. Iz rezultata se može vidjeti i da, za različite instance problema, performanse najboljih modela dobro koreliraju s međusobnim slaganjem označivača. Rezultati također pokazuju da je procjena jednakosti tema teži problem za biološke nego za medijske teme, što je u skladu s opažanjima i statistikama iz postupka označavanje tema (opisanog u 5.6.2). Može se vidjeti i da su relaksirane varijante klasifikacijskog problema lakše i za ljude i za modele – označivači postižu veće slaganje a modeli veću točnost nego za stroge varijante problema. Ovo pokazuje da je lakše razdvojiti slučajeve aproksimativnog poklapanja od nepoklapanja nego razdvojiti slučajeve dobrog poklapanja od aproksimativnog poklapanja i nepoklapanja.

Iz rezultata se može zaključiti da opisani postupak odabira modela dovodi do kvalitetnih modela. Preciznije, ljudske ocjene poklapanja tema, definiranog kao jednakost interpretiranih koncepata, mogu se jako dobro naučiti na temelju značajki udaljenosti vektora riječi i dokumenata tema računatih manjim brojem pogodnih mjera udaljenosti. Posljedično, mjere pokrivenosti tema temeljene na nadziranim modelima očekivano daju dobre odnosno točne procjene pokrivenosti.

5.7 Krivulja pokrivenost-udaljenost

Mjerenje pokrivenosti temeljeno na nadziranom modelu opisano u prethodnom odjeljku aproksimira ljudsko ocjenjivanje pokrivenosti temeljeno na procjeni poklapanja referentnih tema s temama modela, pri čemu se poklapanje tema definira kao jednakost interpretiranih koncepata. Vrednovanje nadziranih modela poklapanja pokazuje da se ljudske ocjene poklapanja mogu jako dobro aproksimirati, iz čega slijedi da su na modelima temeljene ocjene pokrivenosti očekivano dobra aproksimacija ljudskih ocjena. Međutim, postupak izrade nadziranih modela vremenski je zahtjevan zbog potrebe za označavanjem parova tema. Postupak označavanja uključuje treniranje više označivača koji su stručni za interpretaciju tema i tekstova, te provedbu samog označavanja. Nije praktično provoditi označavanje za svaku novu tekstnu domenu za koju je od interesa mjerenje pokrivenosti.

Iz tog razloga potrebna je nenadzirana automatska mjera pokrivenosti koja dobro korelira s nadziranim mjerama a može se brzo primijeniti na nove skupove podataka. Ovdje se predlaže takva mjera pokrivenosti tema koja, kao i nadzirane mjere, mjeri pokrivenost skupa referentnih tema temeljem poklapanja referentnih tema s temama modela. Pri tome se poklapanje računa na temelju mjere udaljenosti i praga udaljenosti koji predstavlja kriterij poklapanja tema. Sama pokrivenost se računa variranjem praga udaljenosti i integriranjem pokrivenosti dobivenih za

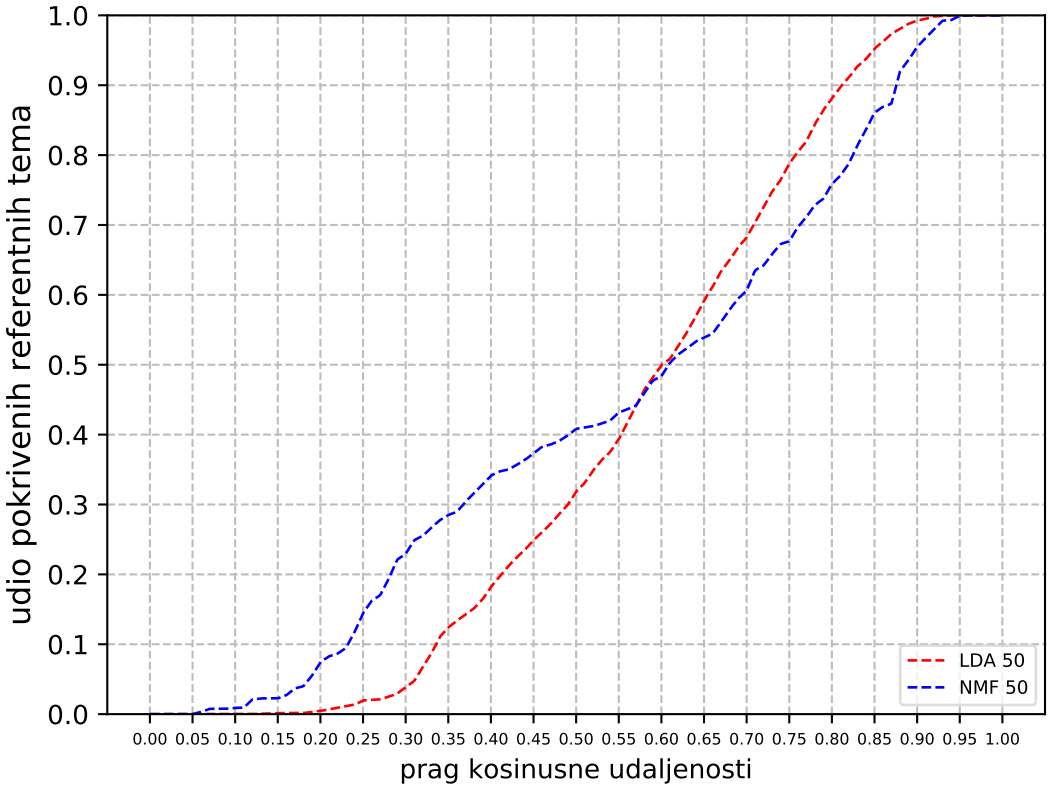
različite vrijednosti praga. Preciznije, variranjem praga udaljenosti formira se krivulja s pragom udaljenosti na x-osi i rezultirajućom pokrivenosti na y-osi, a mjera pokrivenosti računa se kao površina ispod te krivulje. Pri tome je pokrivenost za određeni prag definirana kao udio referentnih tema za koje postoji barem jedna tema modela čija je udaljenost od referentne manja od praga.

Ovaj postupak osim konačne mjere pokrivenosti vodi i do same krivulje koja pruža dodatne informacije o pokrivenosti. Preciznije, ova krivulja pruža sažet grafički prikaz ovisnosti pokrivenosti o strogoći kriterija poklapanja tema i omogućava grafičku usporedbu pokrivenosti većeg broja modela. Primjerice, promatranjem krivulje moguće je brzo dobiti informaciju o razini preciznosti poklapanja potrebnoj da pojedini modeli postignu određenu pokrivenost. Analogni primjer u kontekstu mjerenja klasifikacijskih performansi je ROC-krivulja iz koje se može vidjeti ponašanje niza varijanti klasifikacijskog modela. ROC-krivulja razlikuje se od mjera poput točnosti i F_1 mjere koje daju jednu numeričku ocjenu kvalitete modela. Međutim, ROC-krivulja i krivulja pokrivenost-udaljenost znatno se razlikuju pošto ROC-krivulja nastaje variranjem parametara modela dok krivulja pokrivenost-udaljenost nastaje variranjem kriterija za ocjenu modela.

5.7.1 Definicija i izgradnja krivulje

Krivulja pokrivenost-udaljenost, nadalje *PU-krivulja*, temelji se na *mjeri udaljenosti tema d* odnosno preslikavanju $d : T \times T \to \mathbb{R}$, pri čemu je T skup reprezentacija tema. PU-krivulja temeljena na mjeri d koja prikazuje pokrivenost skupa referentnih tema od strane tema modela definirana je na sljedeći način. Za referentni skup tema $T_{ref} = \{t_r\}$, skup tema modela $T_M = \{t_r\}$, i mjeru udaljenosti d koja poprima vrijednosti unutar intervala $[0, Max_d]$, PU-krivulja se sastoji od svih točaka (x, y) takvih da je $x \in [0, Max_d]$, a $y = |\{t_r \in T_{ref} | \exists t \in T_M : d(t_r, t) \leq x\}|/|T_{ref}|$. Drugim riječima, x-koordinate točaka krivulje su vrijednosti praga mjere udaljenosti, a y-koordinate su definirane kao pokrivenost dobivena za odgovarajući prag – udio onih referentnih tema za koje postoje pokrivajuće teme modela čija je udaljenost od referentne teme manja od praga. Jednostavnije rečeno, PU-krivulja je graf funkcije preslikavanja praga udaljenosti u odgovarajuću pokrivenost. Treba napomenuti da iako sama mjera udaljenosti ne mora biti ograničena odnosno može poprimati proizvoljno velike vrijednosti, za potrebe računanja PU-krivulje potrebno je odrediti maksimalnu udaljenost. To je uvijek moguće napraviti empirijskim računanjem udaljenosti na uzorku parova ili računanjem maksimalne udaljenosti parova tema iz pogodno odabrane reprezentativne klase tema.

Dok PU-krivulja pruža informacije o pokrivenosti referentnih tema s obzirom na različite definicije jednakosti tema, *površina ispod PU-krivulje*, nadalje označena sa *PPU*, daje numeričku mjeru pokrivenosti. PPU mjera može se koristiti za kvantitativno vrednovanje i odabir tematskih modela. Za računanje površine ispod PU-krivulje može se koristiti neka od metoda



Slika 5.1: PU-krivulja pokrivenosti referentih tema od strane modela LDA i NMF sa 50 tema, za zbirku američkih medijskih tekstova.

numeričke integracije.

PPU mjera i PU-krivulja određene su mjerom udaljenosti tema i mogu se automatski izračunati odnosno izgraditi za proizvoljni skup referentnih tema i tema modela. Slika 5.1 prikazuje PU-krivulje za modele LDA i NMF sa 50 tema. Odgovarajuće vrijednosti površine PPU su 0.410 za LDA i 0.407 za NMF. Iako su površine vrlo slične, promatranjem krivulja može se vidjeti razlika u prirodi pokrivenosti – NMF ima bolju pokrivenost za manje pragove udaljenosti, dok je pokrivenost modela LDA bolja za veće pragove. Drugim riječima, NMF daje bolju pokrivenost uz precizno poklapanje tema, dok LDA daje bolju pokrivenost uz aproksimativno poklapanje odnosno pokrivenost kod koje referentna tema može biti "otkrivena" semantički srodnom iako različitom temom modela.

U daljnjim eksperimentima, *udaljenost tema* definirana je kao *udaljenost vektora riječi tema*. Prilikom računanja površine ispod PU-krivulje, izgradnja krivulje provedena je segmentiranjem ukupnog raspona mjere udaljenosti na 50 podintervala jednake širine i računanjem pokrivenosti u rubnim točkama intervala. Površina ispod krivulje aproksimira se primjenom pravila trapeza.

5.7.2 Poželjna svojstva mjere udaljenosti

Mjera površine ispod PU-krivulje temelji se na mjeri udaljenosti između tema koja predstavlja kriterij za određivanje poklapanja dviju tema. Tema, bilo da se radi o temi modela ili referentnoj temi, predstavljena je vektorom dokumenata i vektorom riječi. Koordinate vektora riječi odgovaraju riječima u riječniku a vrijednosti težinama riječi za temu. Koordinate vektora dokumenata odgovaraju dokumentima u zbirci a vrijednosti težinama dokumenata za temu. U svim daljnim eksperimentima, mjera udaljenosti dviju tema računa se primjenom neke od mjera udaljenosti vektora na vektore riječi tema. Ovaj pristup je smislen jer mjera udaljenosti služi kao kriterij semantičkog poklapanja tema a eksperimenti s tematskim modelima pokazali su da udaljenost vektora riječi može dobro aproksimirati semantičku sličnost tema [29, 64, 65, 141]. Dodatno, vektori riječi su glavni način reprezentacije tema, što znači da je opisani pristup lakše primijeniti pošto implementacije tematskih modela očekivano omogućuju brz pristup vektoru riječi, dok je vektore dokumenata često potrebno izračunati. Međutim, udaljenost tema PU-krivulje mogla bi se računati i na temelju dokumenata te mnogih drugih reprezentacija tema dobivenih na temelju informacija sadržanih u modelu. Neovisno o definiciji, postoji nekoliko kriterija koje bi mjera udaljenosti trebala zadovoljiti kako bi bila pogodna za izgradnju PU-krivulje.

Prvo, mjera udaljenosti bi trebala imati dobru *korelaciju s ljudskom percepcijom poklapanja tema*, odnosno manje udaljenosti tema trebale bi odgovarati većoj vjerojatnosti poklapanja, i obrnuto. Za takve mjere variranje praga udaljenosti koji se koristi za definiciju poklapanja tema zaista odgovara variranju vjerojatnosti semantičkog poklapanja.

Drugo, mjera udaljenosti bi trebala biti maksimalno *neovisna o klasi tematskog modela*. Ovo znači da bi odnos između udaljenosti i vjerojatnosti poklapanja tema trebao biti zadržan neovisno da li je tema naučena nekim vjerojatnosnim modelom, faktorizacijskim modelom ili je od ljudi konstruirana referentna tema. Ovo svojstvo je nužno kako bi bilo moguće uspoređivati referentne teme s temama raznih modela, te kako bi mjerenja pokrivenosti različitih modela bila usporediva.

Treće, mjera udaljenosti trebala bi pokrivati *predvidljivi raspon vrijednosti* za široku klasu modela. Ovo svojstvo je poželjno zbog načina izgradnje PU-krivulje – variranjem praga udaljenosti od minimalne do maksimalne udaljenosti. Kako bi pokrivenosti raznovrsnih modela bile usporedive, njihove krivulje se moraju izgraditi na istom rasponu udaljenosti. Iako je uvijek moguće izračunati maksimalnu udaljenost tema na velikom uzorku modela od interesa, mnogo je praktičnije ako je mjera udaljenosti ograničena, ili ako se maksimalna udaljenost može lako izračunati za veću klasu modela.

Prvo svojstvo korelacije mjere s ljudskom intuicijom i drugo svojstvo neovisnosti o klasi modela su povezani. Naime, ljudi pri interpretaciji tema uzimaju u obzir relativne odnose težina riječi i dokumenata a ne njihove apsolutne iznose, pošto interpretiraju teme na temelju uređenih listi riječi i dokumenata. Stoga je očekivano da mjere koje dobro koreliraju s ljudskom percepcijom sličnosti tema koriste podatke o temama na sličan način. S druge strane, opisani način usporedbe tema čini se nužnim za neovisnost mjere o tipu modela. Naime, postoji velik broj tematskih modela s različitim definicijama vrijednosti u vektorima riječi i dokumenata tema te posljedično različitim skalama tih vrijednosti. Primjerice, kod vjerojatnosnih modela te vrijednosti mogu predstavljati vjerojatnosti ili neomeđene parametre apriornih distribucija, a kod faktorizacijskih modela vrijednosti su neomeđene i ovisne o metodi učenja modela.

Svojstvo korelacije mjere udaljenosti s ljudskom intuicijom može se provjeriti promatranjem korelacije mjere i ljudskih ocjena poklapanja tema iz odjeljka 5.6.2. Svojstvo neovisnosti mjere o klasi modela može se provjeriti promatranjem razdiobe udaljenosti velikog broja parova tema uzorkovanih iz različitih klasa modela. U nastavku se predlažu tri mjere udaljenosti koje zadovoljavaju opisana poželjna svojstva i daju se primjeri dviju mjera koje ih ne zadovoljavaju.

5.7.3 Dobre mjere udaljenosti

Sve mjere udaljenosti tema razmotrene u ovom i sljedećem odjeljku računaju udaljenost tema primjenom neke *mjere udaljenosti vektora* na vektore riječi tih tema. Tri mjere udaljenosti tema za koje se pokazalo da zadovoljavaju svojstva poželjna za izradu PU-krivulje koriste jednu od sljedećih mjera za računanje udaljenosti vektora tema: kosinusnu udaljenost, Hellingerovu udaljenost te normaliziranu L_1-udaljenost.

Kosinusna udaljenost (engl. *cosine distance*) je mjera udaljenosti često korištena u dubinskoj analizi podataka za usporedbu visokodimenzionalnih vektora [157]. Za dva vektora v i w

kosinusna udaljenost *cosd* definirana je kao:

$$cosd(v,w) = 1 - \frac{v \cdot w}{\|v\|\|w\|}$$

Sa $v \cdot w$ označen je skalarni produkt, a sa $\|v\|$ euklidska norma vektora. Kosinusna udaljenost je inverz kosinusne sličnosti vektora definirane kao $\frac{v \cdot w}{\|v\|\|w\|}$ i ovisi samo o kutu između dva vektora. To ovu udaljenost čini manje ovisnom o vrijednostima koordinata vektora, pošto ovisi samo o relativnim odnosima vrijednosti koordinata koji definiraju kut. Eksperimenti s tematskim modelima pokazuju da *cosd* dobro korelira s ljudskom percepcijom sličnosti tema [29]. Dodatno, za nenegativne vektore poput vektora tema vjerojatnosnih modela i modela NMF, kosinusna udaljenost poprima vrijednosti u rasponu $[0,1]$ – maksimalna vrijednost 1 odgovara pravom kutu između vektora, a minimalna vrijednost 0 odgovara vektorima na istom pravcu. Ovaj raspon vrijednosti neovisan je o dimenzionalnosti i ostalim svojstvima vektora.

Normalizirana L_1-udaljenost je standardna L_1-udaljenost (engl. *Manhattan distance*) primjenjena na vektore koji su prethodno normalizirani do vjerojatnosne razdiobe. L_1 udaljenost korištena je za detekciju međusobno sličnih tema modela [65]. Za dva vektora, v i w, L_1 udaljenost definirana je na sljedeći način:

$$L_1(v,w) = \sum_i |v_i - w_i|$$

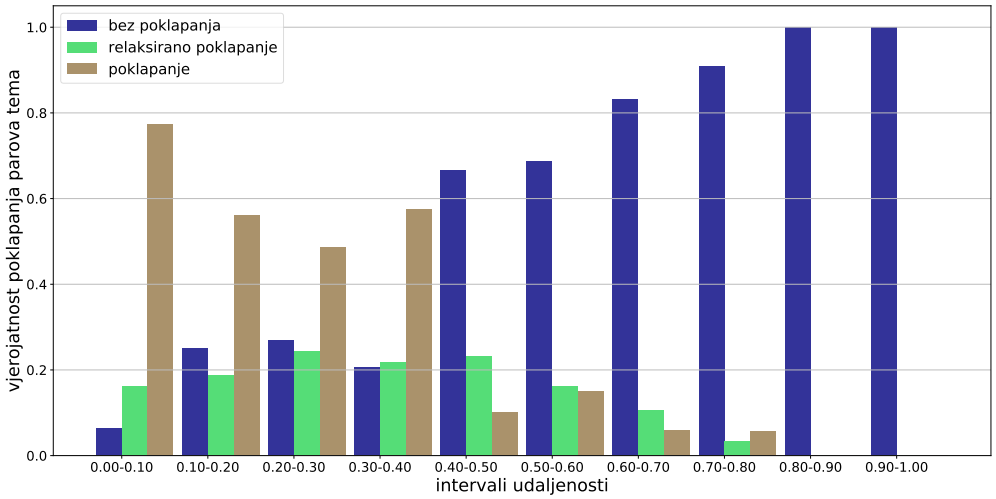
Normalizacija vektora se provodi iz dva razloga – kako bi se mjera učinila manje osjetljivom na vrijednosti komponenti vektora (što smanjuje ovisnost mjere o klasi modela) te kako bi se ograničila maksimalna vrijednost udaljenosti, koja je u slučaju normaliziranih vektora uvijek manja od 2, što se može lako pokazati pomoću nejedankost trokuta.

Hellingerova udaljenost, u oznaci helld, je mjera udaljenosti dviju vjerojatnosnih razdioba i predstavlja simetričnu aproksimaciju KL-divergencije [162]. Prije primjene helld mjere vektori se normaliziraju do vjerojatnosne razdiobe kako bi mjera bila primjenjiva i za faktorizacijske modele i ostale modele koji nisu vjerojatnosni. Za dva vektora v i w, helld udaljenost definirana je kao:

$$helld(v,w) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{v_i} - \sqrt{w_i})^2} = \frac{1}{\sqrt{2}}\|\sqrt{v_i} - \sqrt{w_i}\|$$

Hellingerova udaljenost je zbog normalizacije vektora očekivano manje osjetljiva na klasu modela odnosno na razlike u rasponima vrijednosti koordinata vektora. Može se pokazati da su vrijednosti helld mjere ograničene na interval $[0,1]$.

Provjera koreliranosti opisanih mjera s ljudskim ocjenama jednakost tema provodi se korištenjem označenih parova tema opisanih u odjeljku 5.6.2. Za svaki skup podataka, medijski i biološki, označeni su parovi tema koje se sastoje od referentnih tema (opisanih u odjeljku 5.4) i tema modela iz četiri različite klase (opisanih u odjeljku 5.5). Parovi tema označeni su oz-



Slika 5.2: Korelacija kosinusne udaljenosti i ljudskih oznaka poklapanja tema, za medijske teme.

nakama 1 (poklapanje – isti koncepti), 0.5 (poklapanje uz šum ili vrlo slične koncepte), ili 0 (različiti koncepti).

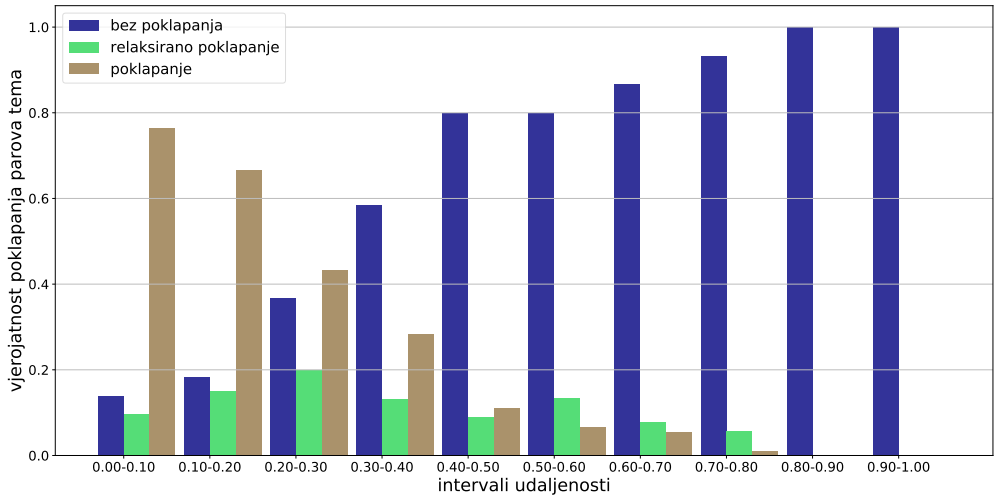
Za svaku od mjera udaljenosti i svaku od zbirki tekstova, izrađen je graf koji prikazuje vjerojatnosti ova tri tipa poklapanja u ovisnosti o udaljenosti. Preciznije, za svaki od intervala udaljenosti mjere (prikazanih na x-osi) izračunate su vjerojatnosti oznaka svih parova tema čije se međusobne udaljenosti nalaze unutar intervala. Intervali su formirani podjelom cijelog raspona vrijednosti mjere na 10 jednakih djelova. Pri tome su za kosinusnu i Hellingerovu udaljenost uzeti puni rasponi vrijednosti (interval $[0,1]$) dok je za mjeru L_1 doljnja granica raspona povećana, radi veće jasnoće prikaza, do minimalne vrijednosti na svim parovima.

Opisani grafovi prikazani su na slikama 5.2 – 5.7. Grafovi jasno pokazuju visoku razinu korelacije mjera udaljenosti i ljudskih ocjena poklapanja za oba skupa parova tema. Manja udaljenost tema korelira s povećanim udjelom poklapanja tema i obrnuto, povećanje udaljenost korelira s povećanjem udjela različitih tema. Pri tome se za svaku od mjera ovaj prijelaz udjela odvija kontinuirano.

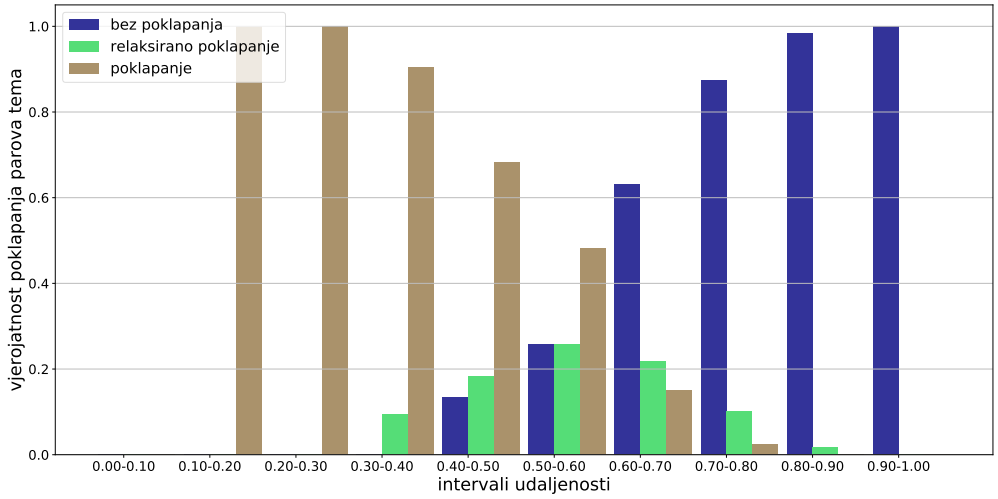
Kako se provjerila neosjetljivost opisanih mjera udaljenosti na klasu tematskih modela, izrađeni su grafovi razdiobe udaljenosti tema uzorkovanih iz modela različitih klasa. Preciznije, cilj je provjeriti neosjetljivost izmjerene udaljenosti tema na razlike u vektorima riječi koje su posljedica klase modela a ne sličnosti ili različitosti tema.

Iz skupa svih parova tema svih modela četiri različite klase (opisanih u 5.5 i tablici 5.2) uzorkovan je skup od 20.000 parova i izrađeni su histogrami razdiobe vrijednosti mjera udaljenosti. Histogrami su normalizirani tako da na y-osi ne prikazuju broj parova već udio u ukupnom broju parova. Ovi histogrami prikazani su na slikama 5.8 – 5.13.

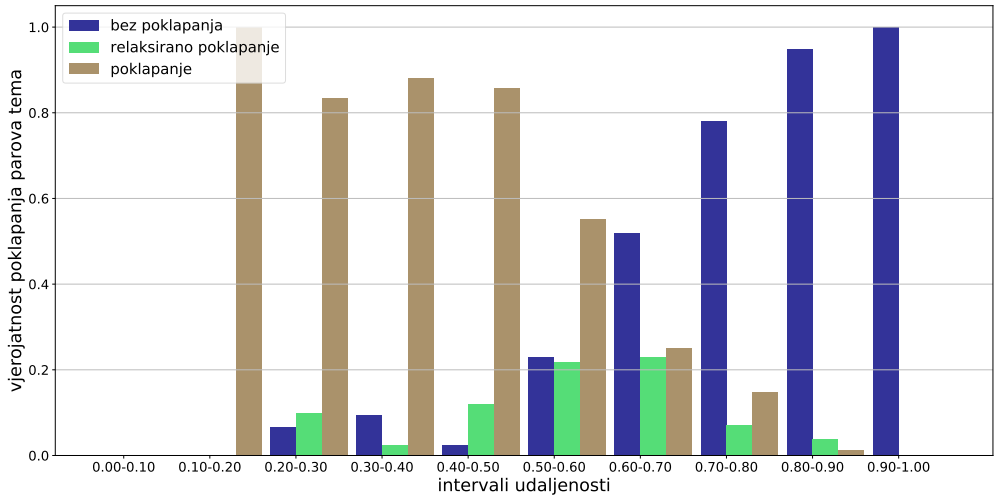
Iz slika se vidi da su razdiobe udaljenosti za sve mjere i za oba skupa podataka unimo-



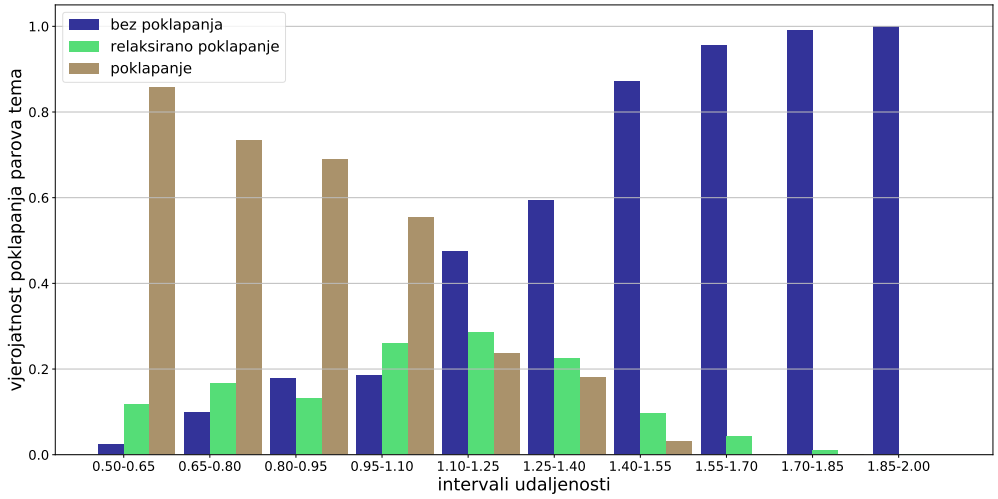
Slika 5.3: Korelacija kosinusne udaljenosti i ljudskih oznaka poklapanja tema, za biološke teme.



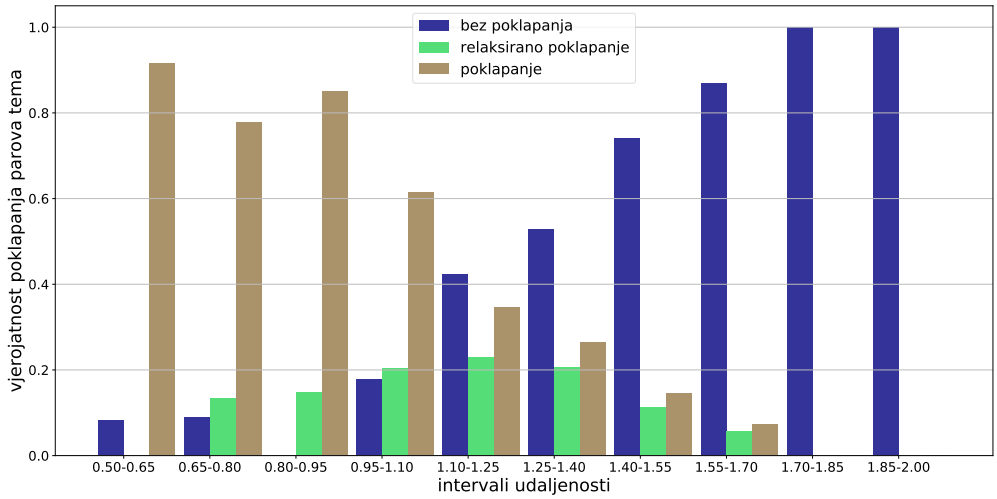
Slika 5.4: Korelacija Hellingerove udaljenosti i ljudskih oznaka poklapanja tema, za medijske teme.



Slika 5.5: Korelacija Hellingerove udaljenosti i ljudskih oznaka poklapanja tema, za biološke teme.



Slika 5.6: Korelacija L_1-udaljenosti i ljudskih oznaka poklapanja tema, za medijske teme.



Slika 5.7: Korelacija L_1-udaljenosti i ljudskih oznaka poklapanja tema, za biološke teme.

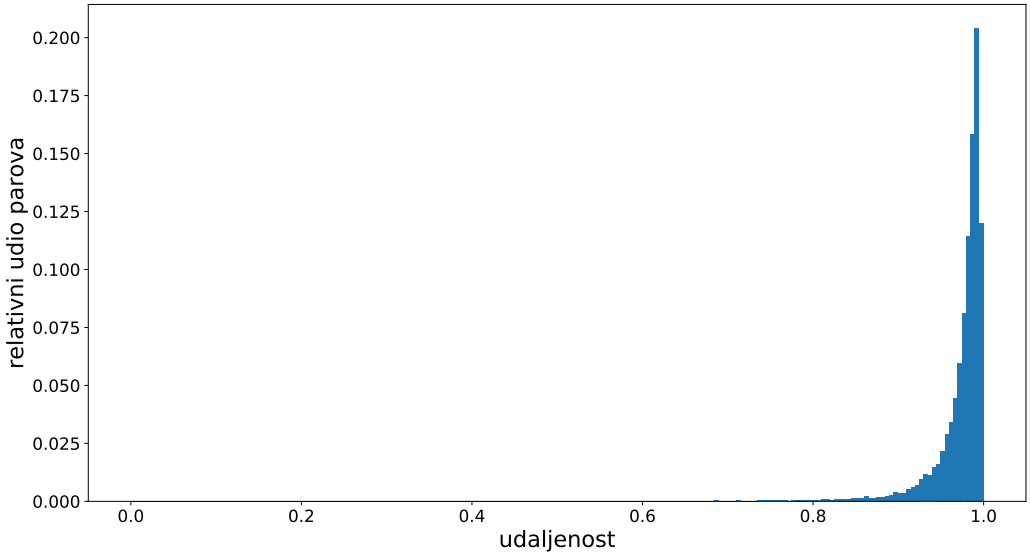
dalne, što ne bi bio slučaj da postoje značajne razlike u razdiobama udaljenosti tema različitih klasa modela. Kada se promotre razdiobe udaljenosti tema unutar svake pojedine klasa modela, rezultati su u skladu s unimodalnošću globalne razdiobe. Preciznije, svaka mjera daje slične razdiobe za sve klase modela i za oba skupa podataka, uz manje razlike koje se mogu objasniti prirodom modela i skupova podataka. Primjerice, za manju biološku zbirku tekstova koja očekivano sadrži manji broj tema dolazi do pojave većeg broja jednakih i sličnih tema, što rezultira udaljenostima nešto manjim od udaljenosti medijskih tema.

Histogrami također pokazuju da su udaljenosti sve tri mjere grupirane oko visokih vrijednosti, odnosno da je izmjerena udaljenost većine parova tema visoka. Ovo svojstvo potvrđuje usklađenost mjera s prirodom poklapanja tema – kao što je opisano u odjeljku 5.6.1, za slučajno uzorkovane parove tema poklapanje tema se očekuje samo za mali udio parova, približno $1/T$ parova za modele s T tema. Pri tome kosinusna udaljenost najbolje prati ovo svojstvo pošto broj parova kontinuirano raste s porastom udaljenosti. Za druge dvije mjere, iako su im razdiobe pomaknutih prema višim vrijednostima, udio parova pada na samom rubu spektra blizu maksimalne vrijednosti. To ukazuje da kosinusna udaljenost nešto bolje modelira poklapanje tema.

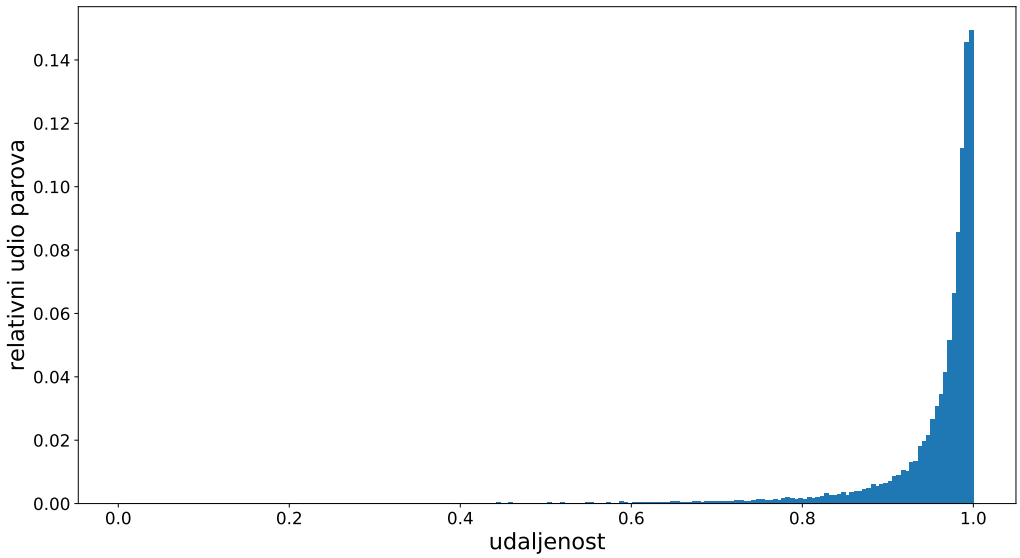
Svojstva koreliranosti s ljudskim ocjenama poklapanja i neosjetljivosti na klasu modela pokazuju da tri opisane mjere imaju sva svojstva potrebna za njihovu primjenu u izgradnji PU-krivulje.

5.7.4 Loše mjere udaljenosti

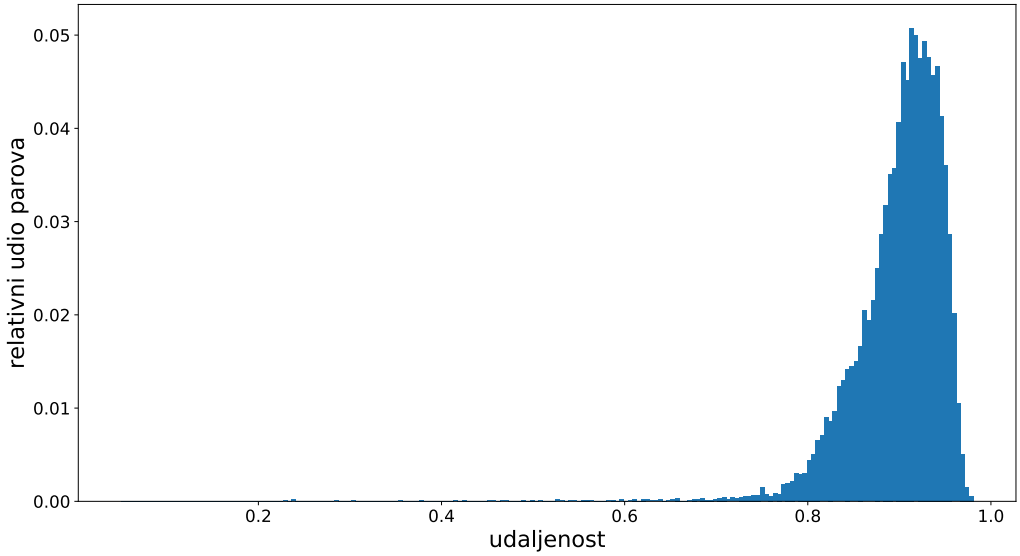
Ovdje se daju primjeri dviju mjera udaljenosti koje nisu dobar izbor za izradu PU-krivulje. Prva mjera je standardna nenormalizirana L_1 udaljenost a druga mjera je Jensen-Shannonova



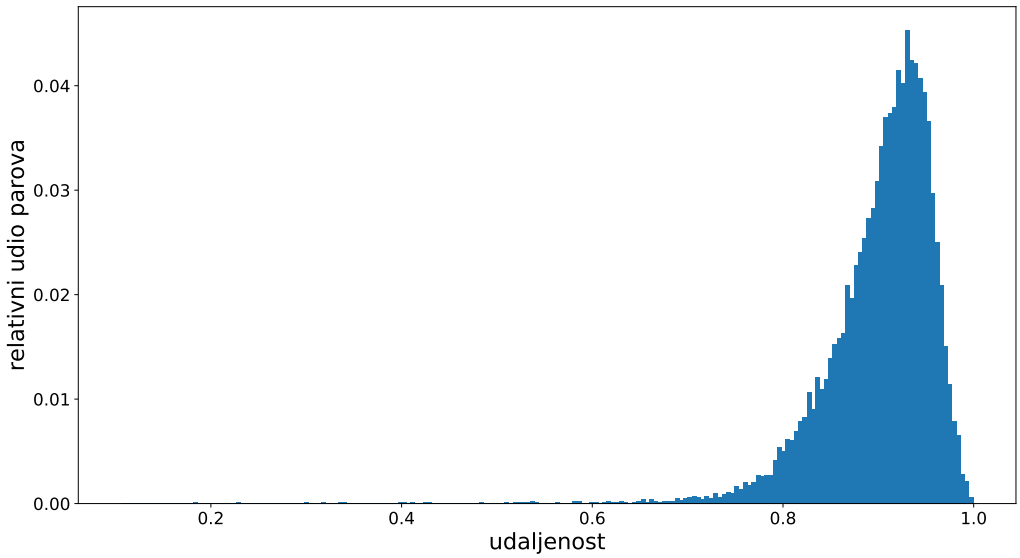
Slika 5.8: Razdioba kosinusne udaljenosti za parove medijskih tema.



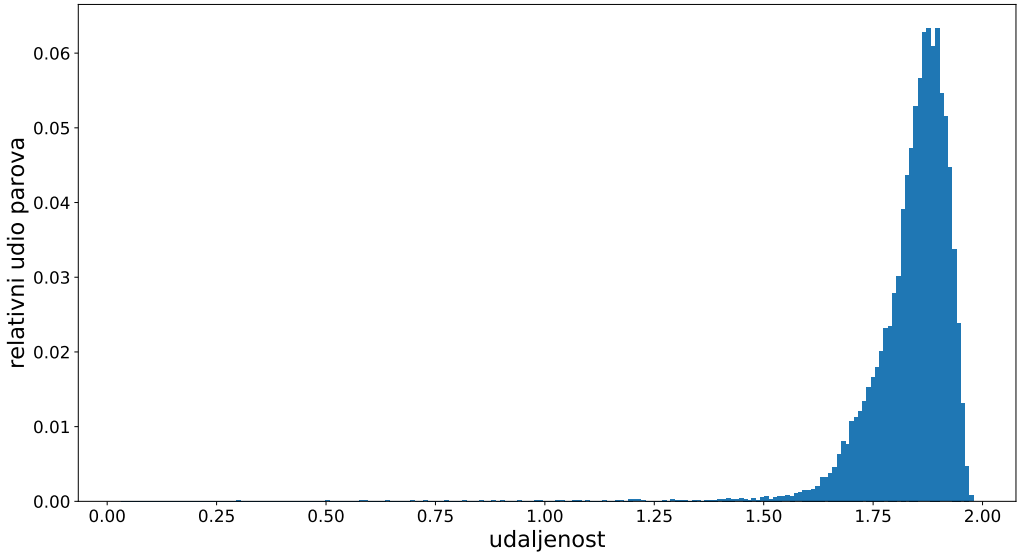
Slika 5.9: Razdioba kosinusne udaljenosti za parove bioloških tema.



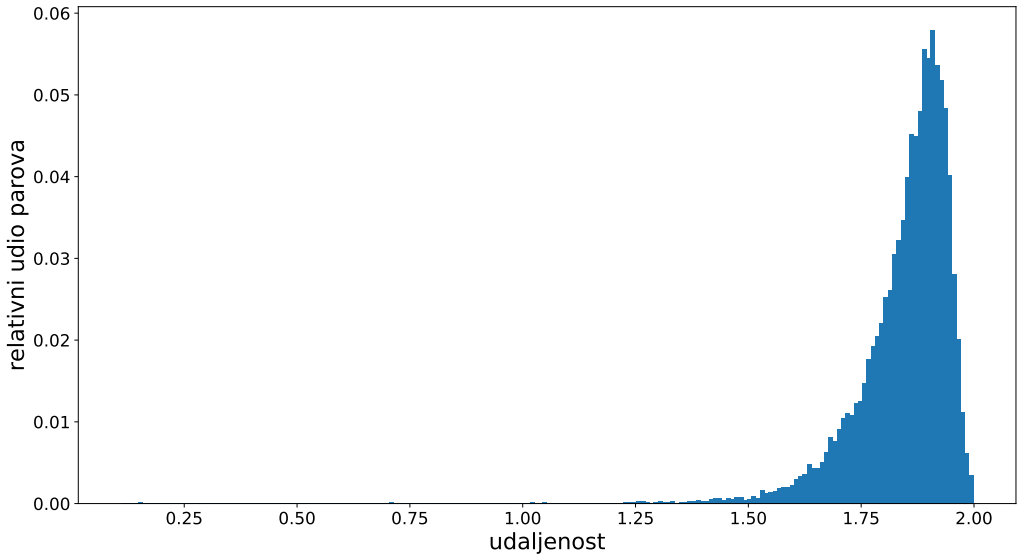
Slika 5.10: Razdioba Hellingerove udaljenosti za parove medijskih tema.



Slika 5.11: Razdioba Hellingerove udaljenosti za parove bioloških tema.



Slika 5.12: Razdioba L_1-udaljenosti za parove medijskih tema.



Slika 5.13: Razdioba L_1-udaljenosti za parove bioloških tema.

divergencija ili JS-divergencija (engl. *Jensen-Shannon divergence*), označena sa JS-div. JS-divergencija je simetrična verzija KL-divergencije i mjeri udaljenost vjerojatnosnih razdioba pa je prije njene primjene potrebno normalizirati vektore tema. JS-divergencija i KL-divergencija koriste za mjerenje semantičke sličnosti tema [141] i ocjenu kvalitete tema [30].

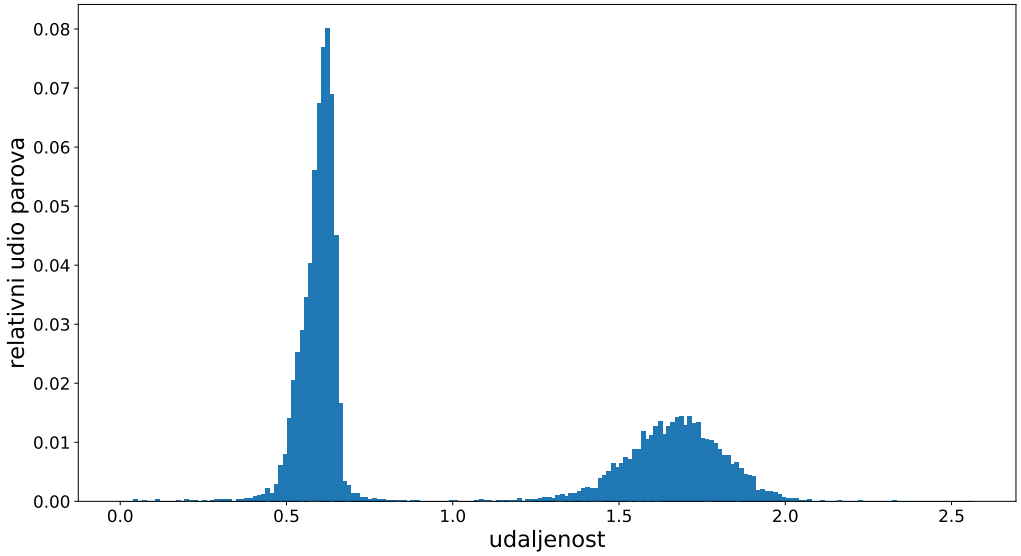
Mjere L_1 i JS-div prije svega nisu neovisne o klasi tematskih modela. Ovo se može vidjeti iz histograma koji prikazuju razdiobe udaljenosti za veliki uzorak parova tema raznorodnih modela (isti uzorak korišten za ispitivanje dobrih mjera udaljenosti). Histogrami su prikazani na slikama 5.14 i 5.15. Ovi histogrami prikazuju udaljenosti medijskih tema, no situacija je ista i za biološke teme.

Iz histograma se može vidjeti da su razdiobe udaljenosti ovih mjera bimodalne, odnosno da se udaljenosti grupiraju oko dviju različitih vrijednosti. To je posljedica osjetljivosti mjera na klasu modela, konkretno osjetljivosti na različitosti između tema vjerojatnosnih modela i tema faktorizacijskog modela NMF. U slučaju JS-divergencije manje vrijednosti udaljenosti (lijeva polovica histograma) odgovaraju parovima tema koje dolaze iz istih tipova modela, dok veće vrijednosti odgovaraju parovima sastavljenim od teme vjerojatnosnog modela i teme modela NMF. Pri tome parovi s obje teme iz modela NMF također odgovaraju manjim vrijednostima. To pokazuje da je JS-divergencija, za razliku od Hellingerove udaljenosti koja također mjeri udaljenosti vjerojatnosnih razdioba, osjetljiva na razlike u strukturi razdioba koje su posljedica vrste modela a ne samo semantičke razlike među temama.

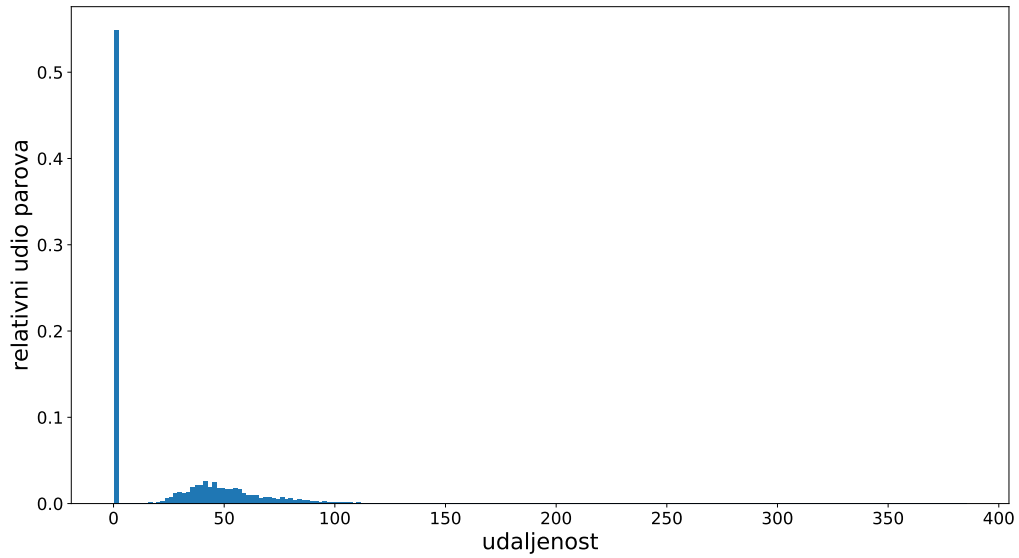
U slučaju L_1 udaljenosti, veće vrijednosti odgovaraju udaljenostima dviju tema iz modela NMF te udaljenostima između tema vjerojatnosnih modela s temama modela NMF. Manje vrijednosti odgovaraju udaljenostima između tema vjerojatnosnih modela. Razlog ove bimodalnosti je taj što vektori tema modela NMF (bez provedene normalizacije) nisu vjerojatnosne razdiobe i sadrže mnogo veće vrijednosti od vektora tema vjerojatnosnih modela.

Ovi nedostaci čine nenormaliziranu L_1 mjeru i JS-divergenciju nepogodnima izgradnju PU-krivulje. U slučaju JS-divergencije za parove NMF tema i tema vjerojatnosnih modela za koje je pregledom utvrđeno da se mogu semantički poklapati, mjera će dati visoke udaljenosti zbog strukturnih različitosti razdioba riječi u temama. Kod izgradnje PU-krivulje, prag udaljenosti koji odgovara različitosti tema neće biti jednak za sve modele već će biti manji za parove tema istih modela i veći za parove NMF tema i vjerojatnosnih tema. Ova pojava, kao i općenito slabija korelacija s ljudskim oznakama od prethodno opisanih dobrih mjera, može se vidjeti na slici 5.16.

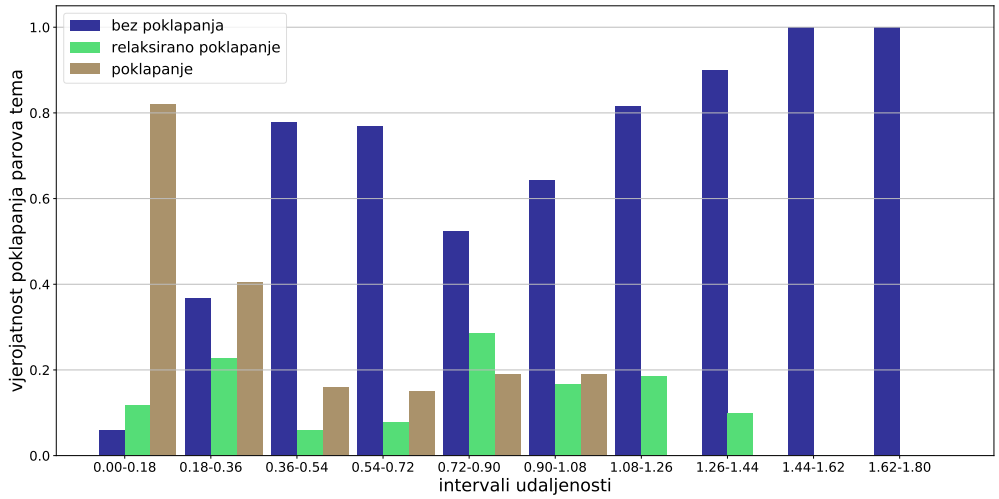
U slučaju nenormalizirane L_1 udaljenosti dodatno vrijedi i to da mjera za različite tipove modela poprima vrijednosti različitih redova veličine, što čini pripadnu PU-krivulju nepogodnom za analizu pokrivenosti pošto je utjecaj udaljenosti manjeg reda veličine (koje odgovaraju vjerojatnosnim modelima) na krivulju zanemariv.



Slika 5.14: Razdioba JS-divergencije za parove medijskih tema.



Slika 5.15: Razdioba nenormalizirane L_1-udaljenosti za parove medijskih tema.



Slika 5.16: Korelacija JS-divergencije i ljudskih oznaka poklapanja tema, za medijske teme.

5.7.5 Vrednovanje pokrivenost-udaljenost mjera

Cilj oblikovanja PPU mjere je pronalazak nenadzirane metode mjerenja pokrivenosti tema koja dobro aproksimira nadzirane mjere temeljene na ljudskim ocjenama poklapanja tema. Predložena PPU mjera računa površinu ispod PU-krivulje i temelji se na pogodnoj mjeri udaljenosti tema odnosno mjeri udaljenosti pripadnih vektora riječi. Definirana su poželjna svojstva ove mjere udaljenosti – korelacija s ljudskim ocjenama poklapanja, neovisnost o klasi modela i ograničeni raspon vrijednosti. Razmotrene su tri mjere udaljenosti – kosinusna udaljenost, Hellingerova udaljenost i normalizirana L_1 udaljenost, i za njih je pokazano da imaju opisana svojstva. Ovdje se ispituje korelacija tri varijante PPU mjere, temeljene na navedenim udaljenostima i označene sa PPU-cosd, PPU-helld, i PPU-L_1, s nadziranim mjerama pokrivenosti. Ispituje se korelacija PPU mjera s obje varijante nadzirane mjere pokrivenosti, mjerom NP-strog koja modelira strogo poklapanje tema i mjerom NP-relaks koja modelira relaksirano poklapanje tema. Obje ove varijante madzirane mjere temeljene su na modelu logističke regresije koji najbolje aproksimira ljudske ocjene poklapanja tema (odjeljak 5.6.3).

Korelacije se računaju na razini tematskih modela – za svaki od skupova podataka, mjere pokrivenosti skupa referentnih tema računaju se za svaki od 100 tematskih modela različitih klasa opisanih u odjeljku 5.5. Korelacije se mjere Spearmanovim koeficijentom korelacije (koji mjeri sličnost rangiranja tematskih modela izvedenih iz mjera pokrivenosti) i Pearsonovim koeficijentom korelacije (koji mjeri linearnu korelaciju mjera pokrivenosti). Ova dva koeficijenta korelacije mogu se koristiti za usporedbu mjera za vrednovanje modela strojnog učenja – u analizi provedenoj u [164] pomoću Spearmanovog i Pearsonovog koeficijenta analizira se srodnost mjera kvalitete klasifikacijskih modela.

Tablica 5.6 sadrži korelacije nadziranih mjera i PPU mjera na oba skupa podataka. Kore-

Tablica 5.6: Koeficijenti korelacije PPU mjera i nadziranih mjera pokrivenosti. Pripadne p-vrijednosti za sve korelacije manje su od 10^{-14}.

	PPU-cosd		PPU-helld		PPU-L_1	
	spearman	pearson	spearman	pearson	spearman	pearson
medijski skup podataka						
NP-strog	0.967	0.963	0.985	0.986	0.943	0.932
NP-relaks	0.872	0.908	0.909	0.925	0.962	0.959
biološki skup podataka						
NP-strog	0.946	0.953	0.759	0.691	0.791	0.692
NP-relaks	0.949	0.976	0.806	0.777	0.836	0.777

lacije između svih varijanti PPU mjera i nadziranih mjera visoke su prema oba kriterija. Pri tome se vrlo visoka razina korelacije s nadziranim mjerama (približno 0.9 ili više) postiže za sve PPU mjere na medijskom skupu podataka te slučaju PPU-cosd mjere na biološkom skupu podataka. Preostale dvije PPU mjere, PPU-helld i PPU-L_1, na biološkom skupu imaju nešto niže korelacije s nadziranim mjerama. Biološka zbirka karakterizirana je apstraktnijim temama koje je teže interpretirati i čije poklapanje je teže ocijeniti (odjeljak 5.6) od poklapanja medijskih tema. Iz tog razloga u slučaju bioloških tema može se očekivati viša razina pogreške kod procjene pokrivenosti, što je vjerojatan uzrok slabije korelacije mjera u ovom slučaju.

Jedino PPU-cosd mjera temeljena na kosinusnoj udaljenosti zadržava visoku razinu korelacije i u težem slučaju bioloških tema. Moguće objašnjenje ove pojave je da kosinusna udaljenost predstavlja bolji model semantičke udaljenosti tema od L_1 i Hellingerove udaljenosti. Međutim, moguće je da je visoka koreliranost PPU-cosd mjere i nadziranih mjera posljedica toga što oba tipa mjera ovise o kosinusnoj udaljenosti – nadzirane mjere koriste kosinusne udaljenosti kao značajke. . Zbog toga su izgrađeni nadzirani modeli koji ne koriste kosinusnu udaljenost za ekstrakciju značajki. Osim te razlike ovi modeli izgrađeni su na isti način kao i najbolji nadzirani modeli na temelju kojih je izračunata tablica 5.6. Nadzirane mjere pokrivenosti temeljene na modelima bez kosinusnih značajki i dalje zadržavaju visoku koreliranost sa PPU-cosd mjerom na oba skupa podataka, kao što je vidljivo iz tablice 5.7. Ovaj rezultat ide u prilog tvrdnji da je visoka korelacija nadziranih mjera i PPU-cosd mjere posljedica toga što kosinusna udaljenost dobro modelira semantičku udaljenost tema.

Na temelju svih prethodnih rezultata, mjera PPU-cosd odabire se kao varijanta PPU mjere najbolja za automatsku aproksimaciju nadziranih mjera pokrivenosti i koristi se u daljnjim eksperimentima vrednovanja modela.

Tablica 5.7: Koeficijenti korelacije PPU-cosd mjere i nadziranih mjera pokrivenosti koje ne koriste cosd mjeru za ekstrakciju značajki. Pripadne p-vrijednosti za sve korelacije manje su od 10^{-40}.

	PPU-cosd			
	medijski podaci		biološki podaci	
	spearman	pearson	spearman	pearson
NP-strog-nocos	0.961	0.928	0.946	0.964
NP-relaks-nocos	0.937	0.954	0.949	0.971

5.8 Vrednovanje pokrivenosti tematskih modela

U ovom poglavlju definiran je problem pokrivenosti tema i predložene su metode vrednovanja pokrivenosti koje uključuju izradu skupa referentnih tema i mjerenje pokrivenosti. Ovdje se analiziraju rezultati primjene tih metoda na mjerenje pokrivenosti referentnih tema od strane tematskih modela.

Struktura eksperimenta Provodi se vrednovanje tematskih modela opisanih u odjeljku 5.5. Cilj izgradnje tog skupa modela bio je obuhvatiti nekoliko standardnih i međusobno različitih klasa modela, te izgraditi varijante modela za različite konfiguracije broja tema, najvažnijeg hiperparametra. Tipovi modela obuhvaćaju seminalni i široko korišteni model LDA, njegovu ekstenziju aLDA, neparametarski model PYP koji sam uči broj tema, te popularni faktorizacijski model NMF. Za svaki skup podataka, medijski i biološki, izgrađeno je 100 različitih modela pri čemu je za svaku kombinaciju klase modela i broja tema naučeno 10 modela koji odgovaraju različitim slučajno odabranim parametrima procesa učenja.

Vrednovanje modela provodi se nadziranim mjerama pokrivenosti (odjeljak 5.6) i nenadziranom PPU mjerom (odjeljak 5.7). Nadzirane mjere pokrivenosti računaju pokrivenost referentnih tema na temelju nadziranog modela koji procjenjuje njihovo poklapanje s temama modela. Nadzirani modeli naučeni su na ljudskim oznakama semantičkog poklapanja tema a njihova izrada i vrednovanje opisani u odjeljku 5.6 pokazuju da je problem poklapanja dobro definiran i da izgrađeni modeli kvalitetno ocjenjuju poklapanje i ne zaostaju mnogo za međusobnim slaganjem označivača. Definirane su dvije vrste nadziranih mjera koje ocjenjuju pokrivenost ili na temelju strogog poklapanja tema ili na temelju relaksiranog poklapanja koje dopušta varijacije u temama zbog slučajnog šuma ili male semantičke različitosti. U ovdje provedenim mjerenjima koriste se nadzirani modeli logističke regresije koji daju najvišu kvalitetu aproksimacije ljudskih ocjena.

Nenadzirane mjere pokrivenosti oblikovane su s ciljem dobre aproksimacije nadziranih a temelje se na ocjenjivanju poklapanja tema pomoću mjere udaljenosti tema. U odjeljku 5.7

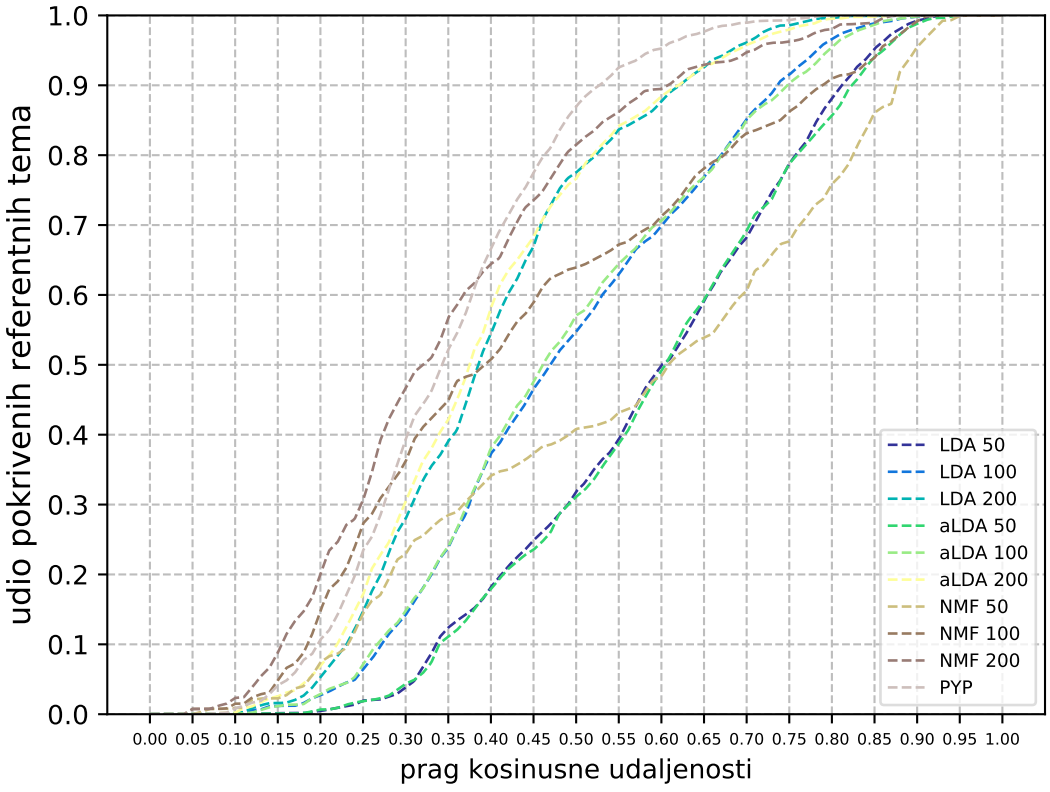
demonstrira se kvaliteta nekolicine mjera udaljenosti, a vrednovanje pokazuje da mjera PPU-cosd temeljena na kosinusnoj udaljenosti odlično aproksimira nadzirane mjere na oba skupa podataka. Nenadzirane mjere temelje se na izgradnji PU-krivulje, grafičkog alata za analizu pokrivenosti koji pokazuje način na koji pokrivenost referentnih tema ovisi o pragu udaljenosti koji određuje njihovu jednakost s temama modela.

Vrednovanje tematskih modela temelji se na referentnim temama koje definiraju koncepte čija pokrivenost se mjeri. U odjeljku 5.4 opisana su dva skupa referentnih tema dobivenih ljudskim pregledom, odabirom i doradom tema automatski naučenih modela – skup medijskih i bioloških tema. Oba skupa tema predstavljaju korisne rezultate eksplorativne analize tekstova, a način njihovog dobivanja jamči da se teme mogu pokriti korištenjem tematskih modela. Medijske teme dobivene su iz tema modela LDA i nisu filtrirane prema tipu koncepata, dok su biološke teme dobivene iz tema modela NMF i sastoje se samo od tema koje odgovaraju konceptima raznih fenotipova.

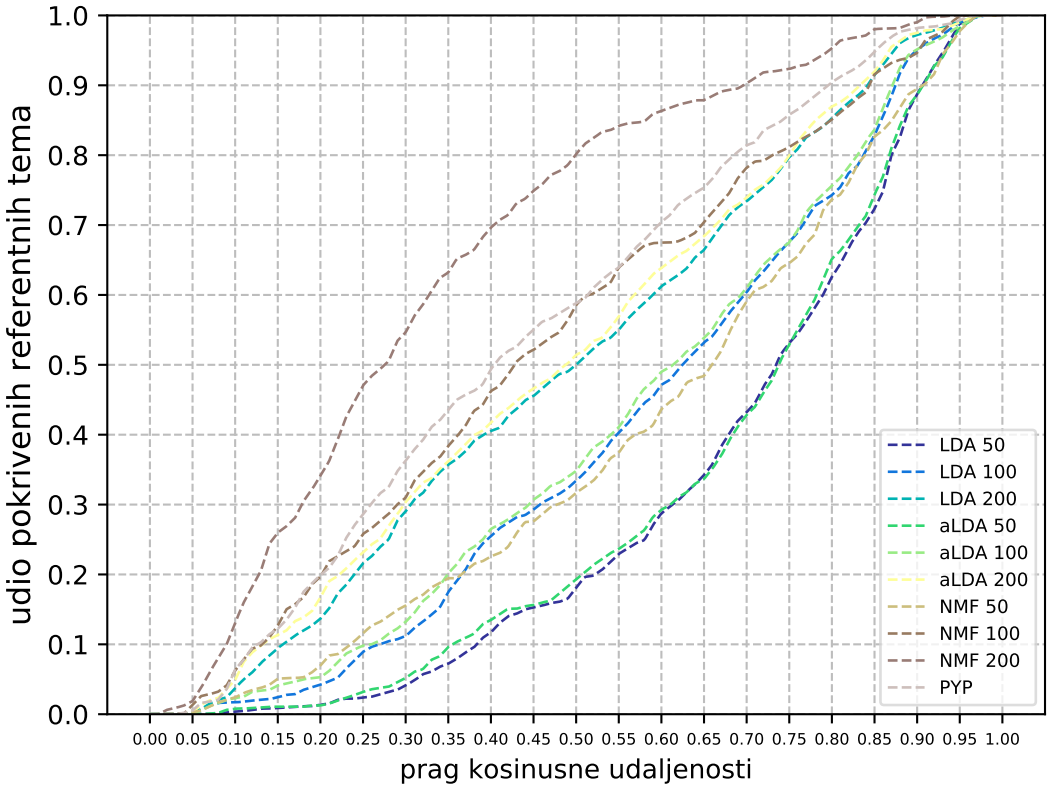
Vrednovanje tematskih modela provodi se na način da se za svaki skup podataka mjeri pokrivenost skupa referentih tema od strane svih izgrađenih tematskih modela. Mjerenje se provodi strogom i relaksiranom varijantom nadzirane mjere, koje su označene sa NP-strog i NP-relaks, te nenadziranom PPU-cosd mjerom. Za svaku mjeru i za svaku kombinaciju klase modela i broja tema izračunati su prosjek i standardna devijacija izmjerene pokrivenosti 10 različitih modela određenih kombinacijom. Rezultati vrednovanja pokrivenosti modela prikazani su u tablici 5.8. Izgrađene su i PU-krivulje za sve kombinacije modela, pri čemu su krivulje, odnosno vrijednosti pokrivenosti za pojedine pragove udaljenost dobivene uprosječivanjem po 10 modela određenih kombinacijom. PU-krivulje svih kombinacija modela prikazane su na slikama 5.17 i 5.18.

Tablica 5.8: Pokrivenost referentnih skupova tema od strane različitih klasa modela mjerena s tri mjere pokrivenosti. Za svaku mjeru i svaki skup podataka označeni su **najbolji rezultati** te najbolji rezultati za pojedini izbor broja tema.

	Medijski skup podataka					
	NP-strog	σ	NP-relaks	σ	PPU-cosd	σ
LDA 50	0.141	0.014	0.383	0.022	0.410	0.006
LDA 100	0.314	0.024	0.565	0.029	0.509	0.009
LDA 200	0.471	0.021	0.716	0.024	0.599	0.006
aLDA 50	0.115	0.017	0.307	0.016	0.407	0.005
aLDA 100	0.265	0.029	0.532	0.026	0.511	0.004
aLDA 200	0.420	0.029	0.717	0.020	0.606	0.008
NMF 50	0.223	0.014	0.320	0.015	0.434	0.004
NMF 100	0.399	0.016	0.510	0.014	0.561	0.006
NMF 200	0.541	0.013	0.619	0.020	**0.647**	0.006
PYP	**0.636**	0.029	**0.842**	0.027	**0.647**	0.007
	Biološki skup podataka					
	NP-strog	σ	NP-relaks	σ	PPU-cosd	σ
LDA 50	0.008	0.007	0.028	0.010	0.308	0.005
LDA 100	0.071	0.012	0.107	0.013	0.402	0.006
LDA 200	0.159	0.015	0.255	0.018	0.504	0.007
aLDA 50	0.009	0.006	0.027	0.006	0.313	0.004
aLDA 100	0.068	0.013	0.107	0.019	0.412	0.007
aLDA 200	0.153	0.015	0.264	0.022	0.516	0.008
NMF 50	0.114	0.008	0.146	0.007	0.394	0.008
NMF 100	0.221	0.012	0.282	0.010	0.536	0.007
NMF 200	**0.442**	0.018	**0.555**	0.022	**0.674**	0.009
PYP	0.229	0.020	0.311	0.018	0.560	0.006



Slika 5.17: PU-krivulje svih klasa modela, za medijski skup podataka.



Slika 5.18: PU-krivulje svih klasa modela, za biološki skup podataka.

Analiza rezultata Pokrivenost dobivena različitim tematskim modelima razlikuje se ovisno o skupu podataka. Kao što je vidljivo iz tablice 5.8, na medijskom skupu najbolju pokrivenost po svim kriterijima daje neparametarski model PYP, dok model NMF sa 200 tema također postiže dobru pokrivenost. Za sve mjere, pokrivenost dobivena parametarskim modelim pozitivno korelira s brojem tema modela. Za parametarske modele vidi se i razlika u prirodi pokrivenosti između faktorizacijskog modela NMF i vjerojatnosnih modela LDA i aLDA. Za svaku vrijednost broja tema NMF daje bolju pokrivenost uz strogo poklapanje tema (NP-strog mjera), dok vjerojatnosni modeli daju bolju pokrivenost uz relaksirano poklapanje koje tolerira šum i malu semantičku varijaciju (NP-relaks mjera). Drugim riječima, modeli NMF precizno otkrivaju referentne teme, dok ih vjerojatnosni modeli otkrivaju manje no još uvijek prepoznatljive varijacije ovih tema. Ova razlika može se vidjeti na odgovarajućim PU-krivuljama na slici 5.17. Za svaku vrijednost broja tema, model NMF daje bolju pokrivenost za manje vrijednosti praga udaljenosti koje odgovaraju preciznom poklapanju tema, dok vjerojatnosni modeli daju bolju pokrivenost za više vrijednosti praga odnosno za aproksimativno poklapanje. Ovaj primjer pokazuje kako se PU-krivulja može iskoristiti za uočavanje finijih razlika u prirodi pokrivenosti koju daju određeni modeli.

Na biološkom skupu podataka modeli NMF dominiraju nad vjerojatnosnim modelima po kriteriju svake od mjera. NMF sa 200 tema daje najbolje pokrivenosti od svih modela, dok za druge vrijednosti broja tema odgovarajući modeli NMF daju bolje rezultate od odgovarajućih parametarskih vjerojatnosnih modela. Neparametarski model PYP postiže više pokrivenosti od svih ostalih vjerojatnosnih modela i usporediv je sa modelom NMF od 100 tema. Svi parametarski vjerojatnosni modeli daju dosta loše rezultate pokrivenosti. Neovisno o klasi modela, rezultati pokrivenosti niži su od rezultata na medijskoj zbirci. Kao i u slučaju medijske zbirke, pokrivenost dobivena parametarskim modelima odlično korelira s brojem tema modela.

Zaključci Rezultati vrednovanja pokrivenosti ukazuju na to da je NMF najbolji izbor modela za otkrivanje tema. model NMF jedini od svih modela ima dobre performanse na oba skupa podataka – dvije različite tekstne domene i dva različita skupa referentnih tema. Za biološki skup podataka NMF je uvjerljivo najbolji model a vjerojatnosni modeli imaju dosta slabe performanse. Na medijskom skupu podataka NMF je kompetitivan s vjerojatnosnim modelima i daje najbolju preciznu pokrivenost referentnih tema, mjerenu NP-strog mjerom, od svih parametarskih modela. S druge strane vjerojatnosni modeli na medijskom skupu daju bolje rezultate relaksirane pokrivenosti.

Ovo ukazuje na to da su NMF teme naučene na medijskim tekstovima interpretabilnije, odnosno da je njihovom interpretacijom lakše vidjeti odgovarajuće semantičke teme nego što je to slučaj kod tema modela LDA i aLDA. S druge strane, modeli LDA i aLDA mogli bi dovesti do aproksimativnog otkrivanja većeg broja semantičkih tema pri čemu bi na interpretaciju tema

očekivano bilo potrošeno više vremena. Kako bi se ove tvrdnje provjerile, trebalo bi provesti kvalitativno vrednovanje tema modela i njihovog odnosa s referentnim temama.

U slučaju medijskog skupa podataka neparametarski model PYP pokazao se najboljim od svih modela i ima vrlo visoke performanse. Na biološkom skupu ovaj model ima najbolje performanse među vjerojatnosnim modelima no zaostaje za modelom NMF.

Model NMF sa 200 tema čini se kao najbolji izbor od svih modela, zbog dobrih performansi na oba skupa podataka i zbog preciznog otkrivanja referentnih tema.

Na medijskom skupu podataka, nadziranim mjerama detektirane razlike u preciznosti pokrivanja između modela NMF i parametarskih vjerojatnosnih modela mogu se jasno vidjeti usporedbom PU-krivulja. Ovo demonstrira korisnost PU-krivulje koja dovodi do zaključka za koji su potrebne dvije različite nadzirane mjere. Također, ovo je dodatni argument u prilog usklađenosti ovih različitih metoda mjerenja.

Na oba skupa podataka pokrivenost raste, za sve mjere, proporcionalno broju tema modela. Ova pojava može se iskoristiti kao argument za validnost mjera pokrivenosti, pošto je razumno očekivati da modeli koji nauče ukupno veći broj tema otkriju više referentnih tema.

Vrlo je zanimljivo pitanje zašto modeli NMF na biološkom skupu imaju uvjerljivo najbolje rezultate dok vjerojatnosni modeli podbacuju. Pri tome treba imati na umu da su biološke referentne teme konstruirane pregledom i izborom tema modela NMF. Moguća su barem dva objašnjenja ove pojave. Prvo objašnjenje je da predložene mjere pokrivenosti nisu neosjetljive na razlike u temama koje su posljedica klase modela. Preciznije, moguće je da je vjerojatnost da mjere procijene poklapanje referentne NMF teme i teme modela veća u slučaju tema modela NMF nego u slučaju vjerojatnosnih modela. S druge strane, mjere su oblikovane tako da budu neosjetljive na klasu modela – nadzirani modeli poklapanja parova tema učeni su na temama uzorkovanih iz svih klasa modela a značajke parova formirane su na temelju mjera neosjetljivih na klasu modela. Nenadzirana PPU mjera temelji se na mjerama udaljenosti tema koje su odabrane upravo zato što su neosjetljive na klasu modela. Opisani način oblikovanja i izgradnje mjera pokrivenosti čini prvo objašnjenje manje vjerojatnim. Drugo moguće objašnjenje je da model NMF bolje otkriva referentne koncepte fenotipova zbog same prirode modela. Ako se zaista radi o toj pojavi, onda su opisani eksperimenti primjer primjene mjera pokrivenosti kako bi se otkrila klasa modela pogodna za dobro modeliranje fenotipnih koncepata. Kako bi se utvrdio točan uzrok ove pojave, potrebno je provesti kvalitativno vrednovanje bioloških tema. Preciznije, trebalo bi utvrditi prirodu modelima otkrivenih tema i prirodu poklapanja tema modela s referentnim temama. Neovisni eksperimenti na srodnim skupovima podataka također bi mogli razjasniti situaciju.

Istraživanje pokrivenosti tema nalazi se u ranoj fazi i provedeni eksperimenti prije svega ukazuju na neke pojave i otvaraju daljnja pitanja. Potrebno je provođenje dodatnih neovisnih eksperimenata koji bi mogli potvrditi ili pobiti neke od navedenih zaključaka, kao i izrada novih

mjera pokrivenosti i usporedba njihovih rezultata s postojećima.

5.9 Rasprava

Problem pokrivenosti tema obuhvaća problematiku mjerenja pokrivenosti skupa referentnih tema (koncepata) od strane automatski naučenih tema modela. Prvi eksperimenti i metode vezane uz pokrivenost predložene su u [29], no unatoč važnosti problema izostala su daljnja istraživanja na ovu temu. U ovom poglavlju definira se problem pokrivenosti tema, predlažu se nove metode vrednovanja tematskih modela s aspekta pokrivenosti i provodi se prvo kvanitativno vrednovanje pokrivenosti na velikom skupu modela i na dvije tekstne domene.

Opisane su dvije metode izrade skupa referentnih tema temeljene na pregledu i doradi tema modela i predložene su dvije metode mjerenja pokrivenosti referentih tema od strane tematskih modela. Prva metoda mjerenja pokrivenosti temelji se na aproksimaciji ljudskih ocjena semantičkog poklapanja tema pomoću nadziranih modela a druga je nenadzirana automatska metoda koja računa pokrivenost na temelju mjere udaljenosti tema. Vrednovanje nadziranih mjera pokazuje vrlo dobru aproksimaciju ljudskih ocjena poklapanja tema na temelju malog skupa značajki udaljenosti tema. Nenadzirane mjere pokrivenosti jako dobro koreliraju s nadziranima a pošto njihova izrada ne zahtijeva označavanje podataka mogu se brzo primijeniti na mjerenje pokrivenosti na novom skupu podataka. Nenadzirane mjere temelje se na izgradnji krivulje pokrivenost-udaljenost, koja se pokazala kao koristan alat za grafičku analizu pokrivenosti koji pruža detaljnije uvide u prirodu pokrivenosti od mjera, koje računaju jednu numeričku ocjenu pokrivenosti.

Na temelju razvijenih mjera i skupova referentnih tema provedeni su prvi kvantitativni eksperimenti mjerenja pokrivenosti referentnih tema od strane velikog skupa tematskih modela iz različitih klasa izgrađenih različitim hiperparametrima. Ovi eksperimenti provedeni su na dva domenski različita skupa podataka – medijskom i biološkom. Eksperimenti demonstriraju smislenost i način primjene predloženih mjera pokrivenosti i daju nove uvide u prirodu tematskih modela s aspekta pokrivenosti tema. Pokazano je da faktorizacijski NMF ima jako dobre performanse odnosno daje visoku pokrivenost uz precizno otkrivanje referentnih tema, te zadržava visoke performanse na oba skupa podataka. Neparametarski model PYP se pokazao kao najbolji izbor među vjerojatnosnim modelima sa osobito visokim performansama na medijskom skupu podataka. Detaljna analiza performansi modela i rezultata mjerenja pokrivenosti provedena je na kraju odjeljka 5.8.

Rezultati eksperimenata pokrivenosti ukazuju na to da bi se za otkrivanje tema na medijskoj agendi trebali koristiti modeli NMF koji nude bolje performanse od jednostavnih vjerojanosnih modela i mogu zadržati visoke performanse pri promjeni skupa podataka. Model NMF trebao bi dovesti i do bržeg postupka otkrivanja tema zbog toga što otkriva velik broj tema koje

se precizno poklapaju s referentnim temama. Zbog navedenih svojstava NMF bi trebao biti dobar izbor za bilo koju primjenu tematskih modela kod koje je poželjna visoka pokrivenost koncepata. Dobar izbor modela za otkrivanje medijskih tema je i neparametarski model PYP koji u eksperimentima pokriva vrlo velik broj referentnih tema, također uz precizno poklapanje. Zanimljivo je da u niti jednom od istraživanja medijske agende, opisanih u odjeljku 3.1.3, u kojima su tematski modeli korišteni za otkrivanje tema, nisu korišteni modeli NMF i PYP, unatoč dostupnosti njihovih implementacija.

Analiza pokrivenosti tematskih modela demonstrira korisnost PPU mjere i korisnost analize pokrivenosti pomoću PU-krivulje. Naime, za razliku od nadziranih mjera čija izrada zahtijeva značajnu količinu ljudskog vremena, za izradu automatske PPU mjeru nisu potrebni označeni podaci, a pokazuje se da PPU pokrivenost odlično korelira s nenadziranim mjerama pokrivenosti. Stoga se PPU mjera može se koristiti za pouzdano određivanje kvalitetnih modela rangiranjem modela. Sama PU-krivulja daje detaljnije uvide u prirodu pokrivenosti i omogućava brzu usporedbu većeg broja modela.

Problem pokrivenosti tema je važan i slabo istražen problem s mnoštvom smjerova za buduća istraživanja. Dva glavna aspekta problema su definicija i izgradnja skupa referentnih tema te izrada mjera pokrivenosti. Buduća istraživanja očekivano će se fokusirati ili na ove aspekte ili na primjene razvijenih metoda na vrednovanje tematskih modela.

Jedan smjer budućih istraživanja odnosi se na poboljšanja ovdje predloženih metoda. Razmatrani postupci izrade referentnih tema mogli bi se pokušati ubrzati vizualnim alatima za pregled tema modela i primjenom metoda vrednovanja tema kako bi se brže eliminirale nekvalitetne teme. Točnost i brzina izrade nadziranih modela za ocjenjivanje jednakosti parova tema vjerojatno bi se mogli poboljšati korištenjem problemu prilagođenih modela klasifikacije i tehnika poput polunadziranog i aktivnog učenja. Aktivno učenje moglo bi dovesti do znatnih ubrzanja i boljeg odabira skupa primjera za učenje u ovom disbalansiranom scenariju u kojem dominiraju parovi međusobno različitih tema. Mjeru PPU bi trebalo učiniti interpretabilnijom, a mogle bi biti korisne i druge modifikacije poput utežavanja pokrivenosti nad određenim intervalima udaljenosti kako bi se naglasio neki aspekt pokrivenosti.

Kada se problem pokrivenosti razmotri u općenitosti, van okvira ovdje predloženih metoda, otkriva se mnoštvo smjerova za daljnja istraživanja i poboljšanja. Skup referentnih tema, koji definira prirodu problema pokrivenosti odnosno vrstu tema koja se očekuje od tematskih modela, mogao bi se definirati na razne načine. Pri tome je prostor mogućih definicija beskonačan pošto odgovara prostoru klasa koncepata. Trebalo bi razmotriti referentne teme na raznim razinama apstrakcije i specifične klase tema poput okvira (engl. *frame*) koji se javljaju u medijskim tekstovima. Općenito, trebalo bi ispitati pokrivenost za skupove referentnih tema reprezentativnih za koncepte kakve je potrebno pokriti u raznim primjenama tematskih modela za eksplorativnu analizu. Ovaj smjer istraživanja mogao bi dovesti do kvantitativnih metoda za

ocjenjivanje relevantnosti tema (problem opisan u odjeljku 3.2), što je važan problem u društvenoznanstvenim primjenama tematskih modela. Svaka nova definicija skupa referentnih tema zahtijeva oblikovanje praktične metode izrade takvog skupa. Stoga je važan smjer istraživanja izrada takvih metoda, koje bi trebale biti automatizirane u što većoj mjeri. Primjerice, skupovi referentnih tema induciranih iz tema modela, poput onih korištenih u ovdje provedenim eksperimentima, mogli bi se pokušati izgraditi automatskom agregacijom i selekcijom tema većeg broja modela. Apstraktnije teme moglo bi se konstruirati na temelju baza znanja poput DBPedia baze ili mnoštva drugih slobodno dostupnih ontologija, mnoge od kojih sadrže koncepte specifične za određenu domenu znanja.

Mjere pokrivenosti skupa referentnih tema određuju prirodu pokrivenosti koja se mjeri i trebale bi biti prilagođene prirodi i reprezentaciji referentih tema. Kod ovdje predloženih metoda pokrivenost se temelji na binarnom ocjenjivanju jednakosti koncepata referentnih tema s konceptima tema modela, no postoje mnogi drugi izgledni pristupi. Poklapanje tema moglo bi se definirati na temelju raznih vrsta semantičke sličnosti koje se mogu modelirati i kategoričkim i numeričkim varijablama. Kod pristupa temeljenih na nadziranom mjerenju poklapanja, definicija poklapanja bi trebala biti smislena za ljudske označivače, a aproksimacija ocjena sličnosti tema može se provesti jednom od mnogih metoda nadziranog, polunadziranog i aktivnog učenja. Osim mjerenja pokrivenosti poklapanjem parova tema pokrivenost bi se mogla mjeriti modelima koji koriste podatke o svim temama i vezama među njima. Jedan primjer je konstrukcija stohastičke matrice koja modelira odnose referentnih tema i tema modela [29], a mogući pristup je adaptacija metoda bipartitnog uparivanja korištenih za ocjenjivanje stabilnosti modela [27]. U slučaju referentnih tema dobivenih iz taksonomija koncepata, mjerenje bi se moglo provesti metodama srodnim metodama označavanja tema modela DBPedia konceptima predloženim u [142].

Konačno, važan smjer istraživanja pokrivenosti je primjena metoda izrade referentnih tema i mjerenja pokrivenosti na vrednovanje tematskih modela. Takve eksperimente trebalo bi provesti za razne tekstne zbirke i pri tome varirati parametre konstrukcije modela poput broja tema, algoritma učenja i metode pretprocesiranja teksta. S jedne strane ovi eksperimenti doveli bi do novih uvida o tematskim modelima a druge strane bi se kroz neovisne eksperimente, promatranjem robusnosti i replikabilnosti rezultata, validirale i metode mjerenja pokrivenosti.

Poglavlje 6

Poboljšanja postupka za analizu medijske agende

U poglavlju 3 predložen je postupak računalne analize medijske agende temeljen na tematskim modelima koji se sastoji od tri koraka: koraka otkrivanja tema, koraka definicije tema i koraka mjerenja tema. Tijekom koraka otkrivanja tema analitičar pregledava tema tematskih modela i interpretira ih kao semantičke teme – tematske koncepte koji se javljaju u tekstovima. U koraku definicije tema analitičar na temelju semantičkih tema odabire postojeće i definira nove semantičke teme te tako definira skup koncepata na kojima se temelji daljnja analiza zbirke. U koraku mjerenja tema provodi se mjerenje pojavljivanja tema iz koraka definicije u tekstovima na način da se gradi model klasifikacije s višestrukim oznakama koji svakom tekstu pridružuje skup tema. Iz tih podataka mogu se izvesti podaci o zastupljenosti tema u tekstovima i supojavljivanju tema, koji se koriste u daljnjim analizama zbirke.

Predloženi postupak analize medijske agende primijenjen je za analizu agende američkih političkih vijesti (odjeljak 3.4) i za analizu agende hrvatskih političkih vijesti u predizbornom i postizbornom razdoblju (odjeljak 3.5). U ovom poglavlju predlažu se poboljšanja tog postupka temeljena na vrednovanjima i opažanjima iz navedenih istraživanja te na metodama vrednovanja dokumentne koherentnosti i pokrivenosti tema predloženih u odjeljcima 3.6.1 i 3.6.2. Razvoj metoda dokumentne koherentnosti i pokrivenosti motiviran je upravo potrebom za metodama vrednovanja ovih aspekata kvalitete tema tematskih modela, koja je uočena tijekom primjene postupka analize agende i opisana u odjeljcima 3.6.1 i 3.6.2.

Glavna poboljšanja postupka analize agende odnose se na korak otkrivanja tema i temelje se na predloženim mjerama dokumentne koherentnosti tema i na vrednovanju pokrivenosti referentnih tema predloženim mjerama pokrivenosti. Mjere koherentnosti koriste se za odabir kvalitetnijih odnosno interpretabilnijih modela te za ubrzanje postupka otkrivanja tema rangiranjem tema prema koherentnosti. Metode vrednovanja pokrivenosti koriste se za odabir kvalitetnijih tematskih modela koji mogu otkriti veći broj semantičkih tema. Svrha ovih poboljšanja je brže

otkrivanje većeg broja semantičkih tema odnosno povećanje efikasnosti postupka otkrivanja tema na medijskoj agendi.

Analiza modela s aspekta koherentnosti i pokrivenosti provedena u odjeljcima 6.1 i 6.2, osim što dovodi do smjernica za poboljšanje otkrivanja tema demonstrira načine korištenja predloženih mjera koherentnosti i pokrivenosti za analizu svojstava tematskih modela.

Ostale preporuke za poboljšanje postupka analize medijske agende odnose se na poboljšanje radnog tijeka otkrivanja tema odnosno postupka pregleda i interpretacije tema, na preporuke za povećanje točnosti mjerenja tema i na prijedlog novog postupka usmjerenog otkrivanja tema koji bi trebao dovesti do mnogo bržeg otkrivanja tema od specifičnog interesa.

6.1 Odabir kvalitetnijih tematskih modela

Preporuke za odabir kvalitetnijih tematskih modela za postupak otkrivanja tema na medijskoj agendi temelje se na mjerenju pokrivenosti skupa referentnih tema od strane tema modela (poglavlje 5) i na mjerenju koherentnosti tema modela (poglavlje 4). Tablica 6.1 prikazuje ocjene pokrivenosti i rangove koherentnosti za tematske modele izgrađene na zbirci američkih medijskih tekstova (odjeljak 5.5).

Pokrivenost referentnih tema, raznorodnog skupa tema tema s medijske agende (odjeljak 5.4), izmjerena je nadziranim mjerama koje računaju pokrivenost na temelju poklapanja referentnih tema i tema modela (odjeljak 5.6). Koriste se mjere strogog poklapanja (mjera NP-strog) i relaksiranog poklapanja (poklapanje uz šum i male semantičke varijacije, mjera NP-relaks). Koherentnost modela ocjenjuje se uprosječivanjem koherentnosti svih tema modela. Koherentnost tema ocjenjuje se mjerom dokumentne koherentnosti GRAF-BROJRIJECI sa najboljim performansama (odjeljak 4.5) te mjerom C_P koja spada među najbolje mjere za procjenu koheretnosti riječi [92].

Za svaku kombinaciju klase modela i broja tema u tablici 6.1, mjere pokrivenosti modela i prosječne koherentnosti tema modela uprosječene su za 10 modela izgrađenih različitim incijalizacijskim parametrima. Tablica prikazuje ocjene pokrivenosti i rangove izračunate na temelju koherentosti.

U istraživanjima medijske agende otkrivanje tema provodi se pretežno na temelju modela LDA (odjeljak 3.1.3). Ovaj model definiran je u odjeljku 3.3.1 kao dobar početni model za postupak otkrivanja tema i eksperimenti analize medijske agende provedeni u odjeljcima 3.4 i 3.5 koriste modele LDA.

Model NMF, kao što se može vidjeti iz tablice 6.1, bolje od modela LDA pokriva referentne teme po kriteriju stroge pokrivenosti NP-relaks. Drugim riječima, za teme modela koje odgovaraju konceptima referentnih tema ti koncepti su detektirani precizno odnosno mogu se lako prepoznati. Dodatno, model NMF ima bolje rezultate dokumentne koheretnosti od svih

Tablica 6.1: Rezultati pokrivenosti i rangovi koherentnosti različitih tipova tematskih modela opisanih u 5.5, izgrađenih na zbirci američkih medijskih tekstova iz 3.4. Za svaki tip modela prikazane su stroga i relaksirana pokrivenost referentnih tema te rang po kriteriju dokumentne koherentnosti i koherentnosti riječi.

	NP-strog	NP-relaks	KohDokum	KohRijeci
LDA 50	0.141	0.383	7	1
LDA 100	0.314	0.565	8	3
LDA 200	0.471	0.716	10	5
aLDA 50	0.115	0.307	5	2
aLDA 100	0.265	0.532	4	4
aLDA 200	0.420	0.717	6	6
NMF 50	0.223	0.320	1	8
NMF 100	0.399	0.510	2	7
NMF 200	0.541	0.619	3	10
PYP	0.636	0.842	9	9

ostalih modela, a dokumentna koheretnost tema je dobar model semantičke koheretnosti medijskih tema (poglavlje 4). Ova dva rezultata pokazuju da NMF uči interpretabilne teme koje precizno pogađaju referentne koncepte, što znači da pregled i interpretacija tema modela NMF očekivano traju kraće od pregleda i interpretacije tema modela LDA i aLDA. Dodatna prednost modela NMF je vrijeme izgradnje modela koje je za red veličine manje nego što je to slučaj kod svih parametarskih vjerojatnosnih modela – nekoliko minuta umjesto nekoliko desetaka minuta u slučaju američke medijske zbirke s približno 25.000 tekstova. Brže učenje modela može biti prednost ukoliko se gradi veći broj modela radi otkrivanja većeg broja semantičkih tema, ili u slučaju da je potrebno eksperimentirati s brojem tema modela kako bi se dobili modeli zadovoljavajuće kvalitete (razmatranja vezana uz broj tema nalaze se u odjeljku 3.3.1). Također, eksperimenti s mjerenjem pokrivenosti u odjeljku 5.8 pokazuju da NMF može zadržati visoke performanse i pri promjeni tekstne domene. Sve opisane razlike u performansama vrijede za svaki izbor broja tema (50, 100 ili 200 tema). Na temelju ovih razmatranja može se zaključiti da je model NMF bolji izbor od modela LDA za početni model (engl. *default model*) na kojem se temelji otkrivanje tema. Primjerice, u slučaju analize agende američkih političkih vijesti prema novim preporukama koristila bi se 3 modela NMF sa 50 tema i 2 modela NMF sa 100 tema.

Neparametarski model PYP ima visoke ocjene pokrivenosti po oba kriterija, bolje od svih drugih modela. Koherentnost tema ovog modela nije visoka po niti jednom kriteriju no kao što se opisuje u sljedećem odjeljku, ovo ne predstavlja problem pošto koherentnost modela

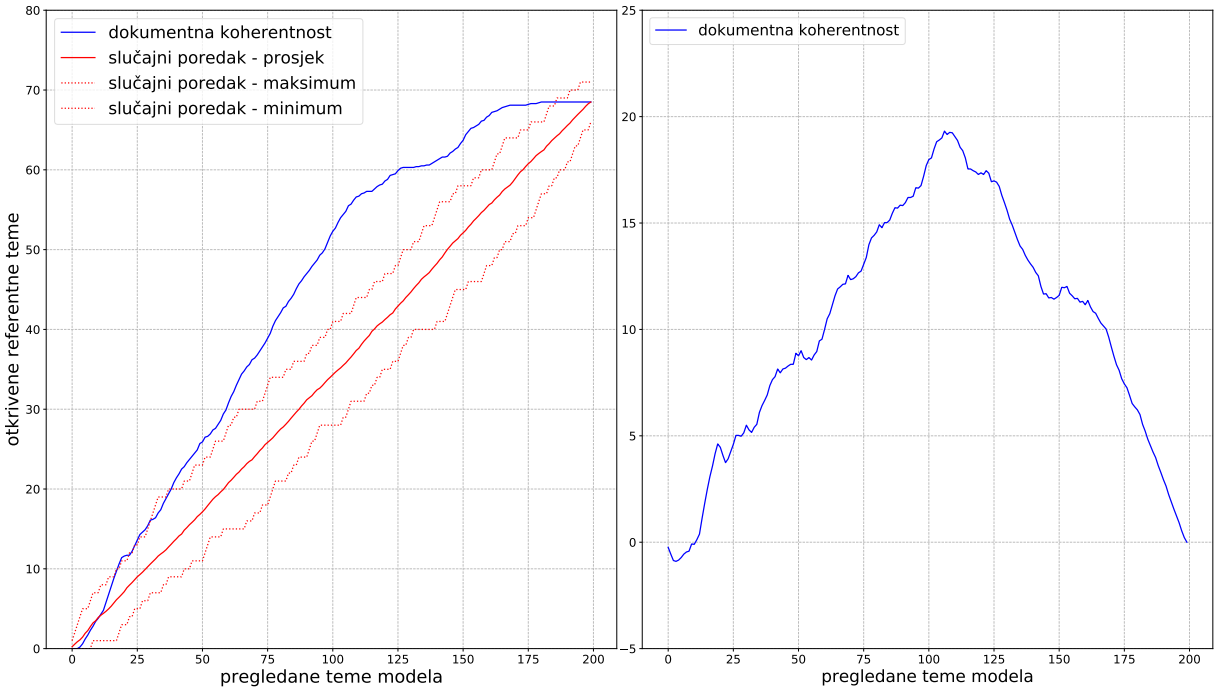
snižavaju "prazne" tj. neodabrane i nenaučene teme koje treba razlikovati od tema koje je neparametarski model odabrao za modeliranje tema u zbirci. Dobri rezultati pokrivenosti pokazuju da je i PYP dobar izbor za postupak otkrivanja medijskih tema, pri čemu proces pregleda tema treba provoditi na način opisan u sljedećem odjeljku. Izbor između modela PYP i NMF ovisi o nekoliko razmatranja. Za učenje modela PYP potrebno je, radi velikog broja varijabli i složene strukture modela, više vremena nego za ostale vjerojatnosne modele – ovo vrijeme iznosi sat i pol u slučaju američke medijske zbirke. Dodatni faktor koji ide u prilog modelu NMF je velik broj dostupnih implementacija.

Model PYP postiže visoku pokrivenost uz nekoliko stotina naučenih tema (između 200 i 300). Stoga, ako analitičar može pregledati samo manji broj tema (manje od 100 tema) ili ako je vrijeme učenja modela važan faktor, NMF je svakako bolji izbor. S druge strane, ako analitičar može pregledati veći broj tema i vrijeme učenja nije ograničavajući faktor, model PYP je vjerojatno bolji izbor. U slučaju da je potrebno pregledati velik broj tema, više nego što ih sadrži jedan model PYP, dobar izbor je izgradnja jednog modela PYP i nekoliko manjih modela NMF.

Parametarski vjerojatnosni modeli LDA i aLDA mogli bi u nekim scenarijima biti dobar izbor za otkrivanje tema. Ovi modeli imaju dobre ocjene relaksirane pokrivenosti (mjera NP-relaks) odnosno dobro otkrivaju referentne koncepte uz pojavu šuma i manje semantičke odmake, što očekivano povećava vrijeme interpretacije tema. U slučaju da vrijeme pregleda tema nije važan faktor a cilj analize je otkrivanje što većeg broja semantičkih tema, LDA s velikim brojem tema bi trebao biti nešto bolji izbor od modela NMF. Dodatno, može se vidjeti da modeli LDA i aLDA imaju najviše ocjene koherentnosti riječi, a kvalitativna analiza tema provedena u odjeljku 4.6.3 ukazuje na dobru korelaciju koherentnosti riječi i apstraktnosti tema. Stoga, ako je od interesa otkrivanje što većeg broja apstraktnih koncepata, modeli LDA s manjim brojem tema su najbolji izbor. Ovi rezultati također pokazuju da bi u slučajevima kada se otkrivanje tema provodi pregledom većeg broja modela, među modele bilo dobro uključiti barem jedan model LDA s manjim brojem tema kako bi se poboljšalo otkrivanje apstraktnijih tema.

6.2 Ubrzavanje otkrivanja tema

Mjere koherentnosti se uobičajeno koriste za vrednovanje tematskih modela, primjerice u eksperimentima analize različitih klasa modela [18, 22] ili, kao primjerice u prethodnom odjeljku, pri odabiru klase modela za neki zadatak. Ovdje se opisuje primjena mjera koherentnosti na ubrzavanje postupka otkrivanja tema, što se postiže na način da se teme modela pregledavaju u padajućem poretku po koherentnosti – koherentnije teme se pregledavaju prije onih manje koherentnih. Mjere koherentnosti u pravilu nisu interpretabilne i namijenjene su relativnoj usporedbi tema i modela, pa je ovo logičan način njihove primjene koji očekivano vodi do bržeg

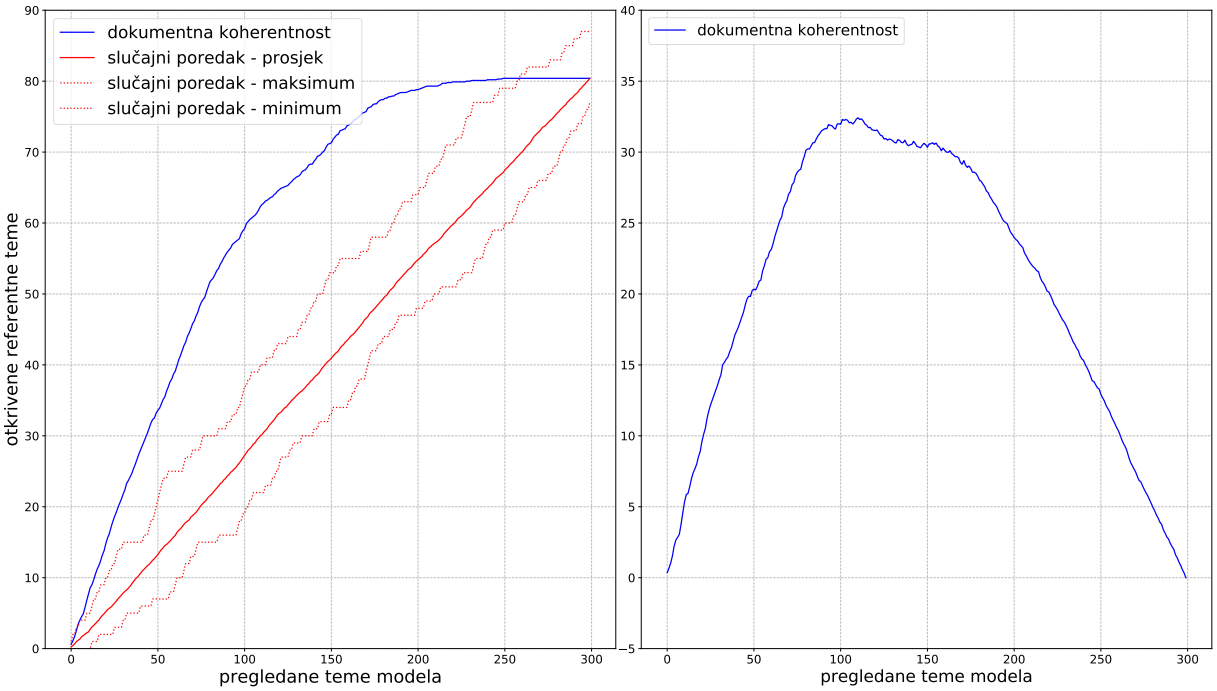


Slika 6.1: Broj otkrivenih tema (y-os) za broj pregledanih tema (x-os): ukupan broj otkrivenih tema (lijevi graf) i prosječna razlika u broju otkrivenih tema (desni graf) između uređaja temeljenog na mjeri koherentnosti i slučajnog uređaja, za model NMF sa 200 tema.

otkrivanja semantičkih tema zbog toga što će koherentnije odnosno interpretabilnije teme biti pregledane prije nekoherentih. Ova metoda predložena je u [93] gdje je pokazano da pregled tema prema padajućoj dokumentnoj koherentnosti vodi do bržeg otkrivanja tema.

U [93] su korišteni modeli LDA iz eksperimenta u odjeljku 3.4 i mjerenje je provedeno na temelju oznaka tema modela semantičkim temama. Ovdje se provodi srodan eksperiment no mjeri se otkrivanje tema iz referentnog skupa medijskih tema (odjeljak 5.4) a za ocjenu poklapanja referentnih tema i tema modela koristi se najbolji nadzirani model poklapanja iz odjeljka 5.6. Slike 6.1 i 6.2 prikazuju grafove ovisnosti broja otkrivenih tema o broju pregledanih tema za teme uređene po koheretnosti i nasumično uređene teme. Pri tome nasumični poredak tema odgovara uobičajenom scenariju pregleda tema po slučajnom poretku tema u modelu. Slike 6.1 i 6.2 prikazuju grafove za model NMF sa 200 tema i za model PYP, dva modela s najboljim rezultatima pokrivenosti na skupu medijskih tema. Grafovi su izgrađeni na temelju skupa modela iz odjeljka 5.5, a svaki broj otkrivenih tema dobiven je uprosječivanjem rezultata 10 različitih instanci modela. Brojevi otkrivenih tema koji odgovaraju slučajnom poretku dobiveni su uprosječivanjem rezultata 5 različitih slučajnih poredaka za svaku od 10 instanci modela (ukupno 50 različitih poredaka).

Kao što se može vidjeti iz grafova, za oba modela pregled tema rangiranih po koherentnosti dovodi do znatnih ubrzanja i većeg broja otkrivenih tema. U slučaju modela NMF, za 100 pregledanih tema modela pregled na temelju koherentnosti u prosjeku dovodi do 50.9 otkrivenih tema, što je poboljšanje od 51% (17.2 tema) u odnosu na pregled po slučajnom poretku koji



Slika 6.2: Broj otkrivenih tema (y-os) za broj pregledanih tema (x-os): ukupan broj otkrivenih tema (lijevi graf) i prosječna razlika u broju otkrivenih tema (desni graf) između uređaja temeljenog na mjeri koherentnosti i slučajnog uređaja, za model PYP.

dovodi do prosječno 33.7 otkrivene teme. U slučaju modela PYP za 100 pregledanih tema upotrebom mjere koherentnosti otkriva se u prosjeku 58.9 tema što je poboljšanje od 32 teme ili 119% u odnosu na 26.9 tema za slučajni poredak.

Za model PYP se dobiva značajno veće ubrzanje otkrivanja tema u odnosu na model NMF. Razlog tome je što je model PYP neparametarski model koji sam uči konačni broj tema T. Učenje broja tema je implementirano korištenjem većeg broja tema od očekivanog, pri čemu model iz podataka nauči T tema dok ostale teme odgovaraju incijalnim vrijednostima odnosno slučajnom šumu. Zbog većeg broja beskorisnih tema u slučaju modela PYP, pregled koherentnih tema prije nekoherentnih stavlja ove teme na kraj liste za pregled, što dovodi do znatnih ubrzanja. Ova pojava je razlog tome što u slučaju modela PYP broj otkrivenih tema ne raste nakon otprilike 250 pregledanih tema – model je ukupno naučio oko 250 tema a ostale teme odgovaraju šumu. Ovo pokazuje kako tek uređivanje tema po koherentnosti čini model PYP dobrim izborom za otkrivanje tema.

Prethodno opisana pojava pomaka beskorisnih tema na kraj liste za pregled može se u manjoj mjeri vidjeti i kod modela NMF kod kojega broj otkrivenih tema prestaje rasti nakon otprilike 175 pregledanih tema. Ovi primjeri pokazuju da se mjere koherentnosti mogu koristiti kako bi se definirao kriterij zaustavljanja postupka pregleda tema – analitičar treba prestati s pregledom kada primijeti da se u listi za pregled počinju pojavljivati samo neinterpretabilne teme.

Koristeći opisani kriterij zaustavljanja, pregled cijelog modela se u slučaju modela NMF

zaustavlja nakon otprilike 175 tema, što je ubrzanje od 25 tema u odnosu na slučajni poredak. Uz pretpostavku da pregled jedne teme modela u prosjeku traje 6 minuta (procjena iz otkrivanja tema provedenog u odjeljku 3.4), primjenom opisane metode analitičar bi uštedio 150 minuta. U slučaju modela PYP koji sadrži više tema koje odgovaraju šumu ušteda je veća – pregled tema završava nakon otprilike 250 tema što dovodi do uštede od 50 tema odnosno 300 minuta.

Prethodni eksperimenti pokazuju da predložene mjere dokumentne koherentnosti mogu učiniti postupak otkrivanja tema znatno efikasnijim. Ista metoda može se primijeniti i u slučaju kada se za otkrivanje koristi veći broj tematskih modela. U tom slučaju očekivano će doći do pojave vrlo sličnih tema iz različitih modela. Pri tome će se takve teme nalaziti blizu jedna drugoj u listi uređenoj po koherentnosti. Kako bi se ubrzao postupak otkrivanja, duplikate je poželjno otkriti i ukloniti iz liste tema. Analiza mjera udaljenosti tema provedena u odjeljku 5.7.3 pokazuje da je dobar kriterij za detekciju vrlo sličnih tema kosinusna udaljenost vektora riječi tema. Ova analiza pokazuje da kosinusna udaljenost vektora riječi jako dobro korelira sa semantičkom sličnosti tema odnosno s ljudskim ocjenama semantičkog poklapanja tema (slike 5.2 i 5.3). Iz ovih grafova može se zaključiti da je kosinusna udaljenost manja od 0.2 dobar kriterij visoke sličnosti tema i taj kriterij može se koristiti za detekciju i uklanjanje duplikata iz liste tema.

Kvalitativna analiza tema iz odjeljka 4.6.3 ukazuje na visoku korelaciju apstraktnosti tema s mjerama koherentnosti riječi. Stoga pregled tema modela padajuće uređenih po koherentnosti riječi očekivano dovodi do izdvajanja apstraktnih tema odnosno ubrzanja otkrivanja apstraktnih semantičkih tema.

6.3 Poboljšanja postupka otkrivanja tema

Ovdje se opisuju poboljšanja postupka otkrivanja tema medijske agende predloženog u odjeljku 3.3.1. Poboljšanja se temelje na odabiru kvalitetnijih tematskih modela (odjeljak 6.1), ubrzanju postupka otkrivanja uređivanjem tema po dokumentnoj koherentnosti (odjeljak 6.2), te na iskustvima primjene postupka otkrivanja na analizu agende američkih političkih vijesti (odjeljak 3.4) i hrvatskih političkih vijesti (odjeljak 3.4).

Poboljšanja temeljena na mjerama vrednovanja modela Ova poboljšanja, izvedena primjenom prethodno predloženih metoda vrednovanja tematskih modela, očekivano dovode do bržeg otkrivanja većeg broja semantičkih tema.

Model NMF je bolji početni izbor modela za otkrivanje tema od modela LDA. Model NMF daje dobru pokrivenost i precizno poklapanje s referentnim temama, uči interpretabilne teme i performanse su mu očekivano robustne s obzirom na promjenu strukture teksta. Model PYP koji daje vrlo visoke ocjene pokrivenosti referentnih medijskih tema je također dobar izbor.

Ako se za otkrivanje koristi veći broj modela preporuča se upotreba jednog modela LDA s manjim brojem tema radi poboljšanja otkrivanja apstraktnijih tema. Detaljnije razlike između ova dva modela i scenariji njihove primjene opisani su u odjeljku 6.1. Pri izgradnji konačnog skupa modela za otkrivanje tema vrijede sva razmatranja iz odjeljka 3.3.1 – trebalo bi koristiti i modele s većim brojem tema i modele s manjim brojem tema.

Nakon što su odabrani modeli koji će se koristiti za otkrivanje tema, postupak otkrivanja može se znatno ubrzati, kao što pokazuje analiza iz odjeljka 6.2, rangiranjem tema modela po padajućoj dokumentnoj koherentnosti pri čemu koheretne teme dolaze na početak liste. Ukoliko se otkrivanje temelji na većem broju modela, preporuča se i primjena kosinusne udaljenosti tema za detekciju i uklanjanje duplikata u listi (detalji primjene nalaze se u 6.2).

Poboljšanja radnog tijeka otkrivanja tema Prilikom pregleda i interpretacije tema preporuča se brzo odbacivanje tzv. spojenih tema (engl. *fused topics*), odnosno tema koje odgovaraju dvama različitim konceptima. Naime, interpretacija spojenih tema zahtijeva više vremena i iako takve teme mogu otkriti veći broj semantičkih tema prethodno opisana poboljšanja dovode do boljeg otkrivanja tema i čine ovu prednost manje važnom. Zbog toga bi spojene teme trebalo interpretirati samo u slučaju da pregled tema koje odgovaraju jednom konceptu nije rezultirao zadovoljavajućim uvidom u medijsku agendu. Upotrebom rangiranja tema po koherentnosti spojene teme će se naći na kraju liste tema, pa se opisana odluka o njihovom pregledu može donijeti nakon pregleda ostalih koherentnijih tema.

Posljedenje poboljšanje se odnosi na upotrebu tablice otkrivenih semantičkih tema, koja je u odjeljku 3.3.1 uvedena kao alat za usklađivanje većeg broja označivača i dobivanje boljeg pregleda otkrivenih tema. Izrada tablice semantičkih tema zahtijeva dodatno vrijeme pošto označivači u postupku otkrivanja moraju utvrditi, za svaku novopregledanu temu modela, odgovara li ona nekoj od postojećih semantičkih tema. Nakon pregleda tema može se dogoditi, unatoč ovoj metodi usklađivanja, da bude potreban još jedan korak dorade i ispravljanja oznaka tema, kao što je to bio slučaj kod analize agende hrvatskih medija opisane u odjeljku 3.5.

Vezano poboljšanje postupka otkrivanja uvodi alternativu upotrebi tablice semantičkih tema. Alternativni postupak je neovisno otkrivanje (pregled, interpretacija i označavanje) tema modela od strane više označivača. Nakon označavanja može se po potrebi provesti dodatni korak usklađivanja oznaka. Ovaj korak trebao bi biti podržan alatima i sučeljem za automatski pronalazak semantički bliskih tema modela i grupiranje tema prema sličnosti. Korak usklađivanja bi se uz ovakve alate provodio pregledom oznaka otkrivenih tema i oznaka njima sličnih tema označenih od strane drugih označivača, te pregledom automatski izgrađenih grupa sličnih tema. I postupak pretraživanja tablice semantičkih tema odnosno pronalazak semantičkih tema sličnih novopregledanim temama modela trebao bi se ubrzati opisanim alatima grupiranja tema i pronalaska sličnih tema.

Odluka koji postupak koristiti ovisi o dostupnom vremenu, broju označivača i potrebnoj kvaliteti konačnih oznaka. Primjerice, ako su potrebne kvalitetne oznake tema modela dobro definiranim semantičkim temama a može se osigurati dosljedna upotreba i popunjavanje tablice semantičkih tema od strane označivača, stari postupak je dobar izbor. Međutim, ako je potrebna brza pregledna analiza agende novopredloženi postupak je bolji izbor.

Novi postupak temeljen na poboljšanjima opisanima u ovom odjeljku trebalo bi vrednovati kroz daljnje eksperimente koje bi trebalo provesti na što većem broju raznolikih medijskih zbirki.

6.4 Preporuke za ostala poboljšanja

U odjeljku 3.3.3 predložena je metoda mjerenja zastupljenosti semantičkih tema pridruživanjem tema tekstnim dokumentima. Metoda pridruživanja temelji se na slabo nadziranim tematskim modelima i postupku izgradnje skupova visokodiskriminativnih riječi koji dobro opisuju semantičke teme. Ova metoda sastoji se od izgradnje skupa riječi za svaku od semantičkih tema i izgradnje tematskih modela čije teme, inicijalizirane skupovima riječi, predstavljaju temelj za odluku o pripadnosti dokumenata semantičkim temama. Vrednovanje ove metode, provedeno u sklopu analize medijske agende opisane u odjeljku 3.4, pokazuje da metoda daje dobre rezultate i da predstavlja bolji izbor od nadziranog modela. Prediktivna validacija metodom dobivenih mjerenja provedena u odjeljcima 3.4 i 3.5 dodatno potvrđuje kvalitetu metode.

Stoga se može zaključiti da je korištenje slabonadziranih tematskih modela za mjerenje zastupljenosti tema dobar pristup i da bi se daljnja poboljšanja postupka mjerenja trebala usmjeriti na poboljšanje ove metode pridruživanja. Prvi smjer poboljšanja odnosi se na postupak odabira skupa riječi. Te riječi trebale bi dobro opisivati samo jednu semantičku temu i biti što manje povezane s ostalim temama. Takve riječi potrebno je izdvojiti iz liste riječi dobivenih na temelju tema modela vezanih sa semantičkom temom (odjeljak 3.3.3). Postupak odabira mogao bi se ubrzati razvojem boljih mjera diskriminativnosti riječi te razvojem alata za vizualizaciju svojstava riječi indikativnih za njihovu diskriminativnost.

Jedan mogući pristup mjerenju diskriminativnosti je promatranje riječi kao značajki za klasifikaciju dokumenata prema pripadnosti temama modela. Pošto se podaci o pripadnosti dokumenata temama modela mogu dobiti iz modela korištenih za otkrivanje tema, postojeće mjere kvalitete klasifikacijskih značajki mogle bi se iskoristiti za mjerenje diskriminativnosti riječi ili skupova riječi. Konačna mjera diskriminativnosti riječi za semantičku temu dobila bi se agregacijom mjera diskriminativnosti riječi za vezane teme modela.

Drugi smjer poboljšanja postupka pridruživanja je poboljšanje strukture tematskih modela za mjerenje. Izrada poboljšanog tematskog modela trebala bi dovesti do poboljšanja klasifikacijskih performansi i ujedno otkloniti potrebu za optimiranjem parametara modela. Struktura

modela trebala bi biti prilagođena mjerenju tema na medijskoj agendi koje su očekivano međusobno srodne odnosno teško razdvojive od drugih sličnih tema. Dobra početna točka za oblikovanje takvih modela su postojeći slabonadzirani tematski modeli [75, 76, 77, 78].

Postupak otkrivanja tema predložen u 3.3.1, kao i prethodna istraživanja u kojima su tematski modeli korišteni za otkrivanje tema (odjeljak 3.1.3), provode neusmjereno otkrivanje tema odnosno ne postavljaju nikakva ograničenja na teme modela. Ovaj pristup svakako je dobar kada je cilj analize dobiti široki pregled medijske agende, no zahtijeva pregled svih tema izgrađenih modela.

S druge strane, u analizi medijske agende od interesa može biti samo manji podskup svih tema. U tim slučajevim postupak otkrivanja tema mogao bi se ubrzati izdvajanjem samo onih tema modela koje su bliske području interesa analitičara. Ovo bi se primjerice moglo postići na način da analitičar definira skup dokumenata i skup riječi koje definiraju tematsku domenu od interesa, nakon čega bi se lista tema modela uredila prema sličnosti s tim riječima i dokumentima. Riječi i dokumenti koji definiraju tematsku domenu pri tome bi se mogli nadopunjavati i mijenjati tijekom postupka pregleda tema. Kriterij "zanimljivosti" tema modela mogao bi se definirati i na temelju sličnosti s prethodno izdvojenim temama od interesa. Na ovaj način analitičar bi mogao u dostupnom vremenu bolje istražiti podskup tema od interesa. Drugi mogući pristup opisanom načinu otkrivanja tema je upotreba interaktivnih tematskih modela [165] (engl. *interactive topic models*). Ovi modeli omogućavaju korisniku da promijeni strukturu tema, primjerice dodavanjem ili uklanjanjem riječi, te da u kratkom vremenu dobije uvid u strukturu ažuriranog modela.

Poglavlje 7

Radni okvir i programska podrška

U ovom poglavlju opisuje se arhitektura i funkcionalnost radnog okvira *pytopia*, koji rješava probleme koji se javljaju pri izgradnji i upravljanju resursima u tematskom modeliranju, a u zadnjem odjeljku opisuje se ostala programska podrška za postupak analize medijske agende koja obuhvaća aplikaciju za sakupljanje tekstova te grafičko sučelje za provođenje postupka otkrivanja tema. Radni okvir *pytopia* razvijen je za potrebe istraživanja provedenog u ovom radu, kao programska podrška postupcima vrednovanja tematskih modela i analize medijske agende.[1]

Važna primjena radnog okvira *pytopia* je proveđenje eksperimenata s tematskim modelima, poput eksperimenata vrednovanja modela u poglavljima 4 i 5. Primjene okvira u analizi medijske agende su izgradnja i odabir modela koji se koriste u postupku otkrivanja tema opisanom u odjeljku 3.3.1, te omogućavanje rada s velikim brojem klasa modela putem standardnih sučelja. Radni okvir može se primijeniti i u mnogim drugim zadacima koji zahtijevaju izgradnju i upravljanje tematskim modelima i vezanim resursima, primjerice za interaktivnu izgradnju modela na temelju korisničkog ulaza ili za implementaciju algoritama koji agregiraju teme većeg broja modela.

U eksperimentima vrednovanja tematskih modela često je potrebno izgraditi veći broj tematskih modela različitih tipova ili varirati parametre procesa izgradnje modela poput zbirke tekstova, pretprocesiranja i hiperparametara [18, 22, 64]. Proces izgradnje tematskih modela, kao i pohrana, učitavanje, dohvat i upravljanje modelima i vezanim komponentama tehnički su složeni zadaci. Radni okvir *pytopia* uvodi oblikovna načela za izradu programske podrške za te zadatke te sadrži potpornu funkcionalnost i niz komponenti[2] koje ubrzavaju razvoj. Arhitektura radnog okvira temelji se na četiri *načela oblikovanja*: načelu standardnih sučelja, načelu

[1] Izvorni kod radnog okvira dostupan je na `https://rebrand.ly/pytopia`.

[2] Pojam komponente označava generičku softversku komponentu koja implementira koherentan skup funkcionalnosti i može ali ne mora odgovarati nekoj Python klasi. Primjeri komponenti su tematski modeli i drugi modeli strojnog učenja, resursi poput riječnika i tekstnih zbirki u raznim stadijima pretprocesiranja, komponente za normalizaciju teksta i izgradnju drugih komponenti, te funkcije poput mjera koherentnosti tema.

identifikabilnosti, načelu dohvata iz konteksta i načelu hijerarhijske kompozicionalnosti. Okvir sadrži potpornu funkcionalnost za implementaciju ovih načela, funkcionalnost za potporu izgradnji, pohrani i učitavanju modela, te niz komponenti za tematsko modeliranje, od temeljnih resursa poput zbirki tekstova i riječnika, preko složenijih resursa poput zbirki vektoriziranih tekstova, do samih tematskih modela i alata za njihovo vrednovanje.

Iako je vrednovanje tematskih modela glavna i motivirajuća primjena radnog okvira, prostor primjena okvira je širi i u načelu obuhvaća bilo koju primjenu tematskih modela. Međutim, radni okvir *pytopia* je najkorisniji u onim primjenama tematskih modela koje zahtijevaju izgradnju većeg skupa tematskih modela i upravljanje tim modelima.

Programski alati za analizu medijske agende mogu iskoristiti načela i funkcionalnosti okvira *pytopia* prije svega za izvedbu koraka otkrivanja tema koji često podrazumijeva izgradnju i pregled većeg broja tematskih modela (odjeljak 3.3.1). Radni okvir se pri tome može koristiti za izgradnju, pohranu i dohvat modela, a standardna sučelja za tematske modele omogućuju eksperimentiranje s nizom klasa modela. U društvenoznanstvenim analizama medijske agende, funkcionalnost pohrane i dohvata modela može se iskoristiti za povećanje replikabilnosti eksperimenata odnosno za pohranu i dijeljenje modela korištenih u eksperimentima.

Platforme koje objedinjuju sakupljanje i pohranu velikih količina medijskih tekstova te alate za njihovu analizu, opisane u odjeljku 3.1.2, mogle bi iskoristiti radni okvir *pytopia* za izgradnju i upravljanje većom količinom modela koji bi se koristili za eksplorativnu analizu tekstova.

U nastavku poglavlja prvo je opisana arhitektura radnog okvira, zatim se daje pregled njegovih funkcionalnosti, a poglavlje završava opisom primjena.

7.1 Arhitektura radnog okvira

Ovdje je opisana arhitektura radnog okvira *pytopia* te softverske tehnologije na kojima se okvir temelji. Prvo se ukratko opisuje programski jezik Python i vezani alati, zatim se opisuje kontekst i opseg okvira, nakog čega slijedi detaljan opis načela oblikovanja.

Radni okvir *pytopia* implementiran je u *programskom jeziku Python* [166, 167]. Programski jezik Python predstavlja dobar izbor za izvedbu ovog okvira pošto se radi o popularnom jeziku i platformi za razvoj i primjenu algoritama strojnog učenja. Drugim riječima, mnogi korisnici i istraživači tematskih modela upoznati su s jezikom Python te postoji velik broj Python alata koji se mogu iskoristiti za izvedbu okvira *pytopia*. Primjeri takvih alata su NumPy i SciPy paketi za znanstveno i tehničko računanje [168] te na tim paketima temeljen radni okvir za strojno učenje scikit-learn [154]. Python ubrzava razvoj programske podrške zbog niza svojstava poput jednostavne sintakse i dinamičkih tipova (engl. *dynamic typing*), te podržava *objektno orijentiranu programsku paradigmu*, metodu oblikovanja koda ključnu za izvedbu okvira *pytopia*. Vrlo je korisno i postojanje Python sučelja prema velikom broju ostalih programskih jezika, uključu-

jući jezike C/C++, Java i R. Ova sučelja olakšavaju primjenu okvira *pytopia* za vrednovanje tematskih modela implementiranih u tim jezicima.

Radni okvir *pytopia* namijenjen je izgradnji resursa i upravljanju resursima u tematskom modeliranju i u kontekstu arhitekture softvera za strojno učenje može se nazvati okvirom srednje razine (engl. *middleware*). Naime, okvir *pytopia* implementira funkcionalnosti smještene između klijentskog koda koji implementira vrednovanje i primjene modela s jedne, te alata koji se koriste za implementaciju algoritama strojnog učenja s druge strane. Primjeri takvih alata su NumPy knjižnica i radni okvir TensorFlow [169]. Preciznije, klijentski programi koriste okvir *pytopia* za izgradnju, pohranu i dohvat modela, te pristupaju modelima putem *pytopia* sučelja, dok *pytopia* koristi funkcionalnosti učenja, pohrane i pristupa modelima implementirane u nekom od alata za izradu i učenje modela.

Kako bi se postigla i zadržala općenitost okvira *pytopia*, te olakšali njegovo korištenje, razvoj i održavanje, opseg okvira *pytopia* ograničen je samo na temeljne funkcionalnosti i manji broj standardnih tematskih modela i uz njih vezanih alata. Temeljene funkcionalnosti obuhvaćaju potporu za primjenu načela oblikovanja, definicije standardnih sučelja, funkcionalnosti za izgradnju, pohranu i upravljanje komponentama te generičke komponente za reprezentaciju i obradu tekstnih podataka. Od tematskih modela okvir bi trebao sadržavati samo implementacije manjeg broja standardnih modela, dok bi svi ostali modeli kompatibilni sa radnim okvirom trebali biti implementirani u vanjskim paketima. Okvir sadrži i manji broj alata za potporu tematskom modeliranju, poput metoda vrednovanja modela i mjera udaljenosti tema.

U nastavku se opisuju *načela oblikovanja radnog okvira pytopia*. Ova načela koriste se za oblikovanje funkcionalnosti okvira, a njihova primjena je nužna ili poželjna pri implementaciji funkcionalnosti temeljenih na radnom okviru. Načela se sastoje od standardnih metoda i načela softverskog inženjerstva no njihova kombinacija definira *općenit i efikasan pristup za oblikovanje i organizaciju programskog koda* za upravljanje tematskim modelima.

7.1.1 Načelo standardnih sučelja

Oblikovanje i upotreba standardnih sučelja softverskih komponenti standardni je postupak softverskog inženjerstva koji omogućava uniformni način rada sa srodnim komponentama [170]. Dobro oblikovanim sučeljima povećava se fleksibilnost i lakoća upotrebe programskog koda i povećava se njegova ponovna iskoristivost (engl. *reusability*) [170]. Primjerice, kod koji koristi jednu komponentu može se primijeniti i na nove komponente istog tipa.

Oblikovanje standardnih sučelja ključno je za radni okvir *pytopia*– standardna sučelja omogućuju dobro definiran i predvidljivi način upotrebe komponenti iz domene tematskog modeliranja. Okvir definira niz standardnih sučelja koja opisuju komponente poput tematskih modela, tema, riječnika i zbirki tekstova, te komponente za izgradnju modela i resursa. Primjer korištenja sučelja je razvoj metode vrednovanja tematskih modela koja modelima pristupa putem

standardnog sučelja – takvu metodu moguće je primijeniti na sve postojeće i buduće modele koji poštuju sučelje.

Primjer sučelja je sučelje za tematski model zadano klasom `TopicModel`. Ovo sučelje definira standardni način korištenja podataka sadržanih u tematskim modelima i njegova upotreba omogućava pristup raznim tipovima modela na predvidljiv način. Okvir 7.1 prikazuje neke od metoda standardnog sučelja tematskog modela.

Okvir 7.1: Python klasa koja definira sučelje tematskog modela

```
class TopicModel(object):
  def numTopics(self)
  def topic(self, topicId)
  def topicVector(self, topicId)
  def topicMatrix(self, dtype=None)
```

Standardna sučelja definiraju se skupom metoda (engl. *method*) i svojstava (engl. *property*) koje moraju sadržavati objekti koji poštuju sučelje. Metode definiraju operacije takvih objekata a svojstva definiraju podatke sadržane u objektima. Metode su definirane imenom i parametrima poziva metode, a svojstva su definirana imenom te tipom objekta na kojeg svojstvo referira. Definicija sučelja ne zahtijeva implementaciju odgovarajuće klase koja definira metode i svojstva sučelja. Umjesto toga, zatijeva se samo da objekti koji poštuju sučelje sadrže te metode i svojstva. Ovaj fleksibilni pristup definiciji tipova naziva se *"duck typing"* i podržan je u jeziku Python. Međutim, sučelje može biti definirano klasom i za važna sučelja implementirane su apstraktne klase koje implementiraju generičke funkcionalnosti sučelja te ujedno služe i za dokumentaciju metoda i svojstava sučelja. Ovakve klase ubrzavaju, mehanizmom nasljeđivanja klasa, implementaciju novih klasa koje poštuju sučelje.

Važna primjena standardnih sučelja je oblikovanje komponenti koje poštuju neko standardno sučelje no izvode većinu operacija pomoću postojeće komponente implementirane u nekom drugom paketu ili programskom jeziku. Ovakve komponente zovu se adapteri i koriste se za ostvarivanje kompatibilnosti vanjske funkcionalnosti s okvirom. Glavna primjena adaptacije je omogućavanje upotrebe postojećih implementacija tematskih modela putem standardnog sučelja `TopicModel`. No adaptacija se može koristiti i za druge komponente povezane s tematskim modeliranjem, poput riječnika i zbirki. Opisana adaptacija je obrazac oblikovanja (engl. *design pattern*) softverskih komponenti važan za široku primjenjivost okvira *pytopia* i naziva se Adapter obrazac [170] (engl. *adapter pattern*). Pri adaptaciji vanjska funkcionalnost može biti implementirana Pythonu ili u nekom drugom programskom jeziku čija funkcionalost se može koristiti putem nekog Python paketa, poput primjerice jezika C, C++, Java i R.

7.1.2 Načelo identifikabilnosti

Svaka komponenta kompatibilna s okvirom *pytopia* mora imati definiran identifikator. Preciznije, svaki objekt kojim okvir upravlja mora imati svojstvo imena `id` kojemu se pristupa naredbom `objekt.id`. Iz perspektive okvira, jednakost objekata odgovara jednakosti odgovarajućih identifikatora. Obavezno definiranje identifikatora komponenti olakšava razne operacije poput izgradnje, dohvata i pohrane, te potiče bolju organizaciju programskog koda. Osim od komponenti-modela i komponenti-resursa poput riječnika i zbirki, identifikabilnost se zahtijeva i od funkcija poput primjerice mjera koherentnosti tema te od procesnih komponenti poput komponenti za normalizaciju teksta.

Ideja je pridružiti komponentama trajni identifikator, trajniji od reference odgovarajućeg python objekta čiji životni vijek korespondira s izvršavanjem programa. Identifikator može biti pridjeljen od strane ljudi ili se može automatski izgraditi na temelju svojstava komponente. Od ljudi pridjeljeni identifikatori olakšavaju rad s važnim komponentama poput temeljnih resursa. Složenost identifikatora ovisi o primjeni okvira a cijena koja se plaća za provedbu identifikabilnosti nije velika – programer koji oblikuje i gradi komponente mora imati na umu nužnost formiranja identifikatora i njihovu namjenu, okvir *pytopia* automatizira izgradnju složenih identifikatora, a Python omogućava lagano pridjeljivanje jednostavnih identifikatora. Okvir 7.2 pokazuje jednostavnost pridjeljivanja od ljudi čitljivog identifikatora zbirci tekstova, primjerice u svrhu imenovanja zbirke u kontekstu eksperimenta.

Okvir 7.2: Jednostavna dodjela od ljudi čitljivog identifikatora resursu.

```
zbirka = TextPerLineCorpus('medijski_tekstovi.txt')
zbirka.id = 'medijska_zbirka'
```

Identifikator je svojstvo (engl. *property*) Python objekta. Tip identifikatora je `str` ili `int` ili bilo koji drugi tip koji se ponaša na sličan način odnosno podržava nekoliko jednostavnih operacija. Preciznije, tip identifikatora mora podržavati usporedbu prema jednakosti (operator ==), računanje hash-vrijednosti (radi mogućnosti upotrebe svojstva `id` kao ključa Python riječnika), te spremanje i učitavanje putem `pickle` mehanizma. Sve ove operacije mogu se definirati za proizvoljnu klasu korištenjem mehanizma preopterećenja operatora [167] (engl. *operator overloading*).

Okvir *pytopia* podržava automatsku izgradnju složenih identifikatora putem klase `IdComposer`. Glavna primjena ove klase je identifikabilnost složenih komponenti definiranih podkomponentama i parametrima. Za takve komponente prirodno je da njihov identifikator bude formiran na temelju identifikatora podkomponenti i vrijednosti parametara, što je lako postići na način da se komponenta definira klasom koja nasljeđuje klasu `IdComposer`. Okvir 7.3 sadrži primjer konstruktora takve klase – identifikator se automatski formira na temelju svih svojstava

konstruiranog objekta, no po potrebi se skup svojstava koja grade svojstvo `id` može kontrolirati.

Okvir 7.3: Automatska izgradnja identifikatora složene komponente.

```
class MyComponent(IdComposer):
  def __init__(self, subComp1, subComp2, param1, param2):
    self.sc1, self.sc2 = subComp1, subComp1
    self.param1, self.param2 = param1, param2
    IdComposer.__init__(self)
```

Prilikom definicije identifikatora i njihove upotrebe potrebno je imati na umu da okvir *pytopia* pretpostavlja da jednakost komponenti odgovara jednakosti njhovih identifikatora. Korisnik mora imati na umu ovo pravilo prilikom definicije temeljnih resursa i drugih važnih resursa čiji identifikatori se pridjeljuju "ručno", dok za automatski formirane identifikatore složenih komponenti ovo svojstvo osigurava okvir pod pretpostavkom da svojstvo vrijedi za sve njihove podkomponente. Primjerice, ako se promijeni identifikator neke zbirke tekstova, okvir će smatrati taj resurs i sve izvedene resurse različitim od starih verzija, što može dovesti do grešaka i bespotrebne izgradnje dupliciranih resursa. Iz opisanih razloga, svi važni resursi s od ljudi pridjeljenim identifikatorima trebali bi imati međusobno različite identifikatore unutar razumnog "konteksta upotrebe", primjerice konteksta jednog projekta.

Opisani princip identifikabilnosti objekata kojima okvir upravlja ključan je za izvedbu okvira i ima mnoge primjene. Identifikabilnost može služiti kao mehanizam imenovanja i referenciranje resursa od posebnog interesa od strane programera. Primjeri ovakvih resursa su tekstne zbirke i riječnici na kojima se temelji izgradnja modela nekog eksperimenta. Identifikator se može koristiti za dohvat resursa, primjerice resursa pohranjenih na disk ili resursa pohranjenih u priručnu memoriju (engl. *cache*). Općenito, objekte je moguće referencirati pomoću identifikatora odnosno varijable mogu sadržavati identifikatore a ne adrese objekata. Ovo primjerice omogućava dohvat objekata iz vanjske memorije prema potrebi te ponovno oslobađanje memorije kada objekt više nije potreban. Također, na ovaj se način olakšava pohrana složenih objekata pohranom jednostavnih identifikatora podkomponenti koji zauzimaju malo memorije.

7.1.3 Načelo dohvata iz konteksta

Svaki objekt kojim se upravlja pomoću okvira *pytopia* ima jedinstveni identifikator koji služi kao trajna referenca tog objekta. Kontekst je mehanizam okvira koji omogućava dohvat objekata na temelju identifikatora i usko je povezan s načelom identifikabilnosti. Kontekst ujedno služi i za grupiranje objekata i na taj način potiče dobru organizaciju resursa i modela koja olakšava njihovu upotrebu. Kontekst je implementiran kao klasa `Context` koja podržava operacije python riječnika [167] (engl. *dictionary*) te se može koristiti kao Pythonov kontekst [167]. Ope-

racije riječnika omogućavaju dohvat na temelju identifikatora te dodavanje objekata u kontekst i njihovo uklanjanje. Mehanizam konteksta u jeziku Python omogućava jednostavnu sintaksu za operacije koje implementiraju "ulazak" u kontekst i "izlazak" iz konteksta. Preciznije, kontekst odgovara skupu postavki koje utječu na kod koji se izvršava unutar konteksta, odnosno koji se nalazi unutar bloka definiranog ključnom riječi `with`.

Identifikatori omogućavaju da se umjesto sa složenim objektom manipulira s manjim identifikacijskim objektom (očekivano znakovnim nizom ili brojem) kojeg je lakše spremati i prenositi i koji prema potrebi može biti zadan i korišten od strane ljudi. No u trenutku kada je potrebno pristupiti objektu na temelju identifikatora, on se dohvaća iz konteksta pomoću metode `resolve(id)`. Ovaj mehanizam izveden je pomoću globalnog konteksta – objekta klase `Context` jedinstvenog za cijeli Python program i dostupnog metodi `resolve`. Globalni kontekst dinamički se mijenja tijekom izvršavanja programa – na početku je prazan a za svaki pojedini kontekst (grupu identifikabilnih objekata) prilikom ulaska u kontekst odnosno izlaska iz konteksta objekti se dodaju u globalni kontekst odnosno uklanjaju iz globalnog konteksta. Okvir 7.4 sadrži jednostavan primjer upotrebe konteksta. Metoda `kontekstZbirki` definira tj. gradi kontekst i nalazi se u vanjskom paketu. Ulaskom u kontekst, koji se provodi izvršavanjem ključne riječi `with`, sve zbirke se dodaju u globalni kontekst i postaju dostupne na temelju njihovih identifikatora. Ovaj primjer ilustrira scenarij u kojem programer-korisnik zna identifikatore određenih tekstnih zbirki (primjerice na temelju dokumentacije) i piše kod za obradu tekstova iz tih zbirki.

Okvir 7.4: Upotreba konteksta koji sadrži tekstne zbirke za projekt.

```
from resursi import kontekstZbirki
with kontekstZbirki():
  zbirka1 = resolve('medijska_zbirka1')
  zbirka2 = resolve('medijska_zbirka2')
  for tekst in zbirka1:
    if zbirka2.textIds(tekst.id):
      print 'tekst se nalazi u obje zbirke:', tekst.id
```

Zamišljeni način korištenja konteksta za organizaciju resursa i drugih komponenti je taj da programer definira i dokumentira kontekste koji odgovaraju srodnim skupovima komponenti te ih učini dostupnim putem metoda ili varijabli. Primjerice, mogu postojati odvojeni konteksti za zbirke, riječnike, komponente za normalizaciju teksta, te za izgrađene tematske modele. Kod koji koristi prethodno definirane kontekste izvšava se unutar odgovarajućih `with` blokova. Kontekst, odnosno objekt klase `Context`, lako je izgraditi dodavanjem identifikabilnih objekata ili svih objekata drugog konteksta putem metoda `Context.add(obj)` i `Context.merge(ctx)`. Okvir 7.5 prikazuje primjer izgradnje konteksta za zbirke i riječnike te njihovo dodavanje u

kontekst s resursima cijelog projekta.

Okvir 7.5: Izgradnja konteksta sa temeljnim resursima.

```
def kontekstRijecnika(): ...
def kontekstZbirki():
  ctx = Context('kontekst zbirki')
  ctx.add(zbirka1)
  ctx.add(zbirka2)
  return ctx
def kontekstResursa():
  ctx = Context('temeljni resursi')
  ctx.merge(kontekstZbirki())
  ctx.merge(kontekstRijecnika())
  return ctx
```

Organizacija resursa i drugih komponenti u kontekste olakšava njihovo dijeljenje. Zamišljeni način dijeljenja je izgradnja i dokumentacija Python paketa koji sadrži metode koje grade kontekste koji odgovaraju nekom skupu podataka. Okvir 7.6 prikazuje primjer izgradnje modela iz temeljnih resursa. Argumenti metode izgradnje modela su identifikatori temeljnih resursa (zbirke i riječnika) te identifikator komponente za normalizaciju teksta.

Okvir 7.6: Izgradnja modela na temelju osnovnih resursa dohvaćenih iz konteksta.

```
from projekt.resursi import *
with kontekstResursa():
  with kontekstNormalizatora():
    izgradiTematskiModel(zbirka='medijska_zbirka2',
    rijecnik='rijecnik1', normalizator='hrvatskoKorijenovanje')
```

7.1.4 Načelo hijerarhijske kompozicionalnosti

Hijerarhijska kompozicionalnost je pristup oblikovanju složenih komponenti pytopia okvira. Ove komponente su određene parametrima izgradnje te podkomponentama, koje pak mogu biti strukturirane na isti način. Ova rekurzivna struktura odgovara stablu u kojem su čvorovi komponente a listovi parametri i jednostavne komponente.

Standardna sučelja, identifikabilnost i konteksti nužni su za implementaciju i korištenje pytopia okvira – očekuje se da komponente poštuju sučelja i imaju definirane identifikatore, a mapiranje identifikatora na objekte ostvaruje se pomoću konteksta. S druge strane, primjena načela hijerarhijske kompozicionalnosti nije nužna za upotrebu pytopia okvira – objekti koji su

instance komponenti ne moraju biti izgrađeni na ovaj način već je dovoljno je da imaju identifikatore i poštuju odgovarajuća sučelja. Međutim, načelo je vrlo korisno kao konceptualni alat za organizaciju koda. Komponente okvira *pytopia* oblikovane su prema tom načelu a preporučeni način izgradnje novih složenih komponenti je upravo primjena načela i korištenje jednostavnijih komponenti.

Hijerarhijska kompozicionalnost u skladu je s prirodom komponenti koje opisuju modele i složene resurse kakvi se javljaju u tematskom modeliranju i općenito u dubinskoj analizi teksta i strojnom učenju. Te komponetne, primjerice tematski modeli, definirane su skupom parametara izgradnje te podkomponentama čiji podaci i operacije se koriste u izgradnji nadkomponente. Treba napomenuti da je oblikovanje komponentni na ovaj način fleksibilno – svaka složena komponenta može se oblikovati hijerarhijom podkomponenti na razne načine i pri tome struktura komponente i njena složenost (definicije podkomponenti i parametara) mogu ovisiti o primjeni i vremenu dostupnom za razvoj. Kod objektno orijentiranog pristupa oblikovanju koda složene komponente očekivano odgovaraju klasama koje definiraju koherentni skup podataka i operacija nad tim podacima. Tablica 7.1 sadrži primjere hijerarhijski kompozicionalnih komponenti okvira *pytopia* i njihovih podkomponenti.

Princip hijerarhijske kompozicionalnosti prije svega se odnosi na postupak izgradnje modela te definira parametre metoda izgradnje modela koji očekivano odgovaraju identifikatorima podkomponenti i parametrima izgradnje. Automatska izgradnja identifikatora složenih objekata podržana je od okvira *pytopia* putem klase `IdComposer`. Ova klasa usklađena je s načelom hijerarhijske kompozicionalnosti i može se koristiti za izgradnju identifikatora iz identifikatora podkomponenti i vrijednosti parametara izgradnje.

Uzmimo kao primjer komponentu koja odgovara tematskom modelu. Izgradnja tematskog modela ovisi o zbirci tekstova koja sadrži podatke za učenje, metodi pretprocesiranja i normalizacije teksta koja definira konačni oblik riječi, te o riječniku koji definira konačni skup riječi i njihove indekse koji određuju matricu tema-riječ. Pri tome svaka od ovih podkomponenti može biti definirana vlastitim parametrima, primjerice parametrima izgradnje riječnika poput načina filtriranja čestih i rijetkih riječi. U slučaju tematskog modela NMF, koji se uobičajeno gradi iz tekstova predstavljenih tf-idf težinama, moguće je definirati podkomponentu koja odgovara zbirci tekstova predstavljenoj matricom tf-idf težina. Tematski model NMF ovisi o toj komponenti koja pak i sama ovisi o riječniku, zbirci tekstova i pretprocesiranju, te o parametrima koji definiraju računanje tf-idf težina. Tematski model ovisi i o raznim parametrima izgradnje, prije svega o parametrima algoritma učenja. Okvir 7.7 prikazuje izgradnju modela NMF definiranog klasom oblikovanom prema principu hijerarhijske kompozicionalnosti.

Okvir 7.7: Inicijalizacija i izgradnja tematskog modela NMF.

```
from sklearn.decomposition import NMF
class NmfModel(TopicModel, IdComposer):
```

```
    def __init__(self, zbirka, rijecnik, normalizator, brojTema,
                 pretprocesiranje='tf-idf', randomSeed=1):
        self.zbirka, self.rijecnik, self.normalizator = \
          zbirka, rijecnik, normalizator
        self.brojTema, self.pretprocesiranje, self.randomSeed = \
          brojTema, pretprocesiranje, randomSeed
        IdComposer.__init__(self)
    def build(self):
        if self.pretprocesiranje == 'vjerojatnosti':
          matrica = self.__matricaVjerojatnosti()
        elif self.pretprocesiranje == 'tf-idf':
          # izgradnja tf-idf podkomponente
          matrica = self.__matricaTfIdf()
        nmf = NMF(n_components=self.brojTema, solver='pg',
                  init='nndsvd', random_state=self.randomSeed)
        w = nmf.fit_transform(matrica)
        self.__nmf = nmf
        self.__w = w
from projekt.resursi import kontekstResursa
with kontekstResursa():
  nmf = NmfModel('medijska_zbirka', 'rijecnik1',
                 'korijenovanje', 100)
  nmf.build()
```

7.2 Funkcionalnost radnog okvira

Dok načela oblikovanja definiraju oblikovanje strukture i način upotrebe komponenti okvira, funkcionalnost okvira sastoji se od implementiranih komponenti i potporne funkcionalnosti te se može koristiti za implementaciju eksperimenata, aplikacija i novih komponenti.

Funkcionalnost radnog okvira *pytopia* sastoji se od niza komponenti koje odgovaraju temeljnim resursima poput zbirki tekstova, izvedenim resursima različite složenosti, i tematskim modelima, te od funkcionalnosti za vrednovanje tematskih modela, funkcionalnosti za izgradnju, pohranu i učitavanje komponenti, i potporne funkcionalnosti poput podrške za testiranje.

Implementirana funkcionalnost poštuje prethodno opisana načela oblikovanja – komponente poštuju standardna sučelja, definiraju identifikatore i strukturirane su hijerarhijski kompozicionalno. Temeljna funkcionalnost okvira za identifikabilnost i kontekst već je navedena i sastoji

se od klase `IdComposer` za automatsku izgradnju identifikatora, klase `Context` koja implementira kontekst, te metode `resolve` za dohvat komponenti iz konteksta na temelju identifikatora. Načelo standardnih sučelja izvedeno je implementacijom niza sučelja koja definiraju razne komponente i potpornu funkcionalnost. Načelo hijerarhijske kompozicionalnosti reflektira se prije svega u samom oblikovanju i implementaciji složenih komponenti i sučeljima za njihovu izgradnju.

7.2.1 Izgradnja objekata

Za svaku komponentu, osim klase koja definira komponentu, potrebno je definirati i metodu ili klasu za izgradnju objekata koji su instance te komponente. Na ovaj način razdvaja se (engl. *decoupling*) implementacija komponente i implementacija njezine izgradnje, čime se postiže fleksibilniji kod.

Primjerice, na ovaj način moguće je modificirati proces izgradnje ili definirati više metoda izgradnje. Samo sučelje izgradnje je jednostavno – funkcijski poziv koji prima parametre izgradnje komponente. Komponenta za izgradnju komponenti, tzv. graditelj (engl. *builder*), mora podržavati funkcijski poziv (engl. *callable object*) i mora na temelju parametara poziva provesti izgradnju odgovarajuće komponente odnosno objekta koji je instanca te komponente. Drugim riječima, graditelj je komponenta koja gradi komponente i može imati definiran identifikator i biti dostupan kroz kontekst.

Uobičajeni pristup izgradnji komponenti je implementacija izgradnje unutar same komponente odnosno klase koja definira komponentu. Ovaj pristup je također podržan od okvira *pytopia* i kompatibilan s prethodno opisanim pristupom sa zasebnim graditeljem. Takva klasa koja sama implementira izgradnju objekta mora definirati konstruktor koji prihvaća parametre izgradnje i metodu `build()` bez parametara. Izgradnja se provodi u dva koraka – koraku inicijalizacije objekta `obj` pomoću konstruktora i koraku izgradnje pozivom metode `obj.build()`. Odvojeni graditelj za ovakve komponente može se lako implementirati na temelju opisanog sučelja i radni okvir sadrži standardnu klasu za definiciju takvih graditelja.

Upotreba opisanih konvencija izgradnje objekata nije obavezna – svaki objekt može se izgraditi na proizvoljan način, a važno je samo da bude identifikabilan i da poštuje standardna sučelja. Međutim, organizacija izgradnje na opisani potiče bolju organizaciju koda, te omogućava automatizaciju procesa izgradnje i upotrebu funkcionalnosti okvira za potporu izgradnji. Primjerice, uz korištenje odvojenih graditelja vrlo je lako implementirati priručnu memoriju (engl. *caching*) i bilježenje (engl. *logging*). Ovo se postiže upotrebom klasa-omotača (engl. *wrapper*) implementiranih u okviru koje mijenjaju ponašanje graditelja. U slučaju priručne memorije modificirani graditelj na temelju identifikatora izgrađenih objekata pohranjuje i dohvaća objekte iz priručne memoriju. U slučaju bilježenja modificirani graditelj nakon svake izgradnje bilježi podatke o izgrađenim komponentama i eventualnim greškama.

Okvir *pytopia* implementira standardne komponente koje se mogu koristiti za izgradnju modela i drugih složenih komponenti. Primjeri takvih komponenti su tf-idf matrica zbirke tekstova ili indeks zbirke odnosno fiksno mapiranje identifikatora tekstova na cijele brojeve. Graditelji tih standardnih komponenti važan su dio okvira i njihovi identifikatori su unaprijed definirani i fiksni. Primjerice, identifikatori graditelja pretodno opisanih komponenti su `"corpus_tfidf_builder"` i `"corpus_index_builder"`. Te graditelje stoga je moguće dohvatiti na temelju standardnih identifikatora, no da bi to bilo moguće za svaki projekt je potrebno definirati kontekst sa tim graditeljima te izvršavati kod izgradnje resursa unutar tog konteksta. Taj korak je lako implementirati pošto okvir sadrži metodu `basicBuildersContext(cacheFolder)` koja gradi kontekst s graditeljima standardnih komponenti koji su konfigurirani da koriste priručnu memoriju za pohranu izgrađenih objekata.

Opisana organizacija izgradnje komponenti ima za posljedicu važnu optimizaciju. Sve složene komponente izgrađene iz istih temeljnih resursa grade iste standardne podkomponente. Primjerice, dva NMF modela izgrađena istim hiperparametrima iz iste zbirke i riječnika i pomoću iste metode normalizacije teksta temelje se na istoj tf-idf reprezentaciji zbirke tekstova. Slično, modeli NMF i LDA izgrađeni iz istih temeljnih resursa koriste iste reprezentacije vreće riječi i isti indeks dokumenata u zbirci. Opisani način izgradnje pomoću standardnih graditelja s djeljenom priručnom memorijom ima za posljedicu to da se svaka od podkomponenti gradi i pohranjuje samo jednom, što dovodi do vremenskih i prostornih ušteda koje su posebno važne kod izgradnje velikog broja modela.

Okvir 7.8 daje primjer izgradnje NMF modela korištenjem zasebne komponente-graditelja. Sam NMF model podržava i samoizgradnju putem konstruktora i metode `build` (okvir 7.7). Izgradnja pomoću graditelja izvedena je u dva koraka - formiranjem osnovnog graditelja na temelju klase NMF modela (operacija `SelfbuildResourceBuilder(NmfModel)`) te formiranjem graditelja-omotača koji implementira priručnu memoriju. Graditelj prihvaća iste parametre izgradnje kao i sama klasa modela (okvir 7.7).

Okvir 7.8: Izgradnja tematskog modela NMF pomoću graditelja.

```
from pytopia.adapt.scikit_learn.nmf import NmfModel
from pytopia.resource.builder_cache import ResourceBuilderCache
from pytopia.resource import SelfbuildResourceBuilder
nmfBuilder = SelfbuildResourceBuilder(NmfModel)
cachedBuilder = ResourceBuilderCache(nmfBuilder, '/model_cache')
from projekt.resursi import kontekstResursa
with kontekstResursa():
  nmf = cachedBuilder('medijska_zbirka', 'rijecnik1',
                      'korijenovanje', 100)
```

7.2.2 Pohrana i učitavanje objekata

Izgrađene objekte potrebno je pohranjivati u i učitavati iz trajne vanjske memorije. Okvir *pytopia* podržava pohranu objekata u sustav datoteka (engl. *file system*) pri čemu se svaki objekt pohranjuje u zasebni direktorij. Sučelje prema ovoj funkcionalnosti sastoji se od dvije metode: metode `saveResource(object, folder)` i metode `loadResource(folder)`.

Da su sve *pytopia* komponente implementirane kao čiste Python klase temeljene na standardnim Python tipovima, pohrana i dohvat njihovih objekata mogli bi se izvesti pomoću Python paketa `pickle`. Međutim, komponente se mogu sastojati od složenih struktura podataka čiju pohranu i učitavanje nije moguće ili poželjno izvesti putem `pickle` paketa. Dodatno, klase mogu sadržavati i strukture implementirane u drugim alatima i programskim jezicima – ovo je očekivano u slučaju klasa koje adaptiraju vanjsku funkcionalnost na *pytopia* sučelja. Kako bi se podržala pohrana i dohvat u svim navedenim slučajevima definiran je jednostavan protokol koji komponente moraju poštovati da bi bile kompatibilne sa standardnim metodama `saveResource` i `loadResource`. Okvir 7.9 prikazuje pohranu i učitavanje NMF modela čija izgradnja je prikazana u okvirima 7.8 i 7.7.

Podaci svake komponente se općenito sastoje od dva dijela – čistih Python podataka koji se mogu pohraniti putem `pickle` paketa, te složenih i vanjskih podataka za čiju pohranu i dohvat je odgovorna sama komponenta. Komponenta mora definirati dvije metode: `save(folder)` i `load(folder)`. Metoda `save` mora prvo pohraniti sve Python podatke u datoteku predefiniranog imena, a zatim pohraniti ostale podatke u druge datoteke. Metoda `load` učitava sve vanjske podatke za objekt čiji su čisti Python podaci, koji definiraju jezgru objekta, već učitani `pickle` paketom. Metoda `saveResource` jednostavno poziva `save` metodu objekta, dok `loadResource` prvo pomoću `pickle` paketa učitava objekt s čistim Python podacima, a zatim poziva `load` metodu tog objekta kako bi se učitali ostali podaci.

Okvir 7.9: Pohrana i učitavanje tematskog modela.

```
from pytopia.resource.loadSave import saveResource, loadResource
from projekt.resursi import kontekstResursa
from projekt.builders import nmfBuilder
with kontekstResursa():
  nmf = nmfBuilder('medijska_zbirka', 'rijecnik1',
                   'korijenovanje', 100)
  nmf.id = 'mojNmfModel'
  saveResource(nmf, '/modeli/nmf/' + nmf.id)
...
nmf = loadResource('/modeli/nmf/mojNmfModel')
```

7.2.3 Komponente za modeliranje teksta

Ove komponente, koje predstavljaju temeljne resurse, složene resurse i same modele, čine jezgru radnog okvira *pytopia*. Ostala potporna funkcionalnost olakšava izgradnju, pohranu, učitavanje, dohvat i upravljanje ovim komponentama. Ovdje opisane komponente poštuju načela oblikovanja radnog okvira – oblikovane su hijerarhijski kompozicionalno, poštuju ili same definiraju standardna sučelja, a njihovi identifikatori se automatski grade na temelju parametara i podkomponenti.

Ove komponente mogu se okvirno podijeliti u tri skupine prema razini apstrakcije: temeljne komponente koje predstavljaju temeljne resurse i nemaju podkomponenti, komponente srednje razine izvedene iz temeljnih resursa i korištene za izgradnju modela, te komponente aplikacijske razine koje sadrže tematske modele i uz njih vezane komponente.

Temeljne komponente Prva skupina komponenti uključuje temeljne resurse i temeljne procesne komponente. To su zbirka tekstova (klasa `Corpus`) koja sadrži objekte klase `Text`, riječnik (klasa `Dictionary`) te komponente koje pretprocesiraju tekst (klasa `Text2Tokens`). Sve ove klase definiraju standardna sučelja za odgovarajuće komponente. Klasa `Text2Tokens` je apstrakcija koja obuhvaća sve vrste pretprocesiranja teksta poput korijenovanja i uklanjanja zaustavnih riječi – komponente ovog tipa transformiraju tekst (znakovni niz) u listu pojavnica (engl. *token*).

Osim klasa koje definiraju sučelja okvir sadrži i implementacije ovih komponenti koje omogućuju korisniku brzu izgradnju resursa. To su primjerice klasa `TextPerLineCorpus` koja čita datoteku-zbirku i interpretira retke datoteke kao `Text` objekte, te komponente pretprocesiranja za engleski i hrvatski jezik.

Komponente srednje razine Ovaj skup komponenti sadrži komponente koje su izgrađene iz temeljnih no njihova glavna namjena nije upotreba u aplikacijskom kodu (kod eksperimenata, kod za analizu podataka, ...) već izgradnja modela više razine, poput tematskih modela. Primjer takve komponente je klasa `BowCorpus` odnosno zbirka tekstova u formatu veće riječi. Ova komponenta, definirana je zbirkom tekstova, komponentom pretprocesiranja (klasa `Text2Tokens`) i riječnikom.

Komponente srednje razine oblikovane su hijerarhijski kompozicionalno i svaka od njih je definirana pomoću komponenti temeljne razine ili drugih komponenti srednje razine. Upotreba ovih komponenti olakšava zadatke analize teksta i ubrzava izgradnju tematskih modela. Kao što je navedeno u opisu mehanizama izgradnje, svaka od ovih komponenti ima odgovarajućeg graditelja koji je očekivano dostupan kroz kontekst putem standardnog identifikatora. Primjerice graditelj za `BowCorpus` komponentu može se dohvatiti identifikatorom `"bow_corpus_builder"`. Sam korisnik okvira je odgovoran za definiciju konteksta s graditeljima standardnih komponenti

Tablica 7.1: Primjeri komponenti srednje i aplikacijske razine. Za svaku komponentu navedene su podkomponente i komponente koje se koriste pri izgradnji ali i samo ovise o podkomponentama (označene podcrtano).

Komponenta	Svrha	Ovisi o komponentama
`CorpusIndex`	indeksiranje tekstova u zbirci	`Corpus`
`BowCorpus`	tekstovi kao vreće riječi	`Corpus`, `Dictionary`, `Text2Tokens`, `CorpusIndex`
`CorpusTfidf`	zbirka kao tf-idf matrica	`Corpus`, `Dictionary`, `Text2Tokens`, `CorpusIndex`, `BowCorpus`
`WordDocIndex`	generička vektorizacija tekstova	`Corpus`, `Dictionary`, `Text2Tokens`, `CorpusIndex`
`InverseTokenizer`	inverzija normalizacije riječi	`Corpus`, `Text2Tokens`
`TopicModel`	tematski model	`Corpus`, `Dictionary`, `Text2Tokens`, `BowCorpus`
`CorpusTopicIndex`	index tema-dokument veza	`Corpus`, `Dictionary`, `Text2Tokens`, `TopicModel`, `BowCorpus`, `CorpusIndex`

i taj kontekst konstruira se pozivom metode `basicBuildersContext(cacheFolder)`.

Tablica 7.1 sadrži primjere komponenti srednje razine. Za svaku komponentu mogu se vidjeti druge komponente na kojima se komponenta temelji. Pri tome se razlikuju podkomponente koje definiraju samu komponentu i pomoćne komponente koje se koriste pri izgradnji ali se i same mogu izvesti iz podkomponenti. Primjerice, `BowCorpus` ovisi o zbirci tekstova, riječniku koji definira indekse riječi i o normalizacijskoj komponenti koja razbija tekstove na pojavnice. `CorpusIndex` je pri tome pomoćna komponenta koja služi za indeksiranje tekstova u zbirci ali ne definira strukturu od `BowCorpus`. Iz tablice se može vidjeti da sve komponente koje predstavljaju neku reprezentaciju zbirke tekstova koriste `CorpusIndex`. Ova važna komponenta gradi fiksno mapiranje sa skupa identifikatora tekstova u zbirci na skup prirodnih brojeva i osigurava da različite komponente koriste kompatibilne indekse tekstova iste zbirke.

Komponente aplikacijske razine Ovaj skup sadrži komponente koje se koriste prije svega u primjenama poput eksperimenata s tematskim modelima, a ne za izgradnju još složenijih komponenti. Ove komponente povezane su s tematskim modeliranjem i obuhvaćaju same tematske modele, komponentu za indeksiranje tekstova u zbirci temama modela, funkcionalnost za vrednovanje tematskih modela i funkcionalnost za grupiranje tema modela.

Tematski modeli sastoje se od klasa koje adaptiraju implementacije raznih tematskih modela na standardno sučelje `TopicModel` okvira *pytopia*. Ovi adapteri omogućuju upotrebu vanjskih implementacija modela prema konvencijama pytopia okvira. Preciznije, omogućuju njihovu izgradnju, pohranu, učitavanje i identifikabilnost, te pristup njihovim podacima modela kroz sučelje `TopicModel`. Trenutno su implementirani adapteri za neke od modela iz paketa scikit-learn[3], hca[4], ARTM[5] i gensim[6]. Treba spomenuti i "umjetni" tematski model implementiran klasom `ArtifTopicModel`, koji omogućava definiciju modela s proizvoljnim tema-riječ i dokument-tema matricama. Ova klasa korištena je za izgradnju modela koji odgovaraju skupovima referentnih tema (odjeljak 5.4), što je kod eksperimenata pokrivenosti omogućilo isto sučelje za oba tipa tema – teme modela i referentne teme.

Komponenta `CorpusTopicIndex` koja implementira dokument-tema indeks definira sučelje za dohvat težina tema za dokumente i težina dokumenata za teme. Ova komponenta gradi se na temelju zbirke tekstova i tematskog modela i koristi se primjerice za dohvat najbolje rangiranih dokumenata za temu. Tablica 7.1 pokazuje kako `TopicModel` i `CorpusTopicIndex` ovise o drugim komponentama niže razine.

Funkcionalnost vrednovanja tematskih modela obuhvaća funkcionalnost za računanje raznih mjera koherentnosti tema te mjere koje ocjenjuju u kojoj je mjeri tema modela spoj dvije semantičke teme (engl. *fused topic*). Funkcionalnost za grupiranje tema omogućava primjenu standardnih algoritama grupiranja na grupiranje tema modela te definira klasu koja adaptira grupirane teme na `TopicModel` sučelje. Okvir sadrži i niz funkcija koje implementiraju razne mjere udaljenosti i sličnosti vektora i primjenjuju se za računanje udaljenosti tema preko udaljenosti pripadnih vektora riječi i dokumenata.

7.2.4 Potporna funkcionalnost

Potporna funkcionalnost obuhvaća bilježenje (engl. *logging*) i testiranje. Bilježenje događaja i grešaka do kojih dolazi prilikom rada s okvirom nužno je zbog velike količine operacija koje se događaju implicitno i nevidljive su korisniku okvira. Uvidom u log-datoteke (engl. *log file*) korisnik može rekonstruirati točan tijek složenih operacija izgradnje te vidjeti zabilježene poruke o greškama.

Testiranje softverskih komponenti standardna je praksa softverskog inženjerstva čiji cilj je osigurati kvalitetu koda i ubrzati otkrivanje grešaka. Testiranje se provodi pisanjem testnih metoda koje simuliraju primjenu komponenti i bilježe eventualne greške. Testne metode omogućuju brzu detekciju grešaka nakon promjena u kodu. Pytopia okvir definira dvije vrste funkcionalnosti za potporu testiranju *pytopia* komponenti – standardne testove te temeljne resurse

[3] `https://scikit-learn.org`
[4] `https://mloss.org/software/view/527/`
[5] `https://github.com/bigartm`
[6] `https://radimrehurek.com/gensim`

izvođenje testova. Primjer standardnog testa je izgradi-spremi-učitaj-usporedi test koji se može brzo primijeniti na svaku komponentu koja poštuje pravila okvira kako bi se provjerilo da je spremljena i učitana komponenta jednaka originalno izgrađenoj, te da su pri tome sve operacije prošle bez grešaka. Primjer resursa za testiranje su male umjetne zbirke tekstova za brzu izgradnju modela i drugih komponenti. Za standardne komponente srednje i aplikacijske razine (primjeri u tablici 7.1) implementirani su testovi kompatibilni sa `pytest` [7] okvirom.

7.3 Primjene radnog okvira

Važna primjena radnog okvira *pytopia*, koja je motivirala njegov razvoj, je analiza i vrednovanje tematskih modela. Primjeri ovakvih eksperimenata su vrednovanja koherentnosti tema opisana u [18, 22] i u poglavlju 4, vrednovanja stabilnosti modela [64], te eksperimenti pokrivenosti opisani u poglavlju 5. Što se više varijanti modela, parametara izgradnje i zbirki tekstova varira u ovakvim eksperimentima dobiveni rezultati otkrivaju više znanja o tematskim modelima. Međutim, provođenje takvih eksperimenata je tehnički zahtjevno i okvir *pytopia* uvelike olakšava njihovu provedbu definicijom standardnih sučelja za komponente korištene u tematskom modeliranju i nizom funkcionalnosti za podršku izgradnji, pohrani, dohvatu i upravljanju većim brojem modela.

U ovom odjeljku opisuju se primjene okvira na eksperimente s mjerama dokumentne koherentnosti iz poglavlja 4 i eksperimente pokrivenosti iz poglavlja 5. Zatim se daje kratak pregled potencijalnih primjena van područja analize tematskih modela. Primjene na postupak analize medijske agende iz poglavlja 3 opisane su u uvodu.

Eksperimenti s mjerama koherentnosti Početni korak implementacije ovih eksperimenata bio je organizacija temeljnih resursa u kontekste. Eksperimenti su provedeni na temelju podataka izgrađenih u eksperimentima analize agende američkih medija (odjeljak 3.4) i hrvatskih medija (odjeljak 3.5). Ovi podaci obuhvaćaju zbirke tekstova, riječnike i izgrađene tematske modele, te komponente za normalizaciju tekstova. Svi ovi resursi i komponente prvo su adaptirani na standardna *pytopia* sučelja. Primjerice, spremljeni modeli bili su izgrađeni radnim okvirom *gensim* pa je implementirana klasa koja adaptira te modele na sučelje `TopicModel`. Adaptirani resursi zatim su organizirani u kontekste.

Ključni element ovih eksperimenata su mjere koherentnosti. Ove mjere implementirane su kao pytopia komponente odnosno klase koje podržavaju funkcijske pozive i identifikabilnost. Ove klase u pravilu su i hijerarhijski kompozicionalne te koriste mogućnost automatske izgradnje identifikatora pomoću klase `IdComposer`. Mjere koherentnosti koriste komponente srednje razine okvira *pytopia* (odjeljak 7.2.3) za izgradnju vektora dokumenata (odjeljak 4.4.2).

[7] `https://docs.pytest.org`

Primjerice, za tf-idf vektorizaciju koristi se komponenta `CorpusTfidf`.

Pri izgradnji vektorskih reprezentacija mjere koherentnosti koriste standardna sučelja okvira za automatsku izgradnju resursa srednje razine poput `CorpusTfidf` komponente. Preciznije, argument mjera koherentnosti je tema odnosno komponenta `Topic` koja sadrži identifikator tematskog modela koji pak sadrži podatke o temeljnim komponentama modela poput zbirke tekstova i riječnika. Na temelju ovih podataka grade se resursi srednje razine koji su kompatibilni sa tematskim modelom odnosno koriste iste temeljne resurse. Okvir 7.10 ilustrira izgradnju resursa na opisani način. Pri tome se izgrađeni resursi ne grade iznova pri svakom pozivu pošto se koriste standardni graditelji definirani na razini projekta koji pohranjuju izgrađene komponente u priručnu memoriju na disku.

Okvir 7.10: Automatska izgradnja resursa srednje razine na temelju teme modela.

```
class MjeraKoherentnosti():
  def __call__(topic):
    model = resolve(topic.model)
    graditelj = resolve('corpus_tfidf_builder')
    tfidf = graditelj(model.corpus, model.dictionary,
                      model.text2tokens)
```

Opisani način primjene načela okvira *pytopia* dovodi do vrlo jednostavne primjene mjera koherentnosti na nove skupove podataka. Naime, sve što je potrebno je definirati kontekst s temeljnim resursima i modelima koji se vrednuju, a same funkcije će automatski izgraditi sve resurse potrebne za vektorizaciju dokumenata, poput tf-idf matrica i matrica vjerojatnosti riječi. Primjerice, implementirane mjere razvijene su na američkom skupu podataka, a eksperimenti s hrvatskim podacima provedeni su na opisani način uz minimalan utrošak vremena.

Od ostalih *pytopia* funkcionalnosti koristi se priručna memorija (engl. *cache*) za pohranu izračunatih ocjena koherentnosti. Preciznije, koristi se komponenta `CachedFunction` koja služi kao omotač (engl. *wrapper*) bilo koje identifikabilne funkcije i pohranjuje rezultate poziva te funkcije na disk. Ova priručna memorija znatno ubrzava opetovano računanje funkcija koherentnosti nužno prilikom njihovog razvoja i vrednovanja.

Eksperimenti pokrivenosti Eksperimenti pokrivenosti (poglavlje 5) provedeni su na dva različita skupa podataka, medijskom i biološkom, pri čemu su svi temeljni resursi svakog od skupova organizirani u *pytopia* kontekste.

U ovom eksperimentu vrednuju se tematski modeli iz različitih klasa. Implementacije ovih modela, opisanih u odjeljku 5.5, adaptirane su na standardno `TopicModel` sučelje. To uključuje i skupove referentnih tema, koji su predstavljeni komponentom `ArtifTopicModel` (umjetni tematski model) koja također poštuje `TopicModel` sučelje. Ovo je omogućilo uniforman način

rada sa svim skupovima tema – temama modela i referentnim temama.

Najveći izazov ovog eksperimenta bila je izgradnja velikog broja tematskih modela. Izgrađeno je ukupno 200 modela, 100 modela po skupu podataka. Ovih 100 modela dolazi iz različith klasa modela i izgrađeno je različitim parametrima (tablica 5.2). Izgradnja modela temelji se na sučeljima izgradnje i pohrane komponenti opisanim u odjeljku 7.2. Za svaku klasu modela definirani su skupovi parametara nastali variranjem hiperparametara i parametara incijalizacije a graditelji modela su automatski pozivani s tim parametrima. Graditelji su konfigurirani na način da pohranjuju izgrađene modele u priručnu memoriju na disku. Sami modeli su izgrađeni na serveru korištenjem Docker[8] platforme – kod za izgradnju, cijeli Python ekosustav i operacijski sustav zapakirani su u Docker kontejner koji je potom pokrenut na serverskom stroju. Direktoriji s priručnom memorijom potom su preneseni na radnu stanicu gdje su graditelji modela konfigurirani tako da umjesto izgradnje učitavaju već izgrađene modele iz priručne memorije.

Važna komponenta prilikom eksperimentiranja bila je, kao i u slučaju eksperimenata koherentnosti, klasa `CachedFunction` koja je omogućila pohranu rezultata mjera pokrivenosti i vezanih funkcija i znatno ubrzala njihovo opetovano izvršavanje. Ove funkcije, implementirane kao identifikabilne pytopia komponente, obuhvaćaju nadzirane mjere pokrivenosti (odjeljak 5.6), klasifikatore za računanje poklapanja tema (odjeljak 5.6.3) te PPU mjere pokrivenosti (odjeljak 5.7).

Ostale potencijalne primjene Iako je upotreba više klasa tematskih modela sa standardnim sučeljem i funkcija za vrednovanje modela potencijalno korisna u svakoj primjeni tematskih modela, pytopia radni okvir je najkorisniji kod onih primjena gdje je potrebna izgradnja i pohrana većeg broja tematskih modela s različitim parametrima.

Jedan primjer ovakve primjene je sustav koji omogućava izgradnju većeg broja tematskih modela od strane analitičara velikih baza medijskih tekstova, poput platformi za agregaciju medijskog sadržaja opisanih u odjeljku 3.1.2. Ovi modeli mogli bi se koristiti za eksplorativnu analizu teksta pri čemu bi se najkorisniji modeli s informativnim temama mogli pohranjivati i dijeliti među korisnicima sustava. Za ovakve sustave potencijalno je korisna i automatska izgradnja modela za tematsku analizu svih tekstova iz određenih vremenskih intervala (tjedana, mjeseci, godina) ili modela za analizu tekstova filtriranih po ključnim riječima ili izvoru vijesti.

Okvir *pytopia* olakšava dijeljenje resursa vezanih uz tematsko modeliranje pošto potiče organizaciju resursa u kontekste te olašava pohranu i učitavanje izgrađenih modela. Na ovaj način moguće je grupirati skup srodnih resursa i modela u Python paket, koji se može iskoristiti za dijeljenje podataka koji definiraju znanstveni eksperiment s tematskim modelima.

Standardna sučelja koja opisuju komponente vezane uz tematsko modeliranje, poput klasa `TopicModel` i `CorpusTopicIndex`, mogla bi olakšati razvoj algoritama koji agregiraju teme

[8] `https://www.docker.com/`

većeg broja modela.

7.4 Ostala programska podrška

U ovom odjeljku opisuje se programska podrška za dva važna aspekta računalne analize medijske agende: izradu tekstne zbirke medijskih tekstova koja predstavlja temelj za analizu, te za korisničko sučelje koje može znatno olakšati i ubrzati postupak analize. Oba aspekta povezana su s radnim okvirom *pytopia*– okvir olakšava pristup tekstovima zbirke, a grafičko sučelje upotrebom okvira može postići lakše i fleksibilnije upravljanje tematskim modelima.

Preciznije, radni okvir sadrži komponentu kompatibilnu sa sučeljem zbirke tekstova (klasa `Corpus`) koja je namijenjena definiciji zbirke na temelju tekstova i njihovih metapodataka pohranjenih u bazi podataka koju gradi u nastavku opisana aplikacija za sakupljanje medijskih tekstova. Grafičko korisničko sučelje koje koristi radni okvir može iskoristiti sve funkcionalnosti okvira za izgradnju i upravljanje resursima te pristupati modelima putem standardnih sučelja, čime se postiže fleksibilnost kod odabira tematskih modela i ostalih komponenti koje se koriste u analizi agende.

7.4.1 Aplikacija za sakupljanje medijskih tekstova

Zbirka medijskih tekstova predstavlja temelj za računalnu analizu medijske agende. U ovom odjeljku opisana je aplikacija *feedsucker* namijenjena izgradnji zbirke medijskih tekstova. Ova aplikacija sakuplja tekstove vijesti iz niza izvora i pohranjuje ih u bazu podataka. *Izvor* (engl. *feed*) je lista nedavno objavljenih vijesti dostupna putem URL adrese. Izvor može biti strukturiran u nekom od standardnih formata poput RSS formata[9] ili može biti web stranica koja sadrži poveznice na vijesti, poput stranice web portala s popisom vijesti iz neke kategorije. Korisnik aplikacije *feedsucker* definira niz izvora putem tekstne datoteke u standardnom formatu i pokreće aplikaciju koja kontinuirano pristupa poveznicama sadržanim u izvorima, i za novoobjavljene vijesti koje još nisu u bazi ekstrahira tekstove i pohranjuje ih u bazu podataka zajedno s metapodacima poput vremena objave.

Motivacija za razvoj ove aplikacije bila je potreba za alatom koji omogućava fleksibilnost pri definciji tekstne zbirke odnosno izradu zbirke točno onih medijskih tekstova koji su od interesa za istraživanje. Ovaj zahtjev podrazumijeva mogućnost praćenja izvora po izboru, neovisno o jeziku ili tematskoj kategoriji.

Jedno moguće rješenje ovog problema su platforme koje sakupljaju medijske tektove iz niza izvora, poput MediaCloud platforme[10] i ProQuest baze medijskih tekstova[11]. Problem kod ovih

[9] `http://www.rssboard.org/rss-specification`
[10] `https://mediacloud.org`
[11] `https://www.proquest.com/libraries/academic/news-newspapers`

rješenja je nemogućnost korisnika da utječe na izvore koji se prate, što smanjuje fleksibilnost pri definiciji zbirke ili čak onemogućava izradu zbirke pošto ovakve platforme u pravilu sakupljaju engleske tekstove. Drugi problem je nemogućnost pristupa cjelovitim tekstovima zbog autorskih prava. S druge strane, komparativna prednost velikih baza je mogućnost pristupa povijesnim podacima – aplikacija *feedsucker* sakuplja samo tekstove objavljene od trenutka pokretanja aplikacije.

Drugo moguće rješenje su aplikacije za iscrpni obilazak web sadržaja (engl. *crawling*) poput alata Scrapy[12]. Međutim, ovakvi alati dohvaćaju veliku količinu URL adresa iscrpnim obilaskom web grafa i te adrese je zatim potrebno obraditi kako bi se otkrio medijski sadržaj od interesa. Sam iscrpni obilazak potrebno je pokretati periodički kako bi se otkrio novi sadržaj, što u pravilu troši znatno više resursa od praćenja fiksnog skupa izvora. Komparativna prednost aplikacije *feedsucker* je da koristi dobro definirane izvore medijskih tekstova. Međutim, korisna primjena alata za iscrpni obilazak u kontekstu upotrebe aplikacije *feedsucker* je pronalaženje URL adresa izvora (engl. *feed*) iscrpnim obilaskom web portala.

Koliko je autoru poznato, aplikacija *feedsucker* je jedina slobodno dostupna aplikacija[13] koja omogućava izradu medijske zbirke web teskstova i nudi potpunu fleksibilnost pri izboru izvora vijesti.

Aplikacija *feedsucker* razvijena je u programskom jeziku Java, za pohranu tekstova koristi relacijsku bazu PostgreSQL[14], a za ekstrakciju teksta iz web stranica koristi Python alat newspaper[15]. Ovaj alat, a posljedično i aplikacija *feedsucker*, primjenjiv je za velik broj različitih jezika pošto vrši ekstrakciju tekstova samo na temelju liste zaustavnih riječi.

Aritektura aplikacije omogućava praćenje izvora vijesti u proizvoljnom formatu te upotrebu proizvoljne metode ekstrakcije teksta iz web stranica. To se postiže definicijom sučelja `IFeedReader` i `IArticleScraper` koja definiraju čitač izvora vijesti i metodu ekstrakcije teksta. Jezgra aplikacije koja provodi radni tijek sakupljanja i pohrane koristi ova sučelja, tako se uvođenje novih metoda svodi na implementaciju klase metode koja poštuje odgovarajuće sučelje. Dokumentacija sadrži opis arhitekture i načina upotrebe aplikacije.

Aplikacija *feedsucker* korištena je u nekoliko navrata za izradu medijskih zbirki. Primjerice, zbirke američkih i hrvatskih medijskih tekstova nad kojima su provedene analize agende opisane u poglavlju 3 sakupljene su pomoću ove aplikacije. Osim ovih manjih zbirki, aplikacija je u nekoliko navrata konfigurirana i pokrenuta u svrhu kontinuiranog sakupljanja većih zbirki. Te zbirke građene su tijekom razdoblja od nekoliko godina i sadrže od nekoliko stotina tisuća tekstova do preko milijun tekstova.

[12] `https://scrapy.org/`

[13] Kod i dokumentacija aplikacije *feedsucker* dostupni su na `https://github.com/dkorenci/feedsucker`.

[14] `https://www.postgresql.org`

[15] `https://github.com/codelucas/newspaper/`

7.4.2 Grafičko korisničko sučelje za otkrivanje tema

Svaki od koraka i podkoraka postupka analize medijske agende opisanog u odjeljku 3.3 moguće je provesti koristeći razne alate koji variraju prema lakoći korištenja. Na jednom kraju spektra su grafička korisnička sučelja koja automatiziraju korake provođenja koraka analize i omogućuju pregledan prikaz podataka. Na drugom kraju spektra su skripte i druge datoteke s kodom nekog programskog jezika pomoću koji se upravlja postupkom analize, te tekstne datoteke koje se koriste kao sučelje prema ulaznim i izlaznim podacima. Drugi pristup je prihvatljiv samo analitičarima koji znaju programirati, međutim pristup ne zahtijeva znatan utrošak vremena za izradu korisničkog sučelja i omogućava fleksibilnost odnosno proizvoljne modifikacije provođenja postupka. Grafičko korisničko sučelje je pak nužno za dostupnost alata za analizu širem krugu korisnika te očekivano daje mnogo bolju preglednost podataka i može znatno ubrzati izvođenje operacija.

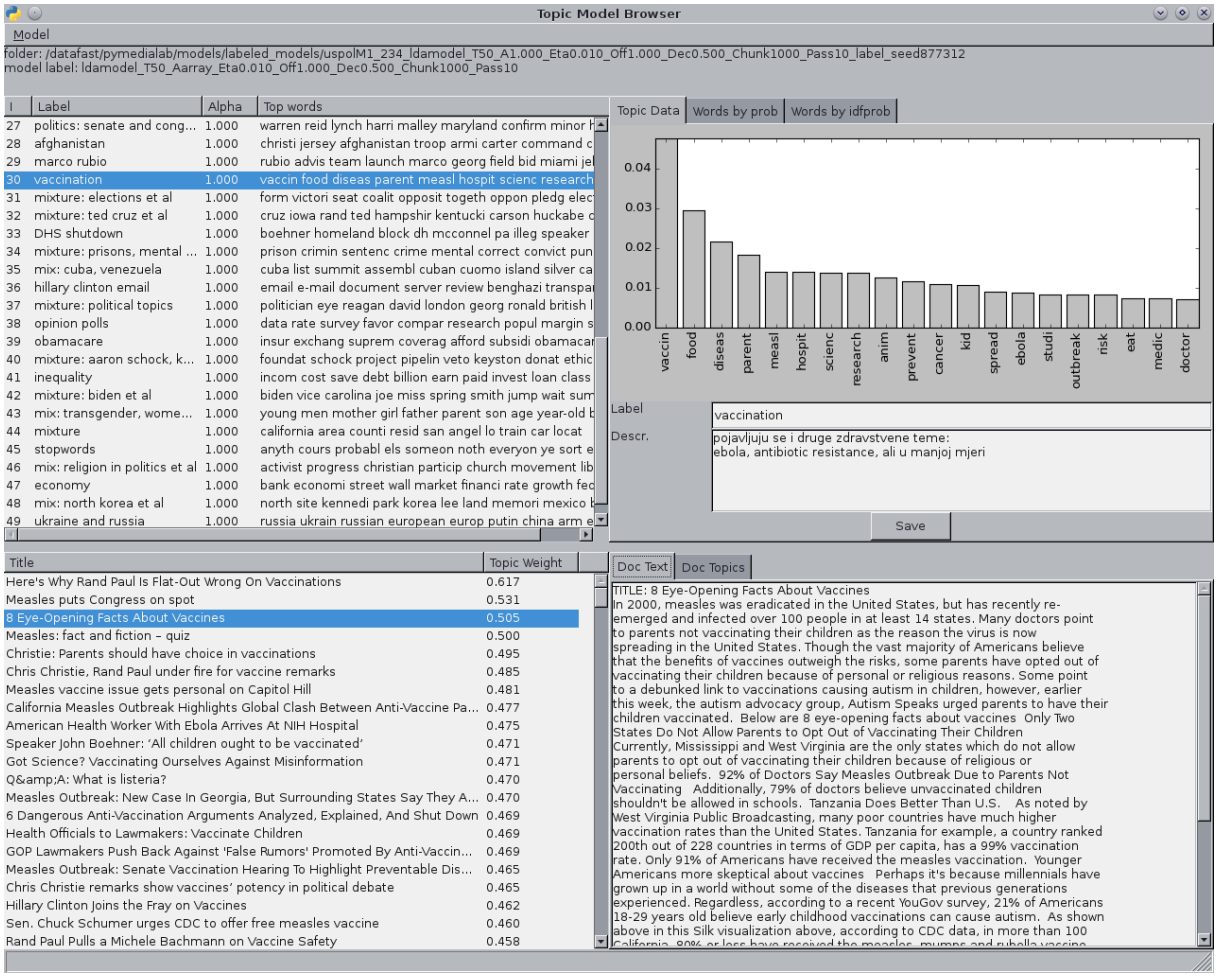
Eksperimenti analize medijske agende, opisani u odjeljcima 3.4 i 3.5, provedeni su kombinacijom oba opisana pristupa. Većina koraka postupka otkrivanja tema (odjeljak 3.3.1) izvedena je pomoću grafičkog korisničkog sučelja, dok je većina koraka mjerenja tema (odjeljak 3.3.3) izvedena putem skriptiranja i tekstnih datoteka, s iznimkom koraka izrade skupova riječi koji opisuju semantičke teme.

Slika 7.1 prikazuje upotrebu grafičkog korisničkog sučelja za pregled i interpretaciju tema modela. Lijevi gornji okvir prikazuje teme modela, njihove oznake i najbolje rangirane riječi. Odabirom pojedine teme podaci koji opisuju temu prikazuju se u gornjem desnom okviru (riječi teme i forma za unošenje oznake i opisa teme), te u doljnjem lijevom okviru (lista uz temu vezanih tekstova označenih naslovima). Odabirom pojedinog dokumenta u doljnjem desnom okviru prikazuje se puni tekst dokumenta.

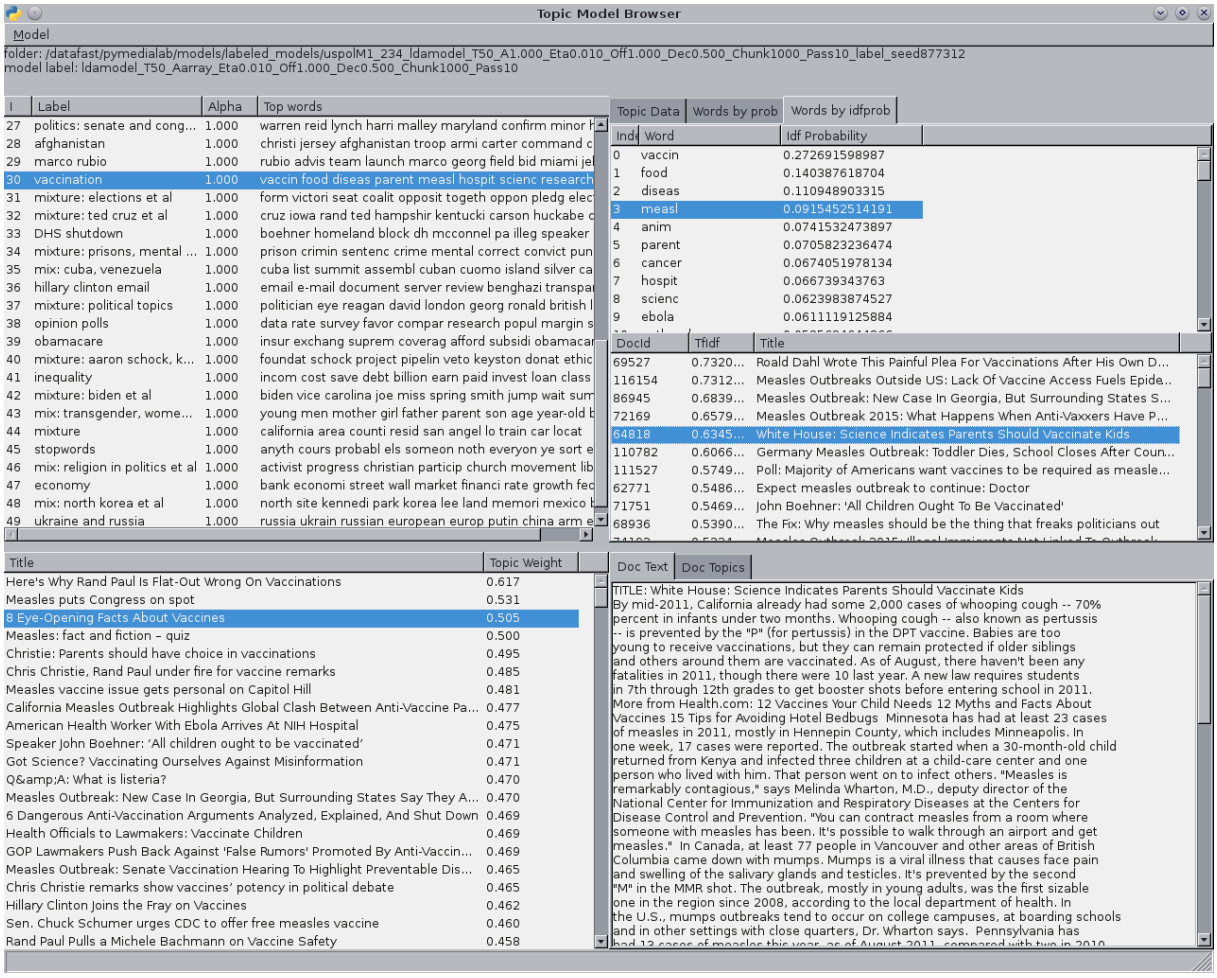
Slika 7.2 prikazuje upotrebu grafičkog korisničkog sučelja za odabir visokodiskriminativnih riječi koje odgovaraju semantičkoj temi. Ove riječi koriste se za inicijalizaciju slabonadziranog tematskog modela koji se koristi za mjerenje zastupljenosti tema (odjeljak 3.3.3). Razlika od prethodnog slučaja je u gornjem desnom okviru koji sadrži listu riječi teme uređenih po heurističkoj mjeri diskriminativnosti. Ovoj listi riječi pristupa se odabirom teme (okvir gore lijevo). Odabirom pojedine riječi u listi dobiva se uvid u listu dokumenata povezanih s tom riječi (povezanost se mjeri tf-idf težinama). Ova lista prikazana je u gornjem desnom okviru, odmah ispod liste riječi.

Opisano grafičko korisničko sučelje implementirano je u jeziku Python pomoću paketa `wx` koji omogućava upotrebu funkcionalnosti radnog okvira `wxWidgets`[16], objektno orijentiranog C++ radnog okvira za brzu izradu grafičkih sučelja sastavljenih od standardnih elemenata poput izbornika, formi, listi, dijaloga i gumbi. Sučelje je razvijeno prije radnog okvira *pytopia* i neki problemi koji su se pojavili pri razvoju, povezani s upravljanjem tematskim modelima,

[16] `https://wxwidgets.org/`



Slika 7.1: Upotreba grafičkog korisničkog sučelja za otkrivanje tema. Prikazan je pregled i obrada teme "cijepljenje".



Slika 7.2: Upotreba grafičkog korisničkog sučelja za odabir visokodiskriminativnih riječi za semantičku temu “cijepljenje”.

utjecali su na oblikovanje djelova okvira *pytopia*– mehanizma izgradnje komponenti te sučelja `TopicModel` i `CorpusTopicIndex`. Nova verzija sučelja trebala bi biti u potpunosti temeljena na okviru *pytopia*, što će prema dosadašnjim iskustiva s implementacijom znatno olakšati razne zadatke upotrebe tematskih modela i omogućiti laganu zamjenu klase modela koja se koristi za analizu.

Poglavlje 8

Zaključak

U ovom radu istraženi su računalni postupci analize medijske agende (engl. *Media Agenda*) koji služe za otkrivanje i mjerenje zastupljenosti tema u velikim zbirkama medijskih tekstova te imaju primjene u nizu znanstvenih i komercijalnih analiza medijskog teksta. Istraženi postupci temeljeni su na tematskim modelima, standardnoj metodi strojnog učenja za analizu tematske strukture teksta. Istraživanje je započelo analizom postojećih metoda analize medija tematskim modelima i na temelju tog istraživanja predložen je postupak koji sistematizira postojeće postupke i otklanja njihove nedostatke. Primjena tog postupka pokazala je potrebu za novim metodama vrednovanja tematskih modela koje bi dovele do povećanja efikasnosti na modelima temeljenih analiza. Iz tog razloga predložene su dvije nove metode vrednovanja tematskih modela – mjere dokumentne koherentnosti tema i metode mjerenja pokrivenosti tema. Ove metode zatim su primijenjene za poboljšanje postupka otkrivanja tema. Razvijen je radni okvir za izgradnju i upravljanje tematskim modelima motiviran problemima uočenim tijekom razvoja programske podrške za analizu agende i vrednovanje modela. Razvijena je i aplikacija za izgradnju medijskih zbirki te grafičko sučelje za potporu postupku analize. U radu je opisan izvorni znanstveni doprinos koji se sastoji od sljedeća četiri dijela.

Prvi dio ostvarenog znanstvenog doprinosa, opisan u poglavljima 3 i 6, je računalni postupak za analizu medijske agende temeljen na tematskim modelima. Ovaj postupak, za razliku od ranije predloženih, provodi otkrivanje tema na temelju većeg broja modela i uvodi važan korak definicije novih tema koji omogućuje preciznu analizu na temelju specifičnih tema od interesa za istraživača. Mjerenje zastupljenosti tema definira se kao problem klasifikacije s višestrukim oznakama (engl. *multilabel classification*), što omogućava precizno kvantitativno vrednovanje mjerenja. Predložena je metoda mjerenja tema temeljena na slabonadziranim tematskim modelima i postupku odabira skupa visokodiskriminativnih riječi koji definira mjerenu temu. Predloženi postupak analize medijske agende vrednovan je kroz primjene na analizu američkih i hrvatskih medija. Ove analize pokazale su da otkrivene teme daju dobar pregled medijske agende, da je mogućnost definicije novih tema na temelju otkrivenih koristan analitički alat,

te da je predložena metoda mjerenja dobar alat za mjerenje zastupljenosti tema. Na temelju razvijenih metoda vrednovanja tematskih modela predložena su daljnja poboljšanja koraka otkrivanja tema s ciljem bržeg otkrivanja većeg broja tema.

Drugi dio ostvarenog znanstvenog doprinosa, opisan u poglavlju 4, je metoda vrednovanja tema dobivenih tematskim modelom temeljena na računanju semantičke povezanosti uz temu vezanih dokumenata. Provedene analize medijske agende pokazale su kako ranije predložene mjere koherentnosti, temeljene na riječima, ne predstavljaju dobar model za medijske teme. S druge strane, pokazalo se kako se te teme mogu dobro interpretirati na temelju dokumenata. Predložene su mjere dokumentne koherentnosti koje vektoriziraju skup uz temu vezanih dokumenata te računaju ocjenu koherentnosti koristeći model vektorskog prostora, vjerojatnosnog prostora ili grafa. Provedeno vrednovanje na skupovima američkih i hrvatskih medijskih tema otkrilo je najbolju na grafu temeljenu mjeru i potvrdilo slabe performanse mjera temeljenih na riječima. Kvalitativna usporedba mjera koherentnosti riječi i mjera koherentnosti dokumenata ukazala je na komplementarnost ovih pristupa i potrebu za njihovim kombiniranjem. Razvijene mjere dokumentne koheretnosti mogu se koristiti ne samo u analizi medijske agende već i u mnogim drugim primjenama tematskih modela za medijski tekst.

Treći dio ostvarenog znanstvenog doprinosa, opisan u poglavlju 5, je metoda vrednovanja pokrivenosti tema koja mjeri poklapanje skupa referentnih tema (koncepata) i skupa tema dobivenih modelom. Vrednovanje ovog važnog aspekta kvalitete tematskih modela slabo je istraženo. Predložena je metoda izrade skupa referentnih tema te dvije metode mjerenja pokrivenosti koje se temelje na računanju semantičkog poklapanja tema. Prva metoda temelji se na izradi klasifikacijskog modela za aproksimaciju ljudskih ocjena poklapanja tema. Vrednovanje je pokazalo da se upotrebom značajki udaljenosti tema i standardnih klasifikacijskih modela može postići vrlo dobra aproksimacija ljudskih ocjena. Druga metoda je nenadzirana automatska metoda koja se temelji na agregaciji ocjena pokrivenosti dobivenih variranjem praga udaljenosti koji određuje poklapanje tema. Vrednovanje ove nenadzirane mjere pokazalo je da odlično aproksimira nadzirane mjere. Metoda računanja ove mjere može se koristiti i za vizualnu analizu i usporedbu pokrivenosti modela. Zbog lakoće primjene na nove zbirke, ova metoda je najvažniji dio doprinosa koji se odnosi na pokrivenost tema.

Četvrti dio ostvarenog znanstvenog doprinosa, opisan u poglavlju 7, sastoji se od radnog okvira za upravljanje tematskim modelima i ostale programske podrške za postupak analize medijske agende. Radni okvir rješava probleme izgradnje, pohrane, dohvata i upotrebe tematskih modela i drugih vezanih komponenti koji su uočeni tijekom razvoja programske podrške za analizu agende. Okvir se temelji na četiri načela oblikovanja: standardnim sučeljima, obaveznoj identifikabilnosti komponenti, dohvatu komponenti iz konteksta te strukturiranju složenih komponenti kao hijerarhija jednostavnijih komponenti. Ova četiri principa motivirana su prirodom komponenti kojima okvir upravlja i standardnim metodama softverskog inženjerstva, a u

kombinaciji rezultiraju općenitim i fleksibilnim pristupm koji znatno olakšava i ubrzava razvoj programske podrške za vrednovanje i primjenu tematskih modela. Ostala programska podrška postupku računalne analize agende sastoji se od grafičkog korisničkog sučelja za otkrivanje tema i od aplikacije za izgradnju zbirki medijskih tekstova.

Znanstveno istraživanje opisano u ovom radu dovelo je do poboljšanog postupka analize medijske agende i do novih metoda vrednovanja tematskih modela. Metode vrednovanja imaju primjene šire od analize medijske agende – mjere dokumentne koherentnosti primjenjive su kod svih analiza medijskog teksta tematskim modelima a metoda mjerenja pokrivenosti tema neovisna je o tekstnoj domeni. Predložene metode vrednovanja predstavljaju nove pristupe koji bi daljnjim razvojem i novim eksperimentima vrednovanja mogli dovesti do novih uvida o tematskim modelima koji su široko korišten alat strojnog učenja. Razvijeni radni okvir za tematsko modeliranje mogao bi se pokazati vrlo korisnim zato što znatno olakšava provođenje opsežnih eksperimenata s tematskim modelima.

Postoje mnogi izgledni smjerovi za nastavak istraživanja opisanih u radu, detaljnije opisani u raspravnim dijelovima odgovarajućih poglavlja. Računalna analiza medijske agende tematskim modelima predstavlja izazovan problem zbog kompleksnosti modelirane domene i tehničke složenosti postupaka. Jedan smjer za daljnje istraživanje je razvoj novih i razmatranje primjena postojećih modela strojnog učenja za ovaj zadatak. Pri tome se ovdje predložene metode mogu primijeniti za vrednovanje i odabir modela. Drugi važan smjer je razvoj kvalitetnih alata s grafičkim korisničkim sučeljem koji bi učinili metode temeljene na strojnom učenju dostupnima širem krugu istraživača medijskog teksta. Ključan doprinos ovakvih alata bilo bi kroz primjene stečeno znanje o modelima i postupcima koji se temelje na modelima. Predloženi radni okvir može znatno olakšati i ubrzati razvoj tih alata.

Predložene metode dokumentne koherentosti mogle bi se adaptirati za nove tekstne domene poput znanstvenih tekstova čija se tematska struktura razlikuje od kraćih i tematski fokusiranih medijskih tekstova. Osim razvoja novih mjera i modela dokumentne koherentnosti, izgledan smjer je razvoj metoda kombiniranja mjera dokumentne koheretnosti i koheretnosti riječi, koje bi mogle dovesti do potpunijeg modela koherentnosti tema.

Problem pokrivenosti tema važan je i neistražen problem u tematskom modeliranju i ovdje opisane metode predstavljaju početne korake u njegovom istraživanju. Postoji niz smjerova za daljnja istraživanja, koji su detaljno opisani u odjeljku 5.9. Neovisno o pristupu, važan je razvoj metoda koje se mogu brzo primijeniti na nove tekstne zbirke iz proizvoljne domene. Što se tiče metoda mjerenja, predložena nenadzirana mjera pokrivenosti mogla bi se pokazati kao zadovoljavajuće rješenje. Međutim, potrebne su metode za brzu izgradnju skupa referentnih tema. Kao dobar pristup tom problemu čini se razvoj na strojnom učenju temeljenih alata i sučelja koja bi omogućila domenskim stručnjacima efikasno definiranje referentnih tema. Poželjno je i provođenje što većeg broja eksperimenata pokrivenosti na raznim tekstnim domenama i s raznim

klasama tematskih modela. Provođenje takvih eksperimenata može se znatno ubrzati i olakšati upotrebom predloženog radnog okvira.

Popis slika

Popis tablica

Literatura

[1] Blei, D. M., Ng, A. Y., Jordan, M. I., "Latent Dirichlet allocation", Journal of Machine Learning Research, Vol. 3, No. Jan, 2003, str. 993–1022.

[2] Chuang, J., Ramage, D., Manning, C., Heer, J., "Interpretation and trust: Designing model-driven visualizations for text analysis", in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012, str. 443–452.

[3] Wei, X., Croft, W. B., "Lda-based document models for ad-hoc retrieval", in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006, str. 178–185.

[4] Chen, M., Jin, X., Shen, D., "Short text classification improved by learning multi-granularity topics", in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 2011, str. 1776–1781.

[5] Boyd-Graber, J. L., Blei, D. M., Zhu, X., "A topic model for word sense disambiguation", in Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, str. 1024–1033.

[6] Lin, C., He, Y., "Joint sentiment/topic model for sentiment analysis", in Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009, str. 375–384.

[7] Griffiths, T. L., Steyvers, M., "Finding scientific topics", Proceedings of the National academy of Sciences, Vol. 101, 2004, str. 5228–5235.

[8] Hoffman, M., Bach, F. R., Blei, D. M., "Online learning for latent Dirichlet allocation", in Advances in Neural Information Processing Systems 23. Curran Associates, Inc., 2010, str. 856–864.

[9] Mimno, D. M., McCallum, A., "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression", in Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, 2008.

[10] Blei, D. M., Lafferty, J. D., “A correlated topic model of science”, The Annals of Applied Statistics, Vol. 1, No. 1, 2007, str. 17-35.

[11] Blei, D. M., “Probabilistic topic models”, Communications of the ACM, Vol. 55, No. 4, 2012, str. 77–84.

[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., “Indexing by latent semantic analysis”, Journal of the American society for information science, Vol. 41, No. 6, 1990, str. 391.

[13] Lee, D. D., Seung, H. S., “Learning the parts of objects by non-negative matrix factorization”, Nature, Vol. 401, No. 6755, 1999, str. 788.

[14] Arora, S., Ge, R., Moitra, A., “Learning topic models–going beyond svd”, in Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on. IEEE, 2012, str. 1–10.

[15] Choo, J., Lee, C., Reddy, C. K., Park, H., “Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization”, IEEE transactions on visualization and computer graphics, Vol. 19, No. 12, 2013, str. 1992–2001.

[16] Greene, D., Cross, J. P., “Unveiling the political agenda of the european parliament plenary: A topical analysis”, in Proceedings of the ACM Web Science Conference. ACM, 2015, str. 2.

[17] Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., Supek, F., “The landscape of microbial phenotypic traits and associated genes”, Nucleic acids research, 2016.

[18] O’Callaghan, D., Greene, D., Carthy, J., Cunningham, P., “An analysis of the coherence of descriptors in topic modeling”, Expert Systems with Applications, Vol. 42, No. 13, 2015, str. 5645–5657.

[19] Wang, Y.-X., Zhang, Y.-J., “Nonnegative matrix factorization: A comprehensive review”, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 6, 2013, str. 1336–1353.

[20] Boyd-Graber, J. L., Blei, D. M., Zhu, J., “Probabalistic walks in semantic hierarchies as a topic model for WSD”, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2007.

[21] Titov, I., McDonald, R. T., “A joint model of text and aspect ratings for sentiment summarization.”, in ACL, 2008, str. 308–316.

[22] Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D., "Exploring topic coherence over many models and many topics", in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012, str. 952–961.

[23] Wallach, H. M., Murray, I., Salakhutdinov, R., Mimno, D., "Evaluation methods for topic models", in Proceedings of the 26th annual international conference on machine learning. ACM, 2009, str. 1105–1112.

[24] Mimno, D., Blei, D., "Bayesian checking for topic models", in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, str. 227–237.

[25] Waal, A. D., Barnard, E., "Evaluating topic models with stability", Pattern Recognition Association of South Africa, 2008.

[26] Koltcov, S., Koltsova, O., Nikolenko, S., "Latent dirichlet allocation: stability and applications to studies of user-generated content", in Proceedings of the 2014 ACM conference on Web science. ACM, 2014, str. 161–165.

[27] Belford, M., Namee, B. M., Greene, D., "Stability of topic modeling via matrix factorization", Expert Systems with Applications, Vol. 91, 2018, str. 159 - 169.

[28] Ramirez, E. H., Brena, R., Magatti, D., Stella, F., "Topic model validation", Neurocomputing, Vol. 76, No. 1, 2012, str. 125–133.

[29] Chuang, J., Gupta, S., Manning, C., Heer, J., "Topic model diagnostics: Assessing domain relevance via topical alignment", in Proceedings of the 30th International Conference on machine learning (ICML-13), 2013, str. 612–620.

[30] AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C., "Topic significance ranking of lda generative models", Machine Learning and Knowledge Discovery in Databases, 2009, str. 67–82.

[31] Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., Blei, D. M., "Reading tea leaves: How humans interpret topic models.", in Proceedings of the 22nd International Conference on Neural Information Processing Systems, 2009, str. 288–296.

[32] Lau, J. H., Newman, D., Baldwin, T., "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.", in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, str. 530–539.

[33] Musat, C., Velcin, J., Trausan-Matu, S., Rizoiu, M.-A., "Improving topic evaluation using conceptual knowledge.", in 22nd International Joint Conference on Artificial Intelligence, Vol. 3, 2011, str. 1866–1871.

[34] Miller, G. A., "WordNet: a lexical database for English", Communications of the ACM, Vol. 38, No. 11, Nov 1995, str. 39–41.

[35] Newman, D., Lau, J. H., Grieser, K., Baldwin, T., "Automatic evaluation of topic coherence", in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010, str. 100–108.

[36] Shaw, D. L., Martin, S. E., "The Function of Mass Media Agenda Setting", Journalism Quarterly, Vol. 69, No. 4, Dec 1992, str. 902–920.

[37] Roberts, M., Wanta, W., Dzwo, T.-H. D., "Agenda setting and issue salience online", Communication Research, Vol. 29, No. 4, 2002, str. 452–465.

[38] Krippendorff, K., Content Analysis: An Introduction to Its Methodology. SAGE Publications, Inc, Apr 2012.

[39] Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., Radev, D. R., "How to analyze political attention with minimal assumptions and costs", American Journal of Political Science, Vol. 54, No. 1, 2010, str. 209–228.

[40] Chuang, J., Wilkerson, J. D., Weiss, R., Tingley, D., Stewart, B. M., Roberts, M. E., Poursabzi-Sangdeh, F., Grimmer, J., Findlater, L., Boyd-Graber, J. *et al.*, "Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations", in Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning, 2014.

[41] Grimmer, J., Stewart, B. M., "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", Political Analysis, Vol. 21, No. 3, 2013, str. 267–297.

[42] McCombs, M. E., Shaw, D. L., "The agenda-setting function of mass media", Public Opinion Quarterly, Vol. 36, No. 2, 1972, str. 176–187.

[43] Scheufele, D. A., "Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication", Mass Communication and Society, Vol. 3, No. 2–3, 2000, str. 297–316.

[44] Bonilla, T., Grimmer, J., "Elevated threat levels and decreased expectations: How democracy handles terrorist threats", Poetics, Vol. 41, No. 6, Dec 2013, str. 650–669.

[45] Papadouka, M. E., Evangelopoulos, N., Ignatow, G., "Agenda setting and active audiences in online coverage of human trafficking", Information, Communication & Society, Vol. 19, No. 5, 2016, str. 655–672.

[46] Entman, R. M., "Framing: Toward clarification of a fractured paradigm", Journal of Communication, Vol. 43, No. 4, 1993, str. 51-58.

[47] Pan, Z., Kosicki, G. M., "Framing analysis: An approach to news discourse", Political communication, Vol. 10, No. 1, 1993, str. 55–75.

[48] DiMaggio, P., Nag, M., Blei, D., "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding", Poetics, Vol. 41, No. 6, 2013, str. 570 - 606, topic Models and the Cultural Sciences.

[49] Gilardi, F., Shipan, C. R., Wueest, B., "Policy diffusion: The issue-definition stage", 2018, dostupno na: https://fabriziogilardi.org/resources/papers/diffusion-policy-perceptions.pdf

[50] Evans, M. S., "A Computational Approach to Qualitative Analysis in Large Textual Datasets", PLoS One, Vol. 9, No. 2, Feb 2014.

[51] Yang, T.-I., Torget, A. J., Mihalcea, R., "Topic modeling on historical newspapers", in LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, Jun 2011.

[52] Riffe, Dr. D., Ellis, B., Rogers, M. K., Van Ommeren, Dr. R. L., Woodman, Dr. K. A., "Gatekeeping and the Network News Mix", Journalism Quarterly, Vol. 63, No. 2, Jun 1986, str. 315–321.

[53] Steinberger, R., Pouliquen, B., der Goot, E. V., "An introduction to the europe media monitor family of applications", CoRR, Vol. abs/1309.5290, 2013, dostupno na: http://arxiv.org/abs/1309.5290

[54] Chuang, J., Fish, S., Larochelle, D., Li, W. P., Weiss, R., "Large-scale topical analysis of multiple online news sources with media cloud", NewsKDD: Data Science for News Publishing, at KDD, 2014.

[55] Devezas, T., Nunes, S., Rodríguez, M. T., MediaViz: An Interactive Visualization Platform for Online Media Studies. ACM, Sep 2015.

[56] Jacobi, C., van Atteveldt, W., Welbers, K., "Quantitative analysis of large amounts of journalistic texts using topic modelling", Digital Journalism, Vol. 4, No. 1, 2016, str. 89–106.

[57] Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., Rand, D. G., "Structural Topic Models for Open-Ended Survey Responses", American Journal of Political Science, Vol. 58, No. 4, Mar 2014, str. 1064–1082.

[58] Rader, E., Wash, R., "Identifying patterns in informal sources of security information", Journal of Cybersecurity, Vol. 1, No. 1, Dec 2015, str. 121–144.

[59] Puschmann, C., Scheffler, T., "Topic Modeling for Media and Communication Research: A Short Primer", dostupno na: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2836478 Aug 2016.

[60] Kim, Y., Kim, S., Jaimes, A., Oh, A., "A computational analysis of agenda setting", in Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. ACM, 2014, str. 323–324.

[61] Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., Blei, D. M., "Hierarchical topic models and the nested Chinese restaurant process", in Advances in Neural Information Processing Systems 16. MIT Press, 2004, str. 17–24.

[62] Kok, Y. H., Goh, A., Holaday, D., "Agenda: A tool for agenda setting research", Information Services & Use, Vol. 19, No. 4, Jan 1999, str. 265–276.

[63] Bai, Q., Wei, K., Chen, M., Hu, Q., He, L., "Mining Temporal Discriminant Frames via Joint Matrix Factorization: A Case Study of Illegal Immigration in the U.S. News Media", in Knowledge Science, Engineering and Management. Springer International Publishing, 2018, str. 260–267.

[64] Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., Heer, J., "TopicCheck: Interactive alignment for assessing topic model stability.", in Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, 2015, str. 175–184.

[65] Roberts, M., Stewart, B., Tingley, D., "Navigating the local modes of big data: The case of topic models", in Computational Social Science. Cambridge University Press, New York, 2016, str. 51–97.

[66] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T. *et al.*, "Applying lda topic modeling in communication research: Toward a valid and reliable methodology", Communication Methods and Measures, Vol. 12, No. 2-3, 2018, str. 93–118.

[67] Koltsova, O., Koltcov, S., "Mapping the public agenda with topic modeling: The case of the Russian livejournal", Policy & Internet, Vol. 5, No. 2, Jul 2013, str. 207–227.

[68] Grimmer, J., "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases", Political Analysis, Vol. 18, No. 1, 2010, str. 1–35.

[69] Boyd-Graber, J., Mimno, D., Newman, D., "Care and feeding of topic models: Problems, diagnostics, and improvementes", in Handbook of Mixed Membership Models and their Applications. CRC Press, 2014, str. 225–254.

[70] Gibaja, E., Ventura, S., "A Tutorial on Multilabel Learning", ACM Computing Surveys, Vol. 47, No. 3, Apr 2015, str. 1–38.

[71] Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S., "An extensive experimental comparison of methods for multi-label learning", Pattern Recognition, Vol. 45, No. 9, Sep 2012, str. 3084–3104.

[72] Chekina, L., Rokach, L., Shapira, B., "Meta-learning for Selecting a Multi-label Classification Algorithm", 2011 IEEE 11th International Conference on Data Mining Workshops, Dec 2011, str. 220–227.

[73] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., "LIBLINEAR: A library for large linear classification", Journal of Machine Learning Research, Vol. 9, 2008, str. 1871–1874.

[74] Druck, G., Mann, G., McCallum, A., "Learning from labeled features using generalized expectation criteria", in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, Jul 2008.

[75] Hingmire, S., Chakraborti, S., "Topic labeled text classification: a weakly supervised approach", in Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval. ACM, Jul 2014.

[76] Chen, X., Xia, Y., Jin, P., Carroll, J., "Dataless text classification with descriptive lda", in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015.

[77] Li, C., Xing, J., Sun, A., Ma, Z., "Effective Document Labeling with Very Few Seed Words: A Topic Model Approach", in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, Oct 2016.

[78] Zha, D., Li, C., "Multi-label Dataless Text Classification with Topic Modeling", Knowledge and Information Systems, Dec 2018.

[79] Korenčić, D., Ristov, S., Šnajder, J., "Getting the agenda right: measuring media agenda using topic models", in Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. ACM, 2015, str. 61–66.

[80] Řehůřek, R., Sojka, P., "Software framework for topic modelling with large corpora", in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, str. 45–50.

[81] Bird, S., "NLTK: The Natural Language Toolkit", in Proceedings of the COLING/ACL Interactive presentation sessions, 2006, str. 69–72.

[82] Sayre, B., Bode, L., Shah, D., Wilcox, D., Shah, C., "Agenda setting in a digital age: Tracking attention to California Proposition 8 in social media, online news and conventional news", Policy & Internet, Vol. 2, No. 2, 2010, str. 7–32.

[83] Weaver, D., Elliott, S. N., "Who sets the agenda for the media?", Journalism Quarterly, Vol. 62, No. 1, 1985, str. 87–94.

[84] Landis, J. R., Koch, G. G., "The measurement of observer agreement for categorical data", Biometrics, Vol. 33, No. 1, 1977, str. 159–174.

[85] Hopkins, D. J., King, G., "A Method of Automated Nonparametric Content Analysis for Social Science", American Journal of Political Science, Vol. 54, No. 1, 2010, str. 229–247.

[86] Takeshita, T., "Agenda-setting effects of the press in a Japanese local election", Studies of Broadcasting, Vol. 29, 1993, str. 193–216.

[87] Korenčić, D., Grbeša-Zenzerović, M., Šnajder, J., "Topics and their salience in the 2015 parliamentary election in Croatia: A topic model based analysis of the media agenda", in Proceedings of the International Conference on the Advances in Computational Analysis of Political Text - PolText 2016, 2016.

[88] Ljubešić, N., Boras, D., Kubelka, O., "Retrieving information in Croatian: Building a simple and efficient rule-based stemmer", Digital Information and Heritage, 2007, str. 313–320.

[89] Zeh, R., Hopmann, D. N., "Indicating mediatization? Two decades of election campaign television coverage", European Journal of Communication, Vol. 28, No. 3, 2013, str. 225–240.

[90] Patterson, T. E., "Out of order: How the decline of the political parties and the growing power of the news media undermine the american way of electing presidents", New York: Alfred Knopf, 1993.

[91] Strömbäck, J., Dimitrova, D. V., "Political and media systems matter: A comparison of election news coverage in Sweden and the United States", The Harvard International Journal of Press/Politics, Vol. 11, No. 4, 2006, str. 131–147.

[92] Röder, M., Both, A., Hinneburg, A., "Exploring the space of topic coherence measures", in Proceedings of the eighth ACM international conference on Web search and data mining. ACM, 2015, str. 399–408.

[93] Korenčić, D., Ristov, S., Šnajder, J., "Document-based topic coherence measures for news media text", Expert Systems with Applications, Vol. 114, 2018, str. 357–373.

[94] Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., Kleis Nielsen, R., "Reuters institute digital news report 2017", 2016.

[95] Steinberger, R., Pouliquen, B., der Goot, E. V., "An introduction to the europe media monitor family of applications", CoRR, Vol. abs/1309.5290, 2013, dostupno na: http://arxiv.org/abs/1309.5290

[96] Vossen, P., Rigau, G., Serafini, L., Stouten, P., Irving, F., Van Hage, W. R., "NewsReader: recording history from daily news streams", in LREC, 2014, str. 2000–2007.

[97] Li, L., Wang, D.-D., Zhu, S.-Z., Li, T., "Personalized news recommendation: a review and an experimental investigation", Journal of computer science and technology, Vol. 26, No. 5, 2011, str. 754–766.

[98] Clerwall, C., "Enter the robot journalist: Users' perceptions of automated content", Journalism Practice, Vol. 8, No. 5, 2014, str. 519–531.

[99] Popescu, O., Strapparava, C., (ur.), Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, dostupno na: http://www.aclweb.org/anthology/W17-42

[100] Entman, R. M., "Framing: Toward clarification of a fractured paradigm", Journal of communication, Vol. 43, No. 4, 1993, str. 51–58.

[101] Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A. J., Teo, C. H., "Unified analysis of streaming news", in Proceedings of the 20th international conference on World wide web. ACM, 2011, str. 267–276.

[102] Newman, D., Chemudugunta, C., Smyth, P., Steyvers, M., "Analyzing entities and topics in news articles using statistical topic models", in Intelligence and Security Informatics. Springer Berlin Heidelberg, 2006, str. 93–104.

[103] Dou, W., Wang, X., Skau, D., Ribarsky, W., Zhou, M. X., "Leadline: Interactive visual analysis of text data through event identification and exploration", in Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012, str. 93–102.

[104] Kucher, K., Kerren, A., "Text visualization techniques: Taxonomy, visual survey, and community insights", in 2015 IEEE Pacific Visualization Symposium (PacificVis), 2015, str. 117-121.

[105] Garcin, F., Dimitrakakis, C., Faltings, B., "Personalized news recommendation with context trees", in Proceedings of the 7th ACM conference on Recommender systems. ACM, 2013, str. 105–112.

[106] Li, L., Zheng, L., Yang, F., Li, T., "Modeling and broadening temporal user interest in personalized news recommendation", Expert Systems with Applications, Vol. 41, No. 7, 2014, str. 3168–3177.

[107] Gao, W., Li, P., Darwish, K., "Joint topic modeling for event summarization across news and social media streams", in Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012, str. 1173–1182.

[108] Shahaf, D., Guestrin, C., "Connecting two (or less) dots: Discovering structure in news articles", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 5, No. 4, 2012, str. 24.

[109] Yi, X., Allan, J., "A comparative study of utilizing topic models for information retrieval", in Advances in Information Retrieval. Springer Berlin Heidelberg, 2009, str. 29–41.

[110] Aletras, N., Stevenson, M., "Evaluating topic coherence using distributional semantics", in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013), 2013, str. 13–22.

[111] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A., "Optimizing semantic coherence in topic models", in Proceedings of the Conference on Empirical Met-

hods in Natural Language Processing. Association for Computational Linguistics, 2011, str. 262–272.

[112] Nikolenko, S. I., Koltcov, S., Koltsova, O., "Topic modelling for qualitative studies", Journal of Information Science, Vol. 43, No. 1, 2015, str. 88-102.

[113] Nikolenko, S. I., "Topic quality metrics based on distributed word representations", in Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2016, str. 1029–1032.

[114] Rosner, F., Hinneburg, A., Röder, M., Nettling, M., Both, A., "Evaluating topic coherence measures", CoRR, Vol. abs/1403.6397, 2014, dostupno na: http://arxiv.org/abs/1403.6397

[115] Ramrakhiyani, N., Pawar, S., Hingmire, S., Palshikar, G. K., "Measuring topic coherence through optimal word buckets", in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. ACL, 2017, str. 437–442.

[116] Schütze, H., Manning, C. D., Raghavan, P., Introduction to information retrieval. Cambridge University Press, 2008, Vol. 39.

[117] Mikolov, T., Chen, K., Corrado, G., Dean, J., "Efficient estimation of word representations in vector space", CoRR, Vol. abs/1301.3781, 2013, dostupno na: http://arxiv.org/abs/1301.3781

[118] Pennington, J., Socher, R., Manning, C. D., "Glove: Global vectors for word representation.", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Vol. 14, 2014, str. 1532–1543.

[119] Zhang, R., Guo, J., Lan, Y., Xu, J., Cheng, X., "Aggregating neural word embeddings for document representation", in Advances in Information Retrieval. Springer International Publishing, 2018, str. 303–315.

[120] Galke, L., Saleh, A., Scherp, A., "Word embeddings for practical information retrieval", in INFORMATIK 2017. Gesellschaft für Informatik, Bonn, 2017, str. 2155–2167.

[121] Salton, G., Buckley, C., "Term-weighting approaches in automatic text retrieval", Information processing & management, Vol. 24, No. 5, 1988, str. 513–523.

[122] Turian, J., Ratinov, L., Bengio, Y., "Word representations: a simple and general method for semi-supervised learning", in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010, str. 384–394.

[123] Ljubešić, N., Erjavec, T., "hrWaC and slWac: Compiling web corpora for Croatian and Slovene", in Text, Speech and Dialogue - 14th International Conference, ser. Lecture Notes in Computer Science. Springer, 2011, str. 395–402.

[124] Schult, D. A., Swart, P., "Exploring network structure, dynamics, and function using networkx", in Proceedings of the 7th Python in Science Conferences (SciPy 2008), Vol. 2008, 2008, str. 11–16.

[125] Freeman, L. C., "Centrality in social networks conceptual clarification", Social networks, Vol. 1, No. 3, 1978, str. 215–239.

[126] Estrada, E., Rodriguez-Velazquez, J. A., "Subgraph centrality in complex networks", Physical Review E, Vol. 71, No. 5, 2005, str. 056103.

[127] Ling, C. X., Huang, J., Zhang, H., "Auc: a statistically consistent and more discriminating measure than accuracy", in Proceedings of the 18th International Joint Conference on Artificial intelligence, Vol. 3, 2003, str. 519–524.

[128] Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., "The role of domain information in word sense disambiguation", Natural Language Engineering, Vol. 8, No. 4, 2002, str. 359–373.

[129] DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L., "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", Biometrics, 1988, str. 837–845.

[130] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., "pROC: an open-source package for r and s+ to analyze and compare roc curves", BMC bioinformatics, Vol. 12, No. 1, 2011, str. 77.

[131] Fitelson, B., "A probabilistic theory of coherence", Analysis, Vol. 63, No. 279, 2003, str. 194–199.

[132] Lau, J. H., Baldwin, T., "An empirical evaluation of doc2vec with practical insights into document embedding generation", CoRR, Vol. abs/1607.05368, 2016, dostupno na: http://arxiv.org/abs/1607.05368

[133] Shi, H., Gerlach, M., Diersen, I., Downey, D., Amaral, L. A. N., "A new evaluation framework for topic modeling algorithms based on synthetic corpora", arXiv, Jan 2019, dostupno na: https://arxiv.org/abs/1901.09848

[134] Saha, B., Getoor, L., “On maximum coverage in the streaming model & application to multi-topic blog-watch”, in Proceedings of the 2009 siam international conference on data mining. SIAM, 2009, str. 697–708.

[135] Yue, Y., Joachims, T., “Predicting diverse subsets using structural svms”, in Proceedings of the 25th International Conference on Machine Learning, ser. ICML ’08. New York, NY, USA: ACM, 2008, str. 1224–1231.

[136] Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., “Large Margin Methods for Structured and Interdependent Output Variables”, Journal of Machine Learning Research, Vol. 6, No. Sep, 2005, str. 1453–1484.

[137] Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., Ma, W.-Y., “Improving web search results using affinity graph”, in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’05. New York, NY, USA: ACM, 2005, str. 504–511.

[138] Gambhir, M., Gupta, V., “Recent automatic text summarization techniques: a survey”, Artificial Intelligence Review, Vol. 47, No. 1, Jan 2017, str. 1–66.

[139] Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S., Mehdiyev, C. A., “Mcmr: Maximum coverage and minimum redundant text summarization model”, Expert Systems with Applications, Vol. 38, No. 12, 2011, str. 14 514 - 14 522.

[140] Takamura, H., Okumura, M., “Text summarization model based on maximum coverage problem and its variant”, in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009, str. 781–789.

[141] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X., “Comparing twitter and traditional media using topic models”, in European conference on information retrieval. Springer, 2011, str. 338–349.

[142] Hulpus, I., Hayes, C., Karnstedt, M., Greene, D., “Unsupervised graph-based topic labelling using dbpedia”, in Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013, str. 465–474.

[143] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., “DBpedia: A Nucleus for a Web of Open Data”, in The Semantic Web. Springer Berlin Heidelberg, Nov 2007, str. 722–735.

[144] Karimzadehgan, M., Zhai, C., Belford, G., "Multi-aspect expertise matching for review assignment", in Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008, str. 1113–1122.

[145] Wallach, H. M., Mimno, D. M., McCallum, A., "Rethinking LDA: Why Priors Matter", in Advances in neural information processing systems, 2009, str. 1973–1981.

[146] Wang, Y., Zhao, X., Sun, Z., Yan, H., Wang, L., Jin, Z., Wang, L., Gao, Y., Law, C., Zeng, J., "Peacock: Learning Long-Tail Topic Features for Industrial Applications", ACM Trans. Intell. Syst. Technol., Vol. 6, No. 4, Aug 2015, str. 1–23.

[147] Buntine, W. L., Mishra, S., "Experiments with non-parametric topic models", in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, Aug 2014.

[148] Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M., "Hierarchical Dirichlet Processes", Journal of the American Statistical Association, Vol. 101, No. 476, 2006, str. 1566–1581.

[149] Pitman, J., Yor, M., "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator", Annals of Probability, Vol. 25, No. 2, Apr 1997, str. 855–900.

[150] Lin, C.-J., "Projected gradient methods for nonnegative matrix factorization", Neural computation, Vol. 19, No. 10, 2007, str. 2756–2779.

[151] Boutsidis, C., Gallopoulos, E., "SVD based initialization: A head start for nonnegative matrix factorization", Pattern Recognition, Vol. 41, No. 4, Apr 2008, str. 1350–1362.

[152] Chen, C., Du, L., Buntine, W., "Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process", in Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, Sep 2011, str. 296–311.

[153] Gilks, W. R., Wild, P., "Adaptive Rejection Sampling for Gibbs Sampling", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 41, No. 2, 1992, str. 337–348.

[154] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J. *et al.*, "Api design for machine learning software: experiences from the scikit-learn project", arXiv preprint arXiv:1309.0238, 2013.

[155] Krawczyk, B., "Learning from imbalanced data: open challenges and future directions", Progress in Artificial Intelligence, Vol. 5, No. 4, Nov 2016, str. 221–232.

[156] Branco, P., Torgo, L., Ribeiro, R. P., "A Survey of Predictive Modeling on Imbalanced Domains", ACM Computing Surveys, Vol. 49, No. 2, Nov 2016, str. 1–50.

[157] Tan, P.-N., Steinbach, M., Kumar, V., "Introduction to Data Mining", Pearson, May 2005.

[158] Murphy, K. P., "Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)", The MIT Press, Aug 2012.

[159] Cortes, C., Vapnik, V., "Support-vector networks", Machine Learning, Vol. 20, No. 3, Sep 1995, str. 273–297.

[160] Breiman, L., "Random Forests", Machine Learning, Vol. 45, No. 1, Oct 2001, str. 5–32.

[161] Nocedal, J., Wright, S., "Numerical Optimization (Springer Series in Operations Research and Financial Engineering)", Springer, Jul 2006.

[162] Jebara, T., Kondor, R., "Bhattacharyya and Expected Likelihood Kernels", in Learning Theory and Kernel Machines. Springer Berlin Heidelberg, 2003, str. 57–71.

[163] Cawley, G. C., Talbot, N. L. C., "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation", Journal of Machine Learning Research, Vol. 11, No. Jul, 2010, str. 2079–2107.

[164] Ferri, C., Hernández-Orallo, J., Modroiu, R., "An experimental comparison of performance measures for classification", Pattern Recognition Letters, Vol. 30, No. 1, 2009, str. 27–38.

[165] Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A., "Interactive topic modeling", Machine Learning, Vol. 95, No. 3, Jun 2014, str. 423–469.

[166] Van Rossum, G., Drake Jr, F. L., Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[167] Lutz, M., "Learning Python, 5th Edition", O'Reilly Media, Jul 2013.

[168] Millman, K. J., Aivazis, M., "Python for scientists and engineers", Computing in Science & Engineering, Vol. 13, No. 2, 2011, str. 9–12.

[169] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. *et al.*, "Tensorflow: A system for large-scale machine learning", in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, str. 265–283.

[170] Gamma, E., Design patterns: elements of reusable object-oriented software. Pearson Education India, 1995.

Životopis

Damir Korenčić rođen je 18. travnja 1983. godine u Zagrebu u Hrvatskoj. Diplomski studij matematike (smjer računarstvo) završio je 2008. godine na Matematičkom odjelu Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu s temom diplomskog rada "Problem usmjeravanja vozila i metaheuristika tabu pretraživanja".

Od listopada 2008. godine do veljače 2010. godine radio je kao programer u tvrtki Beta Tau Beta. Krajem 2010. godine i početkom 2011. godine predavao je matematiku u Klasičnoj gimnaziji u Zagrebu i na Građevinskom fakultetu Sveučilišta u Zagrebu. Od prosinca 2010. godine do prosinca 2016. godine bio je zaposlen kao znanstveni novak na Institutu Ruđer Bošković, gdje se bavio obradom prirodnog jezika i algoritmima kompresije. Od siječnja 2017. godine ima status vanjskog suradnika na Institutu Ruđer Bošković. Od lipnja 2018. godine zaposlen je na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu, kao mlađi istraživač na projektu DATACROSS.

Njegovi istraživački interesi obuhvaćaju područja obrade prirodnog jezika i dubinske analize teksta, uži interesi metode vrednovanja tematskih modela i analize medijske agende, a sudjelovao je i u istraživanjima algoritama kompresije. U suautorstvu je objavio tri rada na međunarodnim znanstvenim skupovima i tri rada u časopisima s međunarodnom recenzijom indeksiranim u bazi Current Contents. Član je strukovne udruge ACM (Association for Computing Machinery). Govori engleski jezik i služi se njemačkim jezikom.

Popis objavljenih radova

Radovi u časopisima

1. Korenčić, D., Ristov, S., Šnajder, J., "Document-based Topic Coherence Measures for News Media Text", Expert Systems with Applications, Vol. 114, prosinac 2018., str. 357–373.
2. Ristov, S., Korenčić, D., "Using Static Suffix Array in Dynamic Application: Case of Text Compression by Longest First Substitution", Information Processing Letters, Vol. 115, No. 2, veljača 2015., str. 175–181.

3. Ristov, S., Korenčić, D., “Fast Construction of Space-optimized Recursive Automaton”, Software: Practice and Experience, Vol. 45, No. 6, travanj 2014., str. 783–799.

Radovi na međunarodnim znanstvenim skupovima

1. Korenčić, D., Grbeša-Zenzerović, M., Šnajder, J., “Topics and their Salience in the 2015 Parliamentary Election in Croatia: A Topic Model Based Analysis of the Media Agenda”, Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), 2016., str. 48–54
2. Korenčić, D., Ristov, S., Šnajder, J., “Getting the Agenda Right: Measuring Media Agenda Using Topic Models”, Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications (TM 2015), ACM 2015., str. 61–66.
3. Glavaš, G., Korenčić, D., Šnajder, J., “Aspect-oriented Opinion Mining from User Reviews in Croatian”, Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013), ACL 2013., str. 18–23.

Biography

Damir Korenčić was born on April 18, 1983 in Zagreb, Croatia. He received his Engineering degree in Mathematics (with specialization in Computer Science) from the University of Zagreb, Faculty of Science, Department of Mathematics in 2008 (thesis title: "The Vehicle Routing Problem and the Tabu Search Metaheuristic").

From October 2008 until February 2010 he was employed as a programmer at Beta Tau Beta ltd. At the end of 2010 and the beginning of 2011 he was teaching mathematics at the Classical Gymnasium in Zagreb and at the Faculty of Civil Engineering, University of Zagreb. From December 2010 until December 2016 he was employed as a research assistant at the Ruđer Bošković Institute, where he worked in the fields of natural language processing and compression algorithms. From January 2017 he has a status of an external research associate at the Ruđer Bošković Institute. From June 2018 he is employed at the Faculty of Electrical Engineering and Computing as a junior researcher on the DATACROSS project.

His research interests include natural language processing and text mining, and specific interests include topic model evaluation and media agenda analysis. He also took part in research of compression algorithms. He has co-authored three conference papers and three journal papers in the journals indexed in the Current Contents. He is a member of the ACM (Association for Computing Machinery). He is fluent in English and speaks German.