Topic Modeling of the SrpELTeC Corpus: A Comparison of NMF, LDA, and BERTopic

Teodora Mihajlov, Milica Ikonić Nešić, Ranka Stanković, Olivera Kitanović

Topic Modeling of the SrpELTeC Corpus: A Comparison of NMF, LDA, and BERTopic | Teodora Mihajlov, Milica Ikonić Nešić, Ranka Stanković, Olivera Kitanović | Annals of Computer Science and Information Systems | 2024 | |

10.15439/2024F1593

http://dr.rgf.bg.ac.rs/s/repo/item/0009167

Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 649–653
DOI: 10.15439/2024F1593
ISSN 2300-5963 ACSIS, Vol. 39

Topic Modeling of the SrpELTeC Corpus: A Comparison of NMF, LDA, and BERTopic

Teodora Mihajlov
0009-0008-8137-6750
Association for Language Resources and Technologies
ul. Studentski trg 3, Belgrade, Serbia
Email: teodoramihajlov@gmail.com

Milica Ikonić Nešić
0000-0002-0835-8889
Univ. of Belgrade, F. of Philology
ul. Studentski trg 3, Belgrade, Serbia
Email: milica.ikonic.nesic@fil.bg.ac

Ranka Stanković, Olivera Kitanović
0000-0001-5123-6273
0000-0002-7571-2729
Univ. of Belgrade, F. of Mining and Geology
ul. Đušina 7, Belgrade, Serbia
Email: {ranka.stankovic, olivera.kitanovic}@rgf.bg.ac.rs,

***Abstract*—Topic modeling is an effective way to gain insight into large amounts of data. Some of the most widely used topic models are Latent Dirichlet allocation (LDA) and Nonnegative Matrix Factorization (NMF). However, new ways to mine topics have emerged with the rise of self-attention models and pre-trained language models. BERTopic represents the current state-of-the-art when it comes to modeling topics. In this paper, we compared LDA, NMF, and BERTopic performance on literary texts in the Serbian language, both quantitatively by measuring Topic Coherency (TC) and Topic Diversity (TD), and by conducting a qualitative evaluation of the obtained topics. Additionally, for BERTopic, we compared multilingual sentence transformer embeddings with the Jerteh-355 monolingual embeddings for Serbian. NMF yielded the best Topic Coherency results, while BERTopic with Jerteh-355 embeddings gave the best Topic Diveristy. The monolingual Serbian Jerteh-355 embeddings also outperformed sentence transformer embeddings in both TC and TD.**

***Index Terms*—topic modeling, LDA, NMF, BERTopic, SrpELTeC, computational literary studies**

I. Introduction

TOPIC modeling has proven to be an effective tool for uncovering common themes and the underlying narratives in texts and for describing copious datasets. In social sciences, one way to leverage topic modeling is to explore topics in literary texts [1], [2].

In this paper, we present an evaluation of statistical and deep learning topic models on the SrpELTeC collection, Serbian part of ELTeC, the European Literary Text Collection, produced within COST Action CA16204 [3], [4]. The aims of the presented project are two-fold: (1) to explore topics in the SrpELTeC collection; (2) to evaluate and contrast the efficacy of conventional topic models, namely Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), with a transformer-based topic model, BERTopic, in analyzing long texts in the Serbian language. To the best of our knowledge, BERTopic has only been used for modeling topics in short texts in the Serbian language, namely on a dataset of tweets expressing hesitancy towards COVID-19 vaccination [5], where it outperformed both LDA and NMF. The model has not yet been applied to long or literary texts in the Serbian language.

To that end, this paper exploits natural language processing methods to obtain information about the Serbian literary texts, both in and outside of canon, which are presented in the SrpELTeC collection. The broader aim of this work is to, in the future, compile a comprehensive connected network of Serbian literary publications, based on the principles of *Wikidata* [6]. The insights that could further be derived from the project could be used not only for testing methods for modeling Serbian literary texts but also to pave the way for the development of approaches for digital humanities for Serbian. In the future, we will aim to rely on the principles of digital humanities that promote using big data paired with carefully curated metadata, following the example of the *MiMoText* research project in computational literary studies [7].

The remainder of the paper is structured as follows: section II delves into related work, covering both traditional and deep learning topic modeling methods and their use thus far; in III we cover data preprocessing steps (III-A) and methods used for obtaining text topics (III-B) and in IV we present the results, both qualitatively IV-A and qualitatively IV-B. Finally, in V, we lay out concluding remarks and propositions for further.

II. Related Work

Several methods provide insight into latent topics in texts. Two of the most widespread methods for topic modeling are Latent Dirichlet Allocation (LDA) [8], and Nonnegative Matrix Factorization (NMF) [9]. LDA is a generative probabilistic model, specifically, a three-level Bayesian model, which models each item in a corpus as a representation of

probabilities of underlying topics [8]. It has proven to be one of the most popular topic modeling algorithms [10]. In contrast, NMF is a non-probabilistic linear, decompositional algorithm, which relies on matrix factorization [9]. In the context of topic modeling, NMF is based on TF-IDF, transforming data by breaking down a matrix into two lower-ranking matrices [10]. Both models call for a predetermined number of topics. Adjusting the number of topics and fitting other parameters accordingly can be challenging [11]. In addition, the models call for extensive data preprocessing. Another downside of traditional methods such as LDA and NMF is that they represent documents in a bag-of-words fashion, which ignores both word order and their semantic relationship [12].

In recent years, the rise of self-attention [13] paved the way for the development of pre-trained language models (PLMs). In turn, this facilitated generating word embeddings and adjusting them for different tasks, such as topic modeling. BERTopic is a BERT-based PLM trained on the topic modeling task, which utilizes pre-trained embeddings to generate text topics [14]. On top of the generated embedding, BERTopic leverages dimensionality reduction and clustering techniques, which are by default UMAP and HDBSCAN, respectively. To create topic representations, the model uses c-TF-IDF, a class-based variation of TF-IDF [10]. One of the perks of BERTopic is its modularity. Although the model has default settings for each of the aforementioned steps, the user can choose different algorithms and parameters for each of the steps, adjusting the model to their data and goals, which makes it a scalable topic modeling solution [14]. Unlike LDA and NMF, BERTopic does not require a predefined number of topics. The main downside of the model is that it assigns only one topic to each document [14].

III. Materials and Methods

A. Data Description and Preprocessing

The data comprises the Serbian part of the ELTeC corpus - a multilingual collection of novels written in the period 1840-1920. The Serbian ELTeC collection, encompasses 100 novels, while the entire collection consists of 157 novels [3]. The 100 novels used here were written by 66 different authors, 62 male, and 4 female, and were published between 1852 and 1920. Two novels in the collection are written by unknown authors. The average novel length is 49,315 words. The remaining novels are currently being prepared and will be a part of an extended sub-collection SrpELTeC-ext.3 [3]. For the purpose of this research we used SrpELTeC TXM Copus [1] of 108 novels in level-2. [2]. Novels in level-2 are annotated with part of speech (POS), lemma, and 7 categories of named entities: persons (PERS),organisations (ORG), locations (LOC), demonyms (DEMO), work of art (WORK), events (EVENT), and titles and professions (ROLE) [15]. Such annotated corpus allows for analysis to be conducted using only nouns (NOUN), of which there are a total of 854,835 in this collection, with 30,684 being unique.

We used a spaCy Python package for the Serbian language for text preprocessing.[3] First, we removed special characters and converted text to lowercase. Next, we converted the text from the Serbian Cyrillic script to Latin script. Finally, we tokenized and lemmatized the text, and removed stopwords. The stopwords consisted of a list of stopwords for the Serbian language, as well as corpus-specific stopwords. The corpus-specific stopwords were extracted by observing keywords while implementing the initial versions of all three models.

B. Models

Latent Dirichlet allocation (LDA). We implemented an LDA model using the Gensim library. Text tokens were obtained using the BoW approach, whereby both bigrams and trigrams were created. Subsequently, we generated a TF-IDF representation of the documents and filtered out words with a frequency < 0.03. To fine-tune the number of topics, we iterated the number of topics between 2 and 10 and evaluated Topic Coherence for each iteration. Finally, we picked 5 topics, since that yielded the best TC score.

Nonnegative Matrix Factorization (NMF). NMF was implemented with the Sci-kit Learn library, as it displayed significantly better results in comparison to its Gensim equivalent. The minimal word frequency was set to 15, and the maximal frequency was an occurrence of a word in 80% of the documents. A full list of stopwords, i.e. Serbian and corpus-specific, was passed to the model, in case some were not initially removed. After calculating topic coherence for the number of topics between 2 and 10, we set the parameter to 7 topics.

BERTopic. For BERTopic, we exploited the modular architecture of the model and tried to best fit each of its components to our data and research aims. We first generated word embeddings. As the default word embeddings, BERTopic uses sentence transformers [16], which support English and include three multilingual sentence transformer models that are trained for 50+ languages including Serbian:

- *distiluse-base-multilingual-cased-v2*: the model maps into 512-dimensional dense vector space and can be used for tasks like clustering or semantic search (480 MB, 135 million parameters).
- *paraphrase-multilingual-MiniLM-L12-v2*: this model maps sentences to a 384-dimensional dense vector space (420 MB, 117 million parameters).
- *paraphrase-multilingual-mpnet-base-v2*: this model maps sentences to a 768-dimensional dense vector space (970 MB, 278 million parameters).

In addition to the three multilingual embedding models, we tested the *Jerteh-355* embeddings. The *Jerteh-355* model is the largest model trained specifically for the Serbian language [17]. Although the model is not fine-tuned for the semantic search task, we wanted to see how it performs in

[1] https://live.european-language-grid.eu/catalogue/corpus/23621
[2] https://github.com/COST-ELTeC/ELTeC-srp/tree/master/level2
[3] https://github.com/procesaur/srpski

TABLE I
TOPIC COHERENCE AND TOPIC DIVERSITY OF THE MODELS

Model	TC	TD
LDA	0.361	0.940
NMF	**0.568**	0.757
BERTopic		
distiluse-base-multilingual-cased-v2	0.427	0.869
paraphrase-multilingual-MiniLM-L12-v2	0.387	0.864
paraphrase-multilingual-mpnet-base-v2	0.299	0.925
Jerteh-355	**0.456**	**0.970**

TABLE II
LDA GENERATED KEYWORDS (NOUN)

Topic	Top 10 keywords
Topc0	knez (*knyaz*), vojvoda (*duke*), vojska (*army*), dvor (*castle*), gospodar (*lord*), junak (*hero*), car (*tsar*), šator (*tent*), pop (*priest*), vlastela (*Medieval Serbian nobility*)
Topc1	gospođa (*mam*), gospodin (*sir*), pop (*priest*), doktor (*doctor*), učitelj (*teacher*), škola (*school*), mati (*mother*), kapetan (*captain*), manastir (*monastery*), dete (*child*)
Topc2	narod (*people*), čovek (*man*), gospodar (*lord*), kapetan (*captain*), vezir (*vizier*), knez (*knyaz*), gospodin (*sir*), kmet (*serf*), čiča (*uncle*), koliba (*hut*)
Topc3	gazda (*lord*), gospodar (*sir*), pop (*priest*), čovjek (*man*), riječ (*word*), planina (*mountain*), dućan (*store*), talijer (*talir*), vrijeme (*time/weather*), djeca (*children*)
Topc4	društvo (*society*), čovek (*man*), reč (*word*), deca (*children*), načelo (*principle*), sloboda (*freedom*), dete (*child*), stanje (*condition*), ženskinja (*woman*), nauka (*science*)

comparison to the aforementioned multilingual models. The model specifics are as follows:

- *Jerteh-355* the model size is 355 million parameters, and it was trained on 4 billion tokens in the Serbian language.

For the dimensionality reduction step, we used UMAP, with the following parameters - *n_neighbors = 5*; *n_components = 5*. For clustering, we used HBDSCAN with the minimal cluster size set to 3. The rest of the UMAP and HBDSCAN parameters were default.

Lastly, to create topic representations, BERTopic utilizes CountVectorizer and class-based c-TF-IDF, to model the importance of each document cluster. We used CountVectorizer to filter out noise from the data: additional stopwords were removed, and all words with frequency <5 and >80% of the documents were filtered out. We looked at both unigrams and bigrams.

After generating topics, BERTopic creates a -1 topic that contains outlier documents. To remove outliers, it offers an *outlier_reduction* function. However, when we tried using this function, we got different topic keywords, which were worse than those originally generated. Therefore, in this phase of research, we opted out of using the *outlier_reduction* option.

C. Evaluation

The models were evaluated quantitatively and qualitatively. For the quantitative evaluation, we used Topic Coherence (TC) and Topic Diversity (TD) measures, both of which are frequently used for evaluating topic models [14], [10], [18]. TC is a measure of semantic relatedness between the words for each topic [19]. We used the C_V coherence measure, which is based on a combination of a sliding window, a one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity [20]. TC ranges from 0 to 1, with values closer to 1 signify more related topic words. TD computes the percentage of unique words for top_n words for each topic. It ranges from 0 to 1, where 1 marks more related, and 0 more redundant topics [18].

IV. RESULTS AND DISCUSSION

A. Quantitative Evaluation

Using all POS, NMF significantly outperformed both LDA and BERTopic in Topic Coherency (TC). BERTopic, however, generated the most diverse topics (Topic Diversity, TD). In topic diversity, LDA came close to BERTopic, while NMF displayed a significant difference in TD measure in comparison with the two other models, as displayed in Table I. The best model performance is presented in bold, while the best performance among different embeddings for BERTopic is presented in bold and underlined font.

For BERTopic, we can see that, although it was not fine-tuned for semantic search, a monolingual embedding model for Serbian, Jerteh-355 outperformed the three sentence transformer models in both TC and TD (Table I).

B. Qualitative Evaluation

For the qualitative analysis of the obtained topics, we will look into *top_n* keywords for each of the models. Translations of all keywords are presented in brackets. Personal names

TABLE III
NMF GENERATED KEYWORDS (NOUN)

Topic	Top 10 keywords
Topc0	vojska (*army*), drum (*road*), vojnik (*soldier*), kapetan (*captain*), borba (*battle*), neprijatelj (*enemy*), bol (*pain*), komanda (*comand*), oficir (*officer*), planina (*mountain*)
Topc1	knez (*knyaz*), đeneral (*general*), gospodar (*sir*), ministar (*minister*), seljak (*peasant*), načelnik (*chief*), otrov (*poison*), doktor (*doctor*), vojvoda (*duke*), svetlost (*duke*)
Topc2	despot (*despot*), vojska (*army*), vojvoda (*duke*), grad (*city*), paš (*pasha*), sultan (*Sultan*), kaluđer (*monk*), car (*emperor*), dvor (*castle*), manastir (*monastery*)
Topc3	slovo (*letter*), đak (*pupil*), učitelj (*teacher*), manastir (*monastery*), kmet (*serf*), iguman (*Hegumen*), majstor (*meister*), gazda (*lord*), kaluđer (*monk*),arhimandrit (*archimandrite*)
Topc4	vezir (*vizier*), ratnik (*warrior*), hajduk (*hajduk*), aga (*agha*), vojvoda (*duke*), družina (*length*), gospodar (*lord*), tvrđava (*fortress*), straža (*watch, as in Night watch*), grad (*city*)
Topc5	gospođa (*mam*), doktor (*doctor*), kćer (*daughter*), gospodar (*lord*), udovica (*widow*), gospođica (*mam*), sahat (*hour*), tetak (*uncle*), advokat (*lawyer*), dama (*lady*)
Topc6	pop (*priest*), kmet (*serf*), učitelj (*teacher*), kapetan (*captain*), baba (*grandmother*), gospoja (*lady*), sokak (*street*), avlija (*courtyard*), ćata (*clerk*), gazda (*lord*)

TABLE IV
BERTOPIC GENERATED KEYWORDS BY THE *jerteh-355* MODEL EMBEDDIGS (NOUN)

Topic	No. of Documents	Top keywords
-1	9	despot (*despot*), kir (*lord*), sultan (*sultan*), ćesar (*emperor*), vezir (*vizier*), bula (*Muslim woman*), knežević (*prince*), vlastelin (*nobleman*), patrijarh (*patriarch*), vjera (*faith*)
0	19	hanum (*lady*), fratar (*friar*), robinja (*slave woman*), gospoja (*madam*), hanuma (*lady*), tatko (*father*), naprednjak (*progressive*), đakon (*deacon*), fala (*thanks*), nena (*grandmother*
1	10	arhimandrit (*archimandrite*), major (*major*), patrijarh (*patriarch*), grof (*count*), nastojatelj (*superior*), djevojka (*girl*), koi (*which*), zadruga (*cooperative*), cigareta (*cigarette*), riječ (*word*)
2	7	đakon (*deacon*), monah (*monk*), kasta (*caste*), frajla (*lady*), ogrlica (*necklace*), presednik (*president*), čika (*uncle*), forinta (*forint*), senat (*senate*), adiđar (*jewelry*
3	6	aga (*aga*), subaša (*overseer*), čorbadži (*chief*), kahva (*coffee*), hodža (*imam*), tatko (*father*), loža (*lodge*), duhovnik (*spiritual father*), riječ (*word*), svijet (*world*
4	6	nazaren (*nazarene*), bukvar (*primer*), gradina (*garden*), tablica (*tablet*), tabla (*board*), cigla (*brick*), slovo (*letter*), apostol (*apostle*), sotona (*Satan*), crep (*roof tile*
5	6	ćata (*boss*), aga (*aga*), gospa (*lady*), front (*front*), stanica (*station*), baterija (*battery*), vagon (*wagon*), dućandžija (*shopkeeper*), divizija (*division*), automobil (*car*
6	5	šator (*tent*), arhimandrit (*archimandrite*), trpezarija (*dining room*), vojvoda (*duke*), knez (*prince*), vlastela (*nobility*), tag (*tag*), prisednik (*president*), beležnik (*notary*), društvo (*society*), knez vojvoda (*duke prince*
7	5	grofica (*countess*), kneginjica (*princess*), dragana (*darling*), grof (*count*), nana (*grandmother*), urednik (*editor*), teta (*aunt*), drama (*drama*), gospa (*lady*), šor (*street*)
8	5	vezir (*vizier*), gospodin ministar (*minister*), vranac (*black horse*), kavana (*tavern*), žandarm (*gendarme*), posluživanje (*service*), česma (*fountain*), aga (*aga*), pobra (*peasant*), glavar (*chief*)

are not translated but are capitalized and indicated with an abbreviation *pers.*. Labels of keywords relating to other named entities such as cities or buildings are also presented in *italic*, preceding word translation.

Analysis of the keywords of the final LDA model presented in Table II determined that the model generated diverse keywords.

However, many keywords are names of the characters in the novels. Even though they are relevant to the corpus, they do not tell us much about latent topics in the texts, especially if one is unfamiliar with novels included in SrpELTeC. By carefully examining NMF keywords (Table III), we ascertained that the model yielded the most informative keywords, which refer to the topics discussed in the novels. The keywords mention words such as **turčin** (eng. Turk), **vojvoda** (duke), **manastir** (eng. monastery),

combined with character names. The keywords are informative of both main topics discussed in the novels and of the main characters that represent the corpus.

No matter the embedding model or parameter setting, BERTopic continually generated names of the characters in the novels as keywords (Table IV), which do not provide us with sufficient information about the topics. After adjusting the model parameters, only 6 documents out of 100 (6%) were assigned the -1 topic, i.e. classified as outliers.

V. CONCLUSION AND FURTHER WORK

In this paper, we compared traditional topic models, NMF and LDA, with a transformer-based BERTopic, on the SrpELTeC collection. Although we expected BERTopic to outperform the traditional models significantly, it only did so when it came to topic diversity. When it comes to TC, it fell somewhat short compared to NMF.

To address the limitations of current work, we plan to try and see the effects of using chunked text on topic extraction. Even though our current corpus comprises 49,315 words, it consists of just 100 documents. Both the length of the texts as well as the small number of documents could be hindering the current model's performance. Therefore, better topics might be obtained by passing shorter chunks to BERTopic might improve results. Lastly, we plan on trying to improve BERTopic performance with chunked documents, as well as see how it performs with other embedding models for the Serbian language, such as BERTić [21].

ACKNOWLEDGMENT

This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings - Serbian Language Applications - TESLA*.

REFERENCES

[1] I. Uglanova and E. Gius, "The order of things. a study on topic modelling of literary texts." *CHR*, no. 18-20, p. 2020, 2020.

[2] K. E. Chu, P. Keikhosrokiani, and M. P. Asl, "A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts," *Pertanika Journal of Science & Technology*, vol. 30, no. 4, pp. 2535–2561, 2022, https://doi.org/10.47836/pjst.30.4.14.

[3] R. Stanković, C. Krstev, B. Š. Todorović, D. Vitas, M. Škorić, and M. I. Nešić, "Distant reading in digital humanities: Case study on the serbian part of the eltec collection," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3337–3345. [Online]. Available: https://aclanthology.org/2022.lrec-1.356

[4] C. Schöch, T. Erjavec, R. Patras, and D. Santos, "Creating the european literary text collection (eltec): Challenges and perspectives," *Modern Languages Open*, 2021, http://doi.org/10.3828/mlo.v0i0.364.

[5] D. Medvecki, B. Bašaragin, A. Ljajić, and N. Milošević, "Multilingual transformer and bertopic for short text topic modeling: The case of serbian," in *Conference on Information Technology and its Applications*. Springer, 2024, pp. 161–173, https://doi.org/10.1007/978-3-031-50755-7_16.

[6] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014, https://doi.org/10.1145/2629489.

[7] C. Schöch, M. Hinzmann, J. Röttgermann, K. Dietz, and A. Klee, "Smart modelling for literary history," *International Journal of Humanities and Arts Computing*, vol. 16, no. 1, pp. 78–93, 2022, https://doi.org/10.3366/ijhac.2022.0278.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[9] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:2095855

[10] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886498, 2022.

[11] ——, "Identifying hidden semantic structures in instagram data: a topic modelling comparison," *Tourism Review*, vol. 77, no. 4, pp. 1234–1246, 2021.

[12] M. Švaňa, "Social media, topic modeling and sentiment analysis in municipal decision support," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023, pp. 1235–1239, http://dx.doi.org/10.15439/2023F1479.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017, https://doi.org/10.48550/arXiv.1706.03762.

[14] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022, https://doi.org/10.48550/arXiv.2203.05794.

[15] R. Stanković, C. Krstev, B. Šandrih Todorović, and M. Škorić, "Annotation of the serbian eltec collection," *Infotheca - Journal for Digital Humanities*, vol. 21, no. 2, pp. 43–59, 2022. [Online]. Available: https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.2.3_en

[16] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019, https://doi.org/10.48550/arXiv.1908.10084.

[17] M. Škorić, "Novi jezički modeli za srpski jezik," *Infoteka*, vol. 24, 2024, https://doi.org/10.48550/arXiv.2402.14379. [Online]. Available: https://arxiv.org/abs/2402.14379

[18] A. B. Dieng, F. J. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020, https://doi.org/10.48550/arXiv.1907.04907.

[19] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 100–108.

[20] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408, https://doi.org/10.1145/2684822.2685324.

[21] N. Ljubešić and D. Lauc, "BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Kiyv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 37–42, https://doi.org/10.48550/arXiv.2104.09243. [Online]. Available: https://www.aclweb.org/anthology/2021.bsnlp-1.5