Analyzing Topic Models: A Tourism Recommender System Perspective

Maryam Kamal[1], Gianfranco Romani[1], Giuseppe Ricciuti[2], Aris Anagnostopoulos[1], and Ioannis Chatzigiannakis[1(✉)]

[1] Department of Computer, Control and Management Engineering (DIAG), Sapienza University of Rome, 00185 Rome, Italy
ichatz@diag.uniroma1.it

[2] Amarena Company Srl, Via Casilina 3T, 00182 Rome, Italy

Abstract. Topic Modeling is a well-known text-mining strategy that detects potential underlying topics for documents. It plays a pivotal role in recommender systems for processing proliferated user-generated content (UGC) for personalized recommendations. Its application presents unique challenges in tourism sector due to the diversity, dynamicity, and multimodality of tourism data. This study presents a comprehensive analysis of selected promising topic models specifically in context of tourism recommender systems. The study conducts experimental evaluation of models' performance on five datasets, and highlights their advantages and unique characteristics based on multiple evaluation parameters. Results reveal no best approach in general, rather optimality of models depend on data characteristics, as thoroughly discussed in this paper. It further discusses open issues for the tourism context-related application of topic models, and future research directions.

Keywords: Topic Modeling · Text Mining · Comparative Analysis · Touristic Experiences

1 Introduction

Over past few years, the technological evolution and increased adoption of web-based platforms have caused sheer expansion in the volumes of user-generated content (UGC) [10,12]. Particularly, for the tourism industry, UCG has become an integral part of all tourism activities. However, the explosion of UGC on web along with broad diversity of content makes it imperative to acquire the interpretation and profiling of content and users. This has necessitated the development of advanced tourism recommender systems primarily relying on tourists' UGC. Processing textual UGC such as tourists' experiences and reviews is crucial for recommender systems, here topic modeling (TM) serves as a pivotal strategy. Topic Modeling links a vast volume of unstructured UGC and the diverse needs for personalized tourism recommendations.

This research is supported by Amarena Company srl.

L. Barolli (Ed.): AINA 2024, LNDECT 200, pp. 250–262, 2024.
https://doi.org/10.1007/978-3-031-57853-3_21

Topic modeling (TM) is a well-known data mining technique that detects potential latent topics for documents based on semantic relevance of words and documents [11]. It plays multifaceted role in context of recommender systems, such as enhancing the personalization aspect of an RS by identifying users' preferences as topics from their shared experiences. At the same time, it extracts the potential interest topics for users from the sheer volume of UGC, streamlining the recommendations. This has made topic modeling one of the most in-demand techniques in the domain of tourism, where topics and labels are required to associate diverse preferences of tourists to related offerings by the travel business, considering the travelers' reviews and user-generated content.

The application of topic models for tourism-related data is particularly unique and challenging. The aim is to acquire topics considering the underlying sentiments, preferences, experiences, and expectations of tourists. Simultaneously, the diversity in tourists' generated content and the strong co-occurrence of emotion-oriented vocabulary are different from the characteristics of blog posts expressing other opinions. Moreover, in comparison to other types of data such as, in microblog services like Twitter, documents reporting touristic experiences are longer, while on the contrary, they are much shorter when compared to articles found in journals or encyclopedias [15]. Such structural differences in the corpus and the vocabulary have a significant impact on the performance of topic models. It is therefore important to evaluate and analyze how topic models perform on tourism-related data and understand the reasons for such performances. To address this need, our study provides a comprehensive analysis of promising topic models in the context of tourism recommender systems. In particular, the Latent Dirichlet Allocation (LDA) [5], Non-Negative Matrix Factorization (NMF) [13], Top2Vec [2], Bidirectional Encoder Representations from Transformers (BERTopic) [7], RoBERTa [14], Contextualized Topic Model (CTM) [4], and Embedded Topic Model (ETM) [6] are analyzed comprehensively.

2 Background

2.1 Preliminaries

This section briefly provides an understanding of the base principles and mechanisms for each selected topic model.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for text, using Dirichlet hyperparameters α and β. The goal is to maximize the probability of document corpus D given these hyperparameters, as in Eq. 1.

$$\text{Maximize } P(D|\alpha, \beta) \tag{1}$$

Top2Vec uses word and document embeddings to discover latent semantic structures in text. It automatically determines the number of topics and does not require preprocessing like stopword removal.

Non-Negative Matrix Factorization (NMF decomposes a term-document matrix A into two non-negative matrices W and H, as shown in Eq. 2.

$$A = W \times H \tag{2}$$

It iteratively updates these matrices to extract topics from data.

BERTopic employs BERT embeddings and transformer embeddings, using class-based TF-IDF ($cTF - IDFscoring$) to evaluate term significance in clusters. *RoBERTa* is an optimized version of BERT, focusing on word context for topic prediction.

Contextualized Topic Model (CTM) includes CombinedTM, which combines contextual embeddings with bag-of-words, and ZeroShortTM, which supports multilingual topic modeling.

Embedded Topic Model (ETM) integrates LDA with a variational autoencoder and word embeddings, generating topic proportions for each document.

2.2 Recent Studies

In recent years, multiple studies and researchers have found topic models significantly helpful to tourism-related concerns. Prominently, topic modeling is used to discover preferences in travel itineraries, to study customers' opinions, and to

Table 1. Application of topic models by recent studies in tourism

Studies using TM in Tourism field			
Study	Objectives	Model(s) Used	Evaluation Metrics
Y. Guo et al. (2017) [8]	Tourist satisfaction analysis	LDA	Jaccard coefficient, human analysis and Standford Topic Modelling Toolbox
J. Bao et al. (2017) [3]	Bikesharing	LDA	Perplexity
H. Quab Vu et al. (2019) [16]	Analysis of travel itineraries	LDA	Perplexity, topic concentration
N. Hu et al. (2019) [9]	Customers' complaints	STM	Several analysis on the topics obtained. No specific metric score
Q. Yan et al. (2022) [17]	interaction actors and experience detection	LDA	Content Analysis
N. Zhao. et al. (2023) [18]	Cultural tourism promotion	LDA	Perplexity and Classification

make recommendations. Since our study involves the application of topic models in the context of touristic experiences, we have summarized some recent relevant studies for topic modeling in tourism, in Table 1.

3 Materials and Methods

3.1 Datasets

For experimental evaluations and analysis of topic models, we have used a total of five unique datasets. Three datasets are exclusively designed for this study, namely AirBnB Touristic Experiences (ATE), TripAdvisor Tourist Activities (TAT), and KuriU (KU). While two datasets are publically available, 20NewsGroup (20NG) and TourPedia (TP). We collected exclusive datasets by web-scraping online posted touristic experiences on leading web-based tourism platforms; AirBnB for ATE and TripAdvisor for TAT. KU is a sampled dataset of the touristic experiences recommender system research project, KuriU, whose module is this study. Note that KU is an Italian language dataset, while we filtered English documents from other datasets. This is to analyze multi-lingual behavior of models. The datasets include touristic experience data for the city of Rome and their statistical summary is mentioned in Table 2.

Table 2. Statistics of the datasets

Dataset	# of Docs	# of Words	Vocabulary Size	Avg. Words Per Doc
ATE	611	111,169	12,661	182
TAT	1860	192,087	13,723	103
KU	5,724	1,556,416	138,095	272
TP	8,000	191,996	27,012	24
20NG	18,846	3,423,145	29,548	182

3.2 Evaluation Parameters

The study evaluates topic models on the following parameters:

Topic Diversity (TD): It measures the distinctiveness of the document clusters produced by the models, using Eq. 3. The value of topic diversity usually ranges between 0 and 1, where a value close to 1 means higher topic diversity while a value closer to 0 means a lower topic diversity. A model is appreciated if it produces higher topic diversity for a given dataset.

$$TD = \frac{n(U)}{K * n(T)} \tag{3}$$

Here, in Eq. (3), n(U) represents the cardinality of the set of unique words U. K represents the top K words for all topics. T represents the set of topics generated by the model where n(T) is the cardinality of the set T.

Inverted RBO (IRBO): It illustrates to what extent topics differ from each other. It ranges from 0 to 1, where 0 means fully identical and 1 means fully diverse topics. It penalizes topics with common words at different rankings less than topics sharing the same words at the highest ranks.

Topic Coherence: It measures the interpretability and coherence of the topics produced by a model and its association with the considered data. A higher value of topic coherence represents better results of a topic model in terms of producing coherent topics. Let N top words of a topic, $P(w_i, w_j)$ refers to the probability of occurrence of words w_i and w_j together, while $P(w_i)$ and $P(w_j)$ is the probability of occurrence of these words individually, we used the following four types of topic coherences:

– C_{uci} that is calculated by using Eq. 4,

$$C_{uci} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \log \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)} \tag{4}$$

– C_v that is estimated by using by Eq. 5 and Eq. 6,

$$\vec{v}(W') = \left\{ \sum_{w_i \epsilon W'} \left(\frac{log \frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log(P(w_i, w_j)+\epsilon)} \right)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{5}$$

$$\Phi_{s_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i.w_i}{\|\vec{u}\|_2.\|\vec{w}\|_2} \tag{6}$$

In Eq. (5) the context vector is $\vec{v}(W')$ and Φ is the confirmation measure.

– C_{umass} calculated by using Eq. 7,

$$C_{umass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, wj) + \epsilon}{P(wj)} \tag{7}$$

– C_{npmi} is an improvisation of the C_{uci} coherence that uses normalized pointwise mutual information (NPMI).

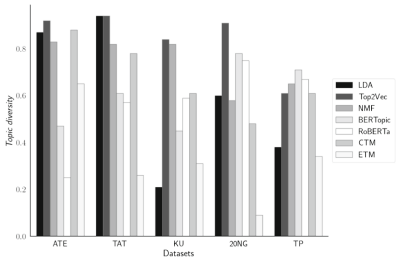
3.3 Experiment and Results

This subsection presents the results and analysis from our experimental evaluations. The implementations are conducted using Python version 3.9.7 on Jupyter Notebook and Google Colab. The coherence evaluation parameters are estimated using Gensim toolkit, while topic diversity measures are estimated using Octis toolkit. Each model is tested with ten iterative runs and the results mentioned in this section are average recorded for each experiment. For the experimentation, we used the default text embedding models for each strategy, which are: *Doc2Vec* for Top2Vec, *roberta-base-nli-stsb-mean-tokens* for RoBERTa, and *all-MiniLM-L6-v2* for BERTopic (English datasets) while *paraphrase-multilingual-MiniLM-L12-v2* for Italian language dataset. Note that we pre-defined the number of

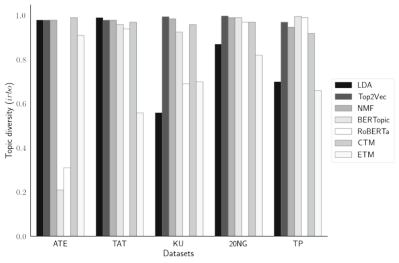
topics for LDA, NMF, CTM, and ETM using the elbow method, while Top2Vec, BERTopic, and RoBERTa are modeled to decide the best suitable number of topics by themselves. Note that as per the requirement of some models, LDA, NMF, and ETM were provided pre-processed data, including removal of stopwords and special characters, and lemmatization.

Topic Diversity: An interesting quality determinant explored in this study is topic diversity. A model is well-appreciated if it estimates higher topic diversity with a suitable number of topics. Figure 1 shows the results obtained in this regard, where Fig. 1a illustrates a comparison of the models with respect to average topic diversity (TD) and Fig. 1b shows average Inverted RBO (IRBO) achieved for each dataset. Here Top2Vec shows higher topic diversity on average, for both cases, considering all datasets. An interesting finding is for TP dataset from Fig. 1a, which illustrates a reduced variation of topic diversity among models and BERTopic as the best method. Similarly, it is interesting to observe from Fig. 1b that BERTopic and RoBERTa show much less IRBO when applied to a small-sized dataset with shorter document lengths like ATE. Note that although Top2Vec provides higher topic diversity on average, the number of clusters (topics) it has produced is also considerably less for almost every dataset (Fig. 1a). This might also indicate a high diversity within a topic cluster which is expected to be less for a good topic model.

Topic Coherence: Remark that the higher the coherence score, the better coherent the topics, except for C_{umass}, where a lower value represents better coherence, according to Gensim implementation [1].



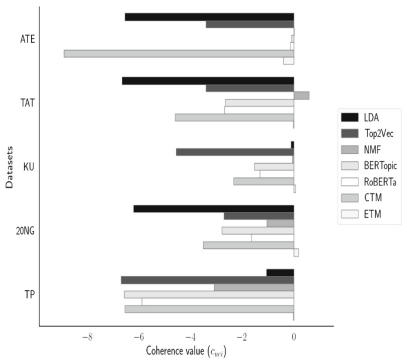
(a) *Comparison of the methods using topic diversity*



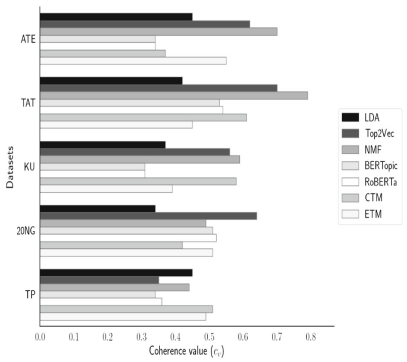
(b) *Comparison of the methods using IRBO score*

Fig. 1. Topic modeling evaluation based on Diversity metrics

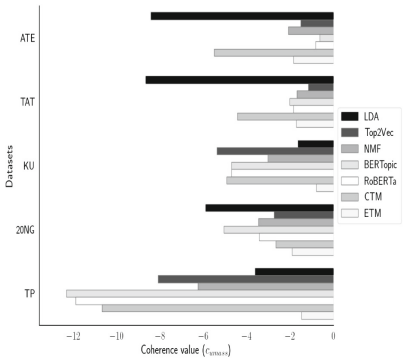
From Fig. 2a, NMF shows better C_{uci} for relatively smaller-sized datasets as ATE and TAT, but as the size of the datasets grows, ETM starts depicting better results. On average ETM concludes to deliver maximum coherence as compared to the others, in terms of C_{uci}. For C_v from Fig. 2b, while NMF shows better coherence on average for 3 out of 5 datasets, its performance degrades when applied to the largest dataset, 20NG, here Top2Vec exhibits better C_v



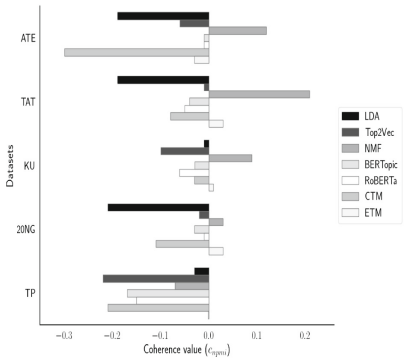
(a) *Comparison of the methods using* C_{uci} *score*



(b) *Comparison of the methods using* C_v *score*



(c) *Comparison of the methods using* C_{umass} *score*



(d) *Comparison of the methods using* C_{npmi} *score*

Fig. 2. Topic modeling evaluation based on Coherence metrics

than others. This may imply a sensibility of NMF to the sizes of the datasets, where this model seems suitable for small to medium-sized datasets when considering C_v. From Fig. 2c for C_{umass}, LDA outperforms others on average, while Top2Vec shows better performance for the Italian language dataset KU. Note that although LDA shows better C_{umass} on average, BERTopic outperforms all in the case of the medium-sized English dataset TP. This implies the adoption of LDA for small and large-sized English datasets when considering C_{umass} coherence. Top2Vec might be applied if dealing with multi-lingual medium-sized datasets while BERTopic is suggested for medium-sized English datasets when C_{umass} is concerned. Another interesting shortcoming is from Fig. 2d, where NMF illustrates better C_{npmi} in almost every dataset (except for TP) and on average as a whole. Notice that for TP, ETM outperforms all in terms of C_{npmi}.

Table 3. Comparisons of the results on ATE dataset.

Models	C_{uci}	C_v	C_{umass}	C_{npmi}	Topic Diversity (TD)	IRBO	Number of topics
LDA	−6.56	0.45	**-8.45**	−0.19	0.87	0.98	14
Top2Vec	−3.42	0.62	−1.52	−0.06	**0.92**	0.98	6
NMF	**0.01**	**0.70**	−2.09	**0.12**	0.83	0.98	14
BERTopic	−0.10	0.34	−0.63	−0.01	0.47	0.21	3
RoBERTa	−0.14	0.34	−0.83	−0.01	0.25	0.31	10
CTM	−8.93	0.37	−5.51	−0.30	0.88	**0.99**	14
ETM	−0.40	0.55	−1.85	−0.03	0.65	0.91	14

Table 4. Comparisons of the results on TAT dataset.

Models	C_{uci}	C_v	C_{umass}	C_{npmi}	Topic Diversity (TD)	IRBO	Number of topics
LDA	−6.68	0.42	**-8.68**	−0.19	**0.94**	**0.99**	16
Top2Vec	−3.42	0.70	−1.17	−0.01	**0.94**	0.98	6
NMF	**0.59**	**0.79**	−1.70	**0.21**	0.82	0.98	16
BERTopic	−2.66	0.53	−2.02	−0.04	0.61	0.96	45
RoBERTa	−2.69	0.54	−1.86	−0.05	0.57	0.94	44
CTM	−4.61	0.61	−4.44	−0.08	0.78	0.97	16
ETM	−0.03	0.45	−1.72	0.03	0.26	0.56	16

Table 5. Comparisons of the results on KU dataset.

Models	C_{uci}	C_v	C_{umass}	C_{npmi}	Topic Diversity (TD)	IRBO	Number of topics
LDA	−0.11	0.37	−1.65	−0.01	0.21	0.56	22
Top2Vec	−4.57	0.56	**−5.38**	−0.10	**0.84**	**0.99**	50
NMF	−0.05	**0.59**	−3.03	**0.09**	0.82	0.98	22
BERTopic	−1.53	0.31	−4.72	−0.03	0.45	0.93	75
RoBERTa	−1.32	0.31	−4.72	−0.06	0.59	0.69	14
CTM	−2.34	0.58	−4.94	−0.03	0.61	0.96	22
ETM	**0.06**	0.39	−0.79	0.01	0.31	0.70	22

Also for 20NG, ETM and NMF deliver the same readings. Hence we can state that NMF performs better for small to medium-sized datasets, while ETM performs better for medium to large-sized datasets if C_{npmi} is concerned.

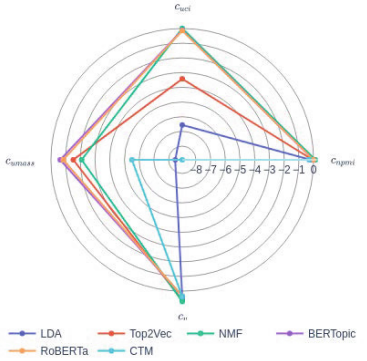
Table 6. Comparisons of the results on TP dataset.

Models	C_{uci}	C_v	C_{umass}	C_{npmi}	Topic Diversity (TD)	IRBO	Number of topics
LDA	−1.06	0.45	−3.62	−0.03	0.38	0.70	14
Top2Vec	−6.72	0.35	−8.10	−0.22	0.61	0.97	41
NMF	−3.10	0.44	−6.27	−0.07	0.65	0.95	14
BERTopic	−6.59	0.34	**-12.35**	−0.17	**0.71**	**0.99**	142
RoBERTa	−5.91	0.36	−11.91	−0.15	0.67	**0.99**	106
CTM	−6.57	**0.51**	−10.70	−0.21	0.61	0.92	14
ETM	**–0.03**	0.49	−1.49	**–0.01**	0.34	0.66	14

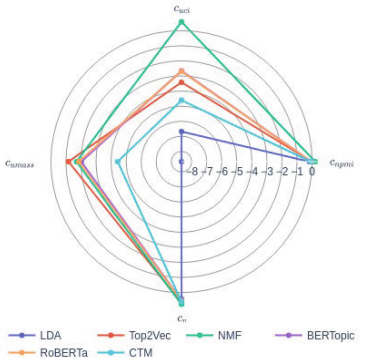
Table 7. Comparisons of the results on 20NG dataset.

Models	C_{uci}	C_v	C_{umass}	C_{npmi}	Topic Diversity (TD)	IRBO	Number of topics
LDA	−6.23	0.34	**–5.92**	−0.21	0.60	0.87	111
Top2Vec	−2.72	**0.64**	−2.74	−0.02	**0.91**	**0.99**	83
NMF	−1.05	0.49	−3.46	**0.03**	0.58	**0.99**	111
BERTopic	−2.80	0.51	−5.06	−0.03	0.78	**0.99**	216
RoBERTa	−1.64	0.52	−3.43	−0.01	0.75	0.97	90
CTM	−3.53	0.42	−2.67	−0.11	0.48	0.97	111
ETM	**0.19**	0.51	−1.91	**0.03**	0.09	0.82	111

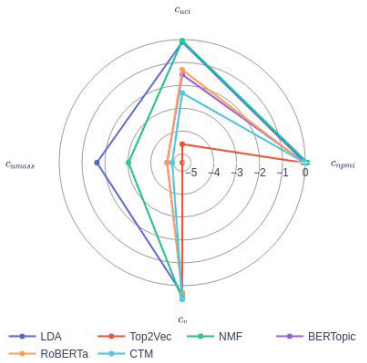
Considering the c_v as the closest coherence measure to human judgment, we can state that NMF produces more human interpretable topics as compared to others. However, the diverse shortcoming points to insightful implicit findings of the study that the coherence of topic models is significantly influenced by the type and size of the datasets along with the number of topics the model uses. This behavior can be observed in Tables 3, 4, 5, 6 and 7, where results are mentioned in detail. A overall view of evaluations are illustrated in Fig. 3.



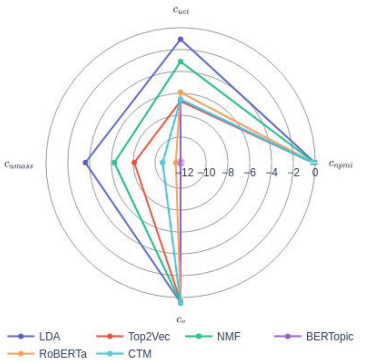
(a) *Comparison of the topic models for ATE dataset*



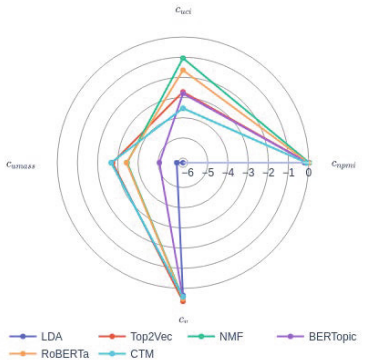
(b) *Comparison of the topic models for TAT dataset*



(c) *Comparison of the topic models for KU dataset*



(d) *Comparison of the topic models for TP dataset*



(e) *Comparison of the topic models for 20NG dataset*

Fig. 3. Evaluation of the topic models based on the results on each dataset

4 Conclusion

Our study delineates a comprehensive review of promising novel and devised topic models. These include LDA, NMF, and Top2Vec, BERTopic, RoBERTa, CTM, and ETM. Further, our study presents an in-detail experimental evaluation-based comparative analysis of these models in a touristic experiences context. The analysis is conducted based on topic coherence and topic diversity in terms of multiple significant parameters. We considered four topic coherence parameters: C_{uci}, C_v, C_{umass} and C_{npmi} along with two diversity parameters: Topic Diversity (TD) and Inverted RBO (IRBO). The experimental evaluations are conducted over five variant and contextually diverse datasets where four are related to touristic experiences, out of which three are exclusively designed for the purpose of this study. The study contributes significant conclusive quantitative results and reveals many valuable implicit deductions. The diverse quantitative findings of the study implicitly reveal that there is no conclusive winner among the considered models and the performance and suitability of the models are correlated to the size and type of data. For this reason, we have concluded the suitability of the models as per the mentioned attributes of the datasets. From Table 3, we observed that for ATE, NMF performs better as compared to others for 3 out of 6 parameters, C_{uci}, C_v and C_{npmi}, followed by LDA, Top2Vec and CTM which performed better for 1 parameter each, C_{umass}, TD and IRBO respectively.

Similarly, Table 4 illustrates results for TAT where NMF performs better on 3 out of 6 parameters, C_{uci}, C_v and C_{npmi}. While LDA also shows better performance for 3 out 6 parameters, C_{umass}, TD, and IRBO. Here LDA outperforms others majorly for diversity while NMF outperforms others majorly for coherence. Top2Vec produces equal TD as LDA for TAT and it also delivers better TD for ATE. Hence we conclude that the use of NMF is preferred for small to medium-sized datasets where document length is moderately shorter on average for better coherence, while Top2Vec or LDA delivers better diversity in such cases.

Further, from Table 5, we conclude that on average Top2Vec outperforms others for medium-sized datasets having multi-lingual documents. Since Top2Vec outperforms others for 3 out of 6 parameters, C_{umass}, TD, and IRBO, followed by NMF that outperformed others for C_v and C_{npmi}, we suggest the suitability of Top2Vec for such cases if moderate coherence is preferred along with high diversity. Conversely, NMF is preferred if good coherence is required irrespective of high diversity. Moreover, from Table 7 we conclude that Top2Vec performs better on average for large-sized English datasets as it delivers better results for 3 out of 6 parameters C_v, TD, and IRBO. Although Table 6 reveals that BERTopic outperforms others quantitatively for majority parameters (C_{umass}, TD and IRBO) for medium-sized datasets, however, RoBERTa exhibits considerably better qualitative aspects than BERTopic for such datasets with marginal difference in readings in terms of C_{umass}, TD and IRBO. Hence, we suggest the use of Top2Vec for large-sized English datasets and RoBERTa for medium-sized English datasets. In both cases, ETM may also be used if only the coherence

parameter is of concern since it delivers better coherence for both cases in terms of C_{uci} and C_{npmi}.

5 Open Issues and Future Research Directions

The diverse domain of touristic experiences causes heterogeneous issues such as the unavailability of versatile and diverse public datasets, changing tourist preferences, data multimodality, and more. Such current limitations can be interesting possible future directions of this study.

References

1. Alenezi, T., Hirtle, S.: Normalized attraction travel personality representation for improving travel recommender systems. IEEE Access (2022)
2. Angelov, D.: Top2vec: distributed representations of topics. arXiv preprint arXiv:2008.09470 (2020)
3. Bao, J., Xu, C., Liu, P., Wang, W.: Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. Netw. Spat. Econ. **17**, 1231–1253 (2017)
4. Bianchi, F., Terragni, S., Hovy, D.: Pre-training is a hot topic: Contextualized document embeddings improve topic coherence (2020)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2001)
6. Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. Trans. Assoc. Comput. Ling. **8**, 439–453 (2020)
7. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
8. Guo, Y., Barnes, S.J., Jia, Q.: Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent Dirichlet allocation. Tour. Manage. **59**, 467–483 (2017)
9. Hu, N., Zhang, T., Gao, B., Bose, I.: What do hotel customers complain about? text analysis using structural topic model. Tour. Manage. **72**, 417–426 (2019)
10. Kamal, M., Chatzigiannakis, I.: Influential factors for tourist profiling for personalized tourism recommendation systems–a compact survey. In: 2021 International Conference on Innovative Computing (ICIC), pp. 1–6. IEEE (2021)
11. Korenčić, D., Ristov, S., Repar, J., Šnajder, J.: A topic coverage approach to evaluation of topic models. IEEE Access **9**, 123280–123312 (2021)
12. Krumm, J., Davies, N., Narayanaswami, C.: User-generated content. IEEE Pervasive Comput. **7**(4), 10–11 (2008)
13. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems 13 (2000)
14. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
15. Lui, M., Lau, J.H., Baldwin, T.: Automatic detection and language identification of multilingual documents. Trans. Assoc. Comput. Ling. **2**, 27–40 (2014)
16. Vu, H.Q., Li, G., Law, R.: Discovering implicit activity preferences in travel itineraries by topic modeling. Tour. Manage. **75**, 435–446 (2019)

17. Yan, Q., Jiang, T., Zhou, S., Zhang, X.: Exploring tourist interaction from user-generated content: topic analysis and content analysis. J. Vacation Mark., 13567667221135196 (2022)
18. Zhao, N., Fan, G., Qi, Z., Shi, J.: Exploring the current situation of cultural tourism scenic spots based on lda model-take Nanjing, Jiangsu province, China as an example. Procedia Comput. Sci. **221**, 826–832 (2023)