Received: 1 October 2024 | Revised: 27 March 2025 | Accepted: 10 September 2025 | Published online: 22 April 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 4(2) 137–154
DOI: 10.47852/bonviewJDSIS62024472

BON VIEW PUBLISHING

Exploring Digital Tourism Through Topic Models: A Review and Experimental Study

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

[1]*Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy*

[2]*Amarena Company Srl, Italy*

Abstract: The surge in the volume and complexity of user-generated content (UGC) and data on digital tourism platforms has raise both opportunities and challenges for its automated analysis. Advanced topic modeling techniques are now necessitated to cater the variety, dynamism, and multifaceted nature of this data, yet their application in digital tourism encounters unique challenges. This study comprehensively reviews prominent and emerging topic models from the categories of probabilistic models, matrix factorization-based models, and neural embedding-based models, describing their systemic architectures and operational mechanisms. In the application context of digital tourism, the study follows an experimental evaluation of the models' performance on five datasets, across multiple coherence and diversity parameters. Results do not reveal optimality of a single model universally; rather, a model's effectiveness depends on size and structure characteristics of the data as extensively analyzed in this article. Additionally, the study presents quantitative and qualitative findings, implicit shortcomings along with conclusive deductions, digital tourism application related open issues of topic models, followed by future directions of research.

Keywords: topic modeling, text-mining, digital tourism, comparative analysis

1. Introduction

The recent evolution of technology and increased use of web applications have caused a significant rise in user-generated content (UGC) and the amount of data available on the web [1]. The significance of UGC in different application domains has been visibly identified, thoroughly explored, and examined over the past years [2, 3]. In tourism sector, UGC is particularly becoming an essential element across all stages including before, during, and after travel [4], where information provided by fellow travelers has become a key influence for other tourists [5]. Although, understanding the subject matter of the UGC is apparent for tourists. However, the massive amount of UGC posted on multiple online platforms coupled with diverse content published by online tourism agencies (OTAs) creates a complex network of associations, necessitating the content to be automatically interpretated and profiled without reliance on human interventions [6, 7]. The automation of these tasks can help with the organization of large-scale datasets for advanced services such as generation of personalized travel recommendations, the identification of the hidden semantics in customer satisfaction-related content, and the delivery of online advertisements based on user interests [8–10].

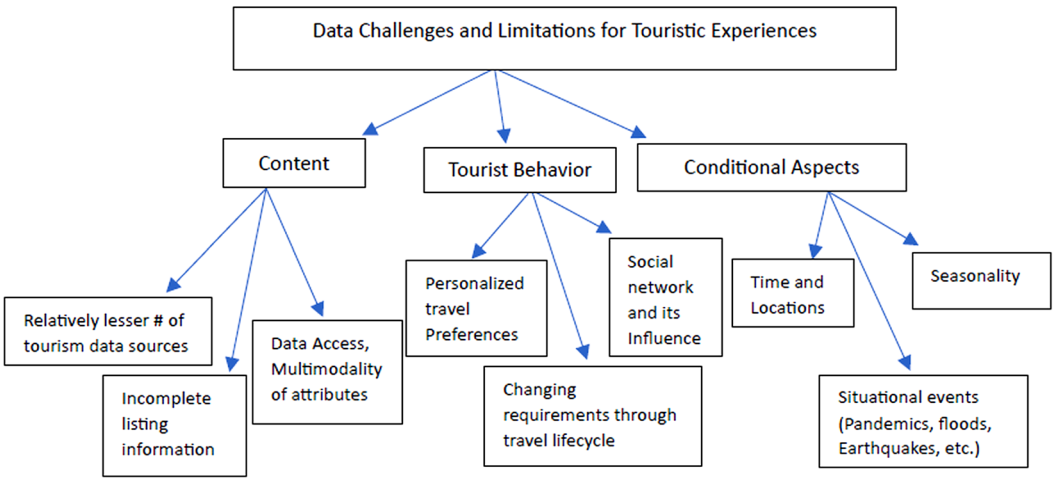
Topic modeling (TM) performs a crucial role to analyze and operate large volume of textual UGC including tourists' reviews and experiences for complex applications like tourism recommender systems. Topic modeling is a prominently acknowledged data mining technique that identifies potential latent topics for a set of documents by semantically relating words and documents [11]. It associates massive volume of unorganized UGC and the varied requirements for customized travel recommendations [11, 12]. This has made topic modeling one of the most sought-after techniques in tourism, where topics and labels are essential for linking diverse tourists' requirements to relevant offerings from tourism businesses, taking into account the reviews and UGC by travelers [13].

The scientific community has extensively explored and evaluated topic models for various domains including marketing and business management [14, 15], biology and medicine [16], and software traceability [17]. However, topic modeling and its potential for tourism industry remain under-explored. Note that topic modeling application is specifically distinct and intricate for data related to tourism. Some prominent challenges associated with tourism related content are illustrated in Figure 1.

As mentioned in Figure 1, tourism related data faces multiple challenges which hinders the application of automated analysis and utilization of strategies such as topic modeling. Tourism content is not only subjected to multimodality, incompleteness, seasonality, situational events but is also very dynamic considering the changing interests of tourists over time and influence of social media. Additionally, the content faces accessibility, privacy concerns, and limited availability concerns. Note that, for topic modeling on tourism data, the goal is to extract and identify topics representing latent sentiments, priorities, experiences, and anticipations of tourists. Simultaneously, diverse content produced by tourists and prominent presence of emotion-focused vocabulary are distinct from conventional opinion-oriented data such as blog posts. Additionally, in contrast to other expressive data types for instance Twitter like microblog services, documents representing touristic experiences are considerably detailed and longer, conversely, these are much precise and shorter in comparison to structured articles from journals or encyclopedias [18]. Such differences of type and structure

*Corresponding author: Ioannis Chatzigiannakis, Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy. Email: ichatz@diag.uniroma1.it

Figure 1
Data challenges and limitations in tourism



in the vocabulary and corpus significantly influence the effectiveness of topic models. Therefore, it is essential to assess and observe the conduct and effectiveness of topic models on tourism-related data and learn the rationale and factors for such performances.

To cater the concern, this article presents an in-depth exploration of topic models broadly used in the relevant bibliography along with very recent state-of-the-art models. Particularly, Latent Dirichlet Allocation (LDA) [19], Non-negative Matrix Factorization (NMF) [20], Topic to Vector (Top2Vec) [21], Bidirectional Encoder Representations from Transformers (BERTopic) [22], Robustly Optimized BERT Pretraining Approach (RoBERTa) [23], Contextualized Topic Model (CTM) [24], and Embedded Topic Model (ETM) [25] are studied in-depth. This set of models not only presents a comprehensive coverage for diverse paradigms, ranging from classical statistical methods to advanced neural network models, but also has proven robustness effectiveness across multiple recent domain oriented studies, as discussed in Section 3. Given this selection, the study presents a thorough review on topic models including summary of systemic architectures, principles, and operational mechanisms. The study also presents a delineated comparative analysis of the models on five domain exclusive datasets, where three out of five datasets are exclusively established for this study while two are public datasets. Based on experimental evaluations for multiple coherence, diversity, and quality parameters, the study discusses the quantitative and qualitative performance analysis of the models and identifies the potential reasons for their certain outcomes. In this sense, our study differentiates from other review articles that focus either on providing a detailed presentation of the methods based only on their theoretical foundations or those articles that focus on other thoroughly investigated different application domains.

The rest of the paper is organized as follows. In Section 2, we have discussed preliminaries and important concepts related to embedding models and topic modeling along with a brief overview of the basics. In Section 3, we have provided a comprehensive literature review with a summary of prior related studies over past 10 years and an in-depth review of topic models including their underlying mechanisms. This review covers both the theoretical foundations and practical implementations of selected topic models. In Section 4, we introduce the datasets and evaluation parameters, followed by experimental exploration of the models, the results and shortcomings of the analysis along with validation of findings and thorough detailed explicit and implicit findings-based discussion. In Section 5, we present the summary and conclusion, followed by current gaps and limitations of topic models when applied in the tourism domain along with future orientations of research.

2. Definition of Terms and Preliminaries

In this section, we provide definitions of some important terms, notations, and basic concepts involved in topic modeling. Note that a text-based dataset is composed of a set of documents (D) which are strings of variable length composed of N words. Here a word (W) or term (T) is considered as the fundamental unit of a sample data. The set of distinct words present in a dataset forms the vocabulary (V) and a topic is then interpreted as a probability distribution over this fixed vocabulary, representing a label for a cluster of documents from a given dataset. Topic models are significantly influenced by the representation of words and documents in a corpus. Traditionally, topic models operate on vector representation of words and documents for input, known as Word Embedding and Document Embedding, respectively. These embeddings are, usually, real-value vectors, representing words or documents in vector space, in such a manner that similar words or documents are positioned closer to each other in spatial proximity. A summary of basic notions is given in Table 1.

We have briefly defined the classification of word embedding and representation techniques based on the study by Selva Birunda and Kanniga Devi [26], as follows:

Traditional word embedding, or count-based embedding: This class comprises methods that use frequency of words, co-occurrence of words, and rarity of words for document representation [27]. A classic representation of documents from this category is Bag of words (BoW). In BoW, each document is described by a vector of dimension equal

Table 1
Notation definition

Scale	Mean interpretation
D	Set of documents
d	Single document
V	Vocabulary
W	Single word
T	Single term
BoW	Bag of words representation
TF-IDF	Term Frequency-Inverse Document Frequency
θ	Topic-document distribution
φ	Term-topic distribution

to the vocabulary size, where each dimension represents the number of times a certain word appears in a document. However, such a text representation has limitations: the vectors tend to be very sparse, addition of new documents having unknown vocabulary may cause scalability concerns, and the context is not considered. Another well-known vector representation method from this class is Term Frequency-Inverse Document Frequency (TF-IDF) where TF measures how frequently a word appears in a document and IDF how much importance weight it carries. Note that IDF is introduced to suppress the impact of terms occurring with prominent frequency across many documents; this also helps to magnify the impact of terms that occur rarely and are important. TF-IDF can be estimated using Equation (1) as follows:

$$tfidf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \times log \frac{|D|}{1+|\{d \in d: t \in d\}|} \quad (1)$$

where $f_{t,d}$ is the count of term t in the document d and D is the dataset. The ith document is then represented as:

$d_i = [tfidf_{0,i}, \ldots, tfidf_{N,i}]$, where N is the number of words in the vocabulary V.

Static word embedding: This category of embedding representation involves prediction-based methods that compute probabilities of the occurrence of the words and map those into fix-sized vectors. These embedding representations do not consider the context, implying that a word embedding remains unaffected irrespective of the different meanings it may convey in different sentences. Consequently, two words frequently appearing together will attain similar embeddings, even if they differ in meanings in different contexts. This category of methods gained popularity after the release of Word to Vector (Word2Vec) [28]. Word2Vec represents words into numeric vectors and also learns word association from the corpus. It may utilize either of its two architectures: (a) Continuous Bag-of-Words (CBOW) and (b) Skip-gram. While CBOW predicts one target word from the surrounding context words, Skip-gram, on the other hand, uses one target word to predict surrounding context words. The Word2Vec method has been used to design Doc2Vec, a well-known document embedding representation method that creates a numeric representation of a document, regardless of its length [29].

Contextualized word embedding: Since context is considered in this class of methods, the representation of a word dynamically varies based on the surrounding words. Methods that use this class of representation, such as Transformers-based embeddings, are considered state-of the-art for most Natural Language Processing (NLP) tasks. These approaches are context-dependent, indicating they can disambiguate polysemes, thanks to the attention mechanism [30]. Embedding representation methods from this category can compute different embeddings for a word depending on its context. BERT is one of the well-appreciated methods from this category [31]. It has been adopted for several applications in NLP [32, 33] and with multiple variations [34]. An interesting variation of BERT used in topic models is Sentence Transformers (SBERT) which uses siamese and triplet network structures to generate embedding representation for sentences [35].

3. Literature Review

3.1. Recent studies

Previously, a number of important studies have focused on topic modeling in addressing digital tourism challenges. For instance, topic models are adopted to identify preferred destinations in itineraries, understand tourists' opinions, and to generate recommendations. As our study focuses in applying topic modeling in the context of tourism domain, the summary of some notable related studies from the past 10 years of literature is presented in Table 2.

Although topic modeling finds initial roots in the 1980s [36], it gained prominence in the 1990s due to appreciable performance recorded by topic models such as Latent Semantic Analysis (LSA) [37], NMF [20] and, in a particular way, LDA [19]. Over the past two decades, models such as LDA have been used to devise various other promising models such as those proposed in Yu et al. [38] as well as EkiNci and Omurca [39]. However, despite their success, conventional Bayesian probabilistic topic models started to show signs of fatigue and could not meet the expectations of big data handling in the era of big data and deep learning [40]. Instead, models based on deep learning are attaining more popularity and appreciation. Deep learning-based models are now applied for topic modeling, document representation [41], computing semantic representations of topics, and dealing with short texts [42].

3.2. Selected models

Topic generation and representation prominently relies on dimensionality reduction techniques adopted by topic models to convert high-dimensional text documents into lower-dimensional topic representations [43]. Hence, for the purpose of review and analysis, we selected seven well-known topic models, categorized based on three key dimensionality reduction approaches: probabilistic models, matrix factorization-based models, and neural embedding-based models. The selected models include LDA, NMF, Top2Vec, BERT, RoBERTa, CTM, and ETM, because of their promising performances and recent widespread adoption in related studies [44–46].

3.2.1. Latent Dirichlet Allocation

LDA is a generative probabilistic model, designed for a given corpus of text documents [19]. The model works on the De Finetti theorem and considers that K latent topics exist in the given N documents corpus, where a multinomial distribution represents each topic over the M words in the vocabulary extracted from the document corpus. It assumes that a document consists of sampling a variant proportional mixture of these topics and the topics sample various words representing those topics. Precisely, the algorithm, in a nutshell, is illustrated as follows:

1) For the ith document d_i in the document corpus D, (where i = 1,2,...,N), choose $\theta_i \sim$ Dirichlet(α).
2) For each word wi,m in the document d:
   a. Draw topic $z_{i,m} \sim$ multinomial(θi)
   b. Estimate topic distribution $\varphi_{zi,m} \sim$ Dirichlet(β)
   c. Estimate word wi,m ~ multinomial ($\varphi_{zi,m}$)

Here α and β are Dirichlet hyperparameters. These are used to estimate the probability of document corpus D using Equation (2) as follows:

$$P(D|\alpha, \beta) = \prod_{i-1}^{N} \int P(\theta_i|\alpha) F(\theta, \varphi) d\theta_i \quad (2)$$

By maximizing the probability in Equation (2), the model learns topic-document distribution θ and term-topic distribution φ, thus generating suitable topics for documents. The model considers the following assumptions for its processing:

1) Each document is an unordered collection of words, namely bag-of-words (BOWs). This indicates that the model does not consider the grammatical and contextual structure of the sentences.
2) Number of topics is pre-decided. This indicates that the model takes a number of topics as input and assigns topics to documents accordingly. This may vary for a different number of topics.
3) The assignments of topics to documents and words to topics are random and the updates are iterative. This assumes that all topic assignments except the current word are correct.

Table 2
Literature summary using topic modeling in the tourism field

Study	Objectives	Model(s) used	Model(s) category	Data source	Evaluation metrics
[47]	Locations recommendations	Geo Topic Model	Probabilistic	Tabelog and Flickr	5-best accuracy
[48]	Travel recommendations	Author Topic Model	Probabilistic	Flickr	MAP
[49]	Rating prediction and recommendation, suggest ratings for reviews and interpretation of users and items	LDA, Topic-Sentiment Criteria	Probabilistic	TripAdvisor, Yelp	RMSE, two-sample Kolmogorov-Smirnov test
[50]	Analysis of user satisfaction	LDA	Probabilistic	TripAdvisor	human analysis, Jaccard coefficient, and Stanford Topic Modeling Toolbox
[13]	Travel itineraries analysis	LDA	Probabilistic	Twitter, Foursquare	Topic Concentration, and perplexity
[51]	Visitor's perception mining	LDA	Probabilistic	TripAdvisor	Qualitative analysis
[52]	Guest satisfaction identification	LDA	Probabilistic	Booking.com, Hotels.com, Agoda	Chi-square goodness of fit tests, Wilcoxon signed rank tests, two proportions z-tests
[53]	Analyze news discourse role in forecasting tourism arrivals	STM	Probabilistic	Hong Kong Tourism Board	Cointegration tests, granger causality tests, MAE, RMSE, MAPE, RMSPE
[54]	Analyze tourists culture related dining experiences	LDA	Probabilistic	TripAdvisor	Visual analysis
[55]	Restaurant recommendation to tourists	NMF, LDA	Matrix Factorization, Probabilistic	Yelp	Salience, valence
[46]	Category travel personality representation	LDA, ETM, and Top2Vec	Matrix Factorization, Probabilistic, and Neural Embedding	TripAdvisor	RMSE, NDCG@K
[56]	Opinion aspects extraction by customer-evaluated parameters	BERTopic	Neural Embedding	Tripadvisor, IRecommend, and Otzovik	Precision, Recall
[57]	Connect computational linguistics with historical methods	NMF	Matrix Factorization	ProQuest Historical Newspapers, GALE Primary Sources, New York Times	Qualitative analysis
[58]	Generate comprehensive destination image	BERTopic	Neural Embedding	Google	UMass Coherence
[59]	Analyze link between content type and its engagement for hotels	LDA	Probabilistic	Twitter	C_v Coherence

3.2.2. Top2Vec

Top2Vec is a neural-embedding based model that uses text data vectorization to identify semantically similar documents, words, or sentences within joint embedding spatial proximity [21, 54]. As word vectors that appear semantically nearest to the document vectors best describe the documents' topic, the number of document clusters represents the number of topics, where each topic is represented by multiple closest words. In short, it leverages joint document and word semantic embedding to find topic vectors.

Mathematically, the general representation of Top2Vec can be summarized as follows:

The word embedding training is conducted to maximize the observation likelihood of word w_i given document d_j:

$$\max \sum_{i,j} \log P(w_i|d_j) \tag{3}$$

where $P(w_i|d_j)$ is the conditional probability of word w_i given document d_j.

Topic centroid c_T is calculated as the average of the document embeddings in the cluster C:

$$c_T = \frac{1}{|C|} \sum_{d_j \in C} d_j \tag{4}$$

where $|C|$ is the number of documents in the cluster, and d_j is the document j embedding vector.

Cosine similarity between a word embedding w_i and a topic centroid c_T is computed as:

$$\text{CosineSimilarity}(w_i, c_T) = \frac{w_i^\top \cdot c_T}{|w_i||c_T|} \tag{5}$$

where $w_i^\top c_T$ is the product of the word embedding w_i and the centroid c_T, and $\|w_i\|$ and $\|c_T\|$ are their respective Euclidean norms.

The model makes the following assumptions:

1) It considers joint document and word vectors, keeping the track of semantics rather than BOWs.
2) It automatically suggests the number of topics.
3) It does not require data pre-processing such as stopwords removal, lemmatization, and stemming.

3.2.3. Non-negative Matrix Factorization

NMF is an unsupervised matrix factorization based model that operates on linear algebra to transform the high-dimensional data into a reduced semantic space with non-negative hidden matrix structures [20, 60]. It works on the *TF-IDF* transformed data and decomposes the term-document matrix A, a form of the original document matrix, into the product of *W* and *H*, that are two matrices as denoted in Equation (6):

$$A = WH \tag{6}$$

where *W* and *H* are positive matrices such as $W \geq 0$, and $H \geq 0$. Here *W* represents terms mapped to topics and H represents topics mapped to documents.

Equation (7) shows that the weighted sum of the components in matrix *A* is:

$$A_i = \sum_{j=1}^{k} W_{ij} * H_j \tag{7}$$

The values of *W* and *H* are updated iteratively as follows:

$$W \leftarrow W \frac{AH^T}{WHH^T} \tag{8}$$

$$H \leftarrow H \frac{W^T A}{W^T WH} \tag{9}$$

The model iterates the above Equations (8) and (9) until it achieves convergence then achieves final term–topic matrix W and topic–document matrix H for topics extraction.

The model works on the following assumptions:

1) Considers original documents as a matrix that is an inner product of two matrices, say *W* and *H*. Here *W* represents the Documents-Topics matrix, while *H* represents the Topics-Terms matrix.
2) Considers non-negative matrix values.
3) It requires pre-defining of a number of topics as input
4) It requires data pre-processing such as stopwords removal, lemmatization, special characters removal, and stemming.

3.2.4. BERTopic

Proposed in 2023, BERTopic is a recent promising neural embedding-based topic modeling approach that uses BERT embeddings and transformer embeddings [22]. It is similar to Top2Vec regarding its algorithmic structure. BERTopic provides embedding extraction for the document corpus with a sentence-transformers model for more than 50 languages. Similar to Top2Vec, BERTopic also offers dimensionality reduction using Uniform Manifold Approximation & Projection (UMAP) and then clusters the documents using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). However, unlike Top2Vec, it applies a variation of TF-IDF, Equation (1), called class-based term frequency-inverse document frequency (cTF-IDF), shown in Equation (10). This variation efficiently evaluates the significance of terms within a cluster or class followed by the creation of term representation [61]. Here, the higher score a term gets, the better it represents its topic [62].

$$\mathrm{cTF-IDF_i} = \frac{\mathrm{t_i}}{\mathrm{w_i}} \times \log \frac{\mathrm{m}}{\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{t_j}} \tag{10}$$

where, *t* is the frequency of each word for each class *i*, w is the total number of words, and m is the total number of documents being divided by the total frequency of word *t* across all classes *n*.

BERTopic offers continuous instead of discrete topic modeling, which makes it different from other approaches [63]. The model leads to different results with repeated execution due to its stochastic nature. The model offers the following features:

1) It does not require a number of topics in advance. It estimates the number of topics automatically.
2) It offers several multi-lingual models to extract document embeddings. Usually, in practice, it uses sentence-transformers package with two default models; Distilbert for English and Cross-Lingual Language Model with RoBERTa architecture (XLM-R) for any other language. The XLM-R models support 50+ languages.
3) The approach mentions outliers in the resulting output as Topic 0 with the label of -1.

3.2.5. Robustly Optimized BERT Pretraining Approach

RoBERTa is a devised strategy from the BERT embedding model, known to be a robustly optimized variant of the BERT model [23]. It is a transformer-based neural embedding model that takes into consideration the context of a given word for each occurrence. RoBERTa uses a dynamic version of BERT's masking strategy [31], where the model learns to predict hidden sections and topics for the text documents and modifies key hyper-parameters of BERT. The model, like BERT, encodes substantial information about lexical semantics.

In comparison to BERT, RoBERTa is equipped with dynamic mask generation, full sentences without Next Sentence Prediction (NSP) objective, larger batches, and a larger byte-level byte pair encoding (BPE). It has been trained for longer and on a bigger dataset. The original study of RoBERTa found it to be outperforming BERT and eXtreme Language Net (XLNet); however, it is interesting to observe how it performs in the context of touristic experiences, which is the scope of this study.

3.2.6. Contextualized Topic Model

CTMs are devised from the Neural Product of Dirichlet Latent Dirichlet Allocation (Neural-ProdLDA) variational autoencoding approach and pre-trained embedding models [24]. It aims to combine traditional topic models with contextual embedding models such as BERT or RoBERTa. Mathematically, CTM can be generalized as:

Each document d_i is represented as a combination of its contextualized word embeddings z_i by contextual embedding model represented as CONTXT and a latent topic distribution θ_i as shown in Equation (11):

$$z_i = \mathrm{CONTXT}(d_i) \tag{11}$$

CTM considers that each document d_i has a latent topic distribution θ_i, modelled by a variational distribution, as shown in Equation (12):

$$\theta_i \sim \mathrm{Dirichlet}(\alpha) \tag{12}$$

For each topic *k*, topic distribution θ_i and the topic-word distribution β_k generate words in the document using Equation (13):

$$P(w_j|\theta_i) = \sum_{k=1}^{K} \theta_{ik}\beta_{k,j} \quad (13)$$

where β_k denotes distribution of words for topic $_k$, and θ_{ik} is the topic proportion for document d_i. The two major categories of CTM include Combined Topic Model (CombinedTM) and Zero-Shot Topic Model (ZeroShortTM).

CombinedTM uses contextual embeddings, SBERT, with the BOW to produce coherent topics. The framework trains a neural inference network that maps the BoW document representation into a continuous latent representation. Then, a decoder network reconstructs the BoW by generating its words from the latent document representation. A hidden layer represents documents with the same dimensions as the vocabulary size and the BOW representation. On the other hand, ZeroShotTM is a variation of CTM that works for missing words in data and also offers multilingual topic modeling (if trained with multi-lingual embeddings) [64]. It is a neural variational topic model that combines deep learning-based topic models with embedding techniques such as SBERT. Once the model is trained by reconstructing BOW from a neural network, it can generate the representations of the documents and predict their topic distributions even for the unknown words in test data. Although CTMs are a promising addition, however, these have some constraints including the maximum size of BOW (not to be more than 2000 elements), multi-lingual model not being trained on English data and pre-processing required to generate BOW.

3.2.7. Embedded Topic Model

The ETM is a generative topic model devised from LDA [25]. It combines LDA with a variational auto-encoder (VAE). The basic idea is to optimize and use LDA with word embeddings (Word2Vec). It produces word embedding similar to the CBOW word embeddings. However, ETM uses an assigned topic vector instead of a context vector. ETM offers two versions, native ETM which learns its own topics and word embeddings, and ETM SG which uses pre-trained word embeddings.

ETM functions in a simple manner. It uses categorical distribution to model each word. The parameter for each modelled word is the inner product between a word embedding and its assigned topic embedding. The fitting of the model uses amortized variational inference algorithm. The generative process ETM for a dth document can be summarized as follows, where LN(.) represents the logistic normal distribution:

1) Draw topic proportions $\theta_d \sim$ LN(0,I).
2) For each word n in the document:
   a. Generate topic assignment $z_{dn} \sim$ Cat(θ_d).
   b. Generate $w_{dn} \sim$ softmax($\rho^T \alpha_{zdn}$)

Note that the initial steps of the approach, 1 and 2a, are similar to traditional LDA. The difference can be found in step 2b, where the model uses vocabulary embedding ρ and assigned topic embedding α_{zdn} to get the words from the topic zdn.

4. Comparative Evaluation

In this section, we have presented the experimental evaluation and comparative analysis of the considered topic models, that includes LDA, Top2Vec, NMF, BERTopic, RoBERTa, CTM, and ETM. The comparison is performed using 1 generic dataset and 4 tourism-focused datasets, out of which 3 are collected exclusively for this study. The statistical summary of the datasets is mentioned in Table 3. We have introduced the datasets and evaluation parameters in the following subsections.

Table 3
Statistics of the datasets

Dataset labels	# of docs	# of words	Vocabulary size	Avg. words per doc
ATE	737	126,450	2629	68
TAT	2765	284,050	4555	152
KU	5724	1,556,416	138,095	272
TP	8000	191,996	27,012	24
20NG	18,846	3,423,145	29,548	182

4.1. Datasets

In this subsection, we introduce the five datasets we have used in our analysis. The datasets were particularly selected for rigorous evaluation of topic models across a diverse characteristic of data within the context. Three of the tourism-oriented datasets deliver variation of size and structure of the data. One dataset adds a multi-lingual dimension, while one dataset offers a broad, generic benchmark. Our selection of datasets ensures a thorough analysis of method effectiveness within the context of tourism with varied thematic environments. The datasets include the following:

20NewsGroup (20NG) is a well-established generic benchmark dataset having more than 18,000 newsgroup articles based on 20 different topics. The dataset is primarily in the English language and is versatile to serve a split for training and testing data. It has been widely used to evaluate topic models in many studies such as Churchill and Singh [65] as well as Taylor and du Preez [66].

TourPedia (TP) is a publicly available dataset related to tourism attractions and reviews about those attractions. The attractions include accommodations, restaurants, and points of interest. The dataset contains more than 490,000 places and 577,000 reviews. It consists of data for 8 cities: Amsterdam, Barcelona, Berlin, Dubai, London, Paris, Rome, and Tuscany. TourPedia was contributed by the project OpeNER, funded by the 7th Framework Program of the European Commission [67]. It has been used in many data analysis studies such as Mishra et al. [55] as well as Patel and Urolagin [68].

TripAdvisor Tourist Activities (TAT): It is one of the exclusive datasets we collected for this study. Applying web-scrapping to TripAdvisor, we collected the data of all the touristic activities of the tourism destinations in Rome, Italy. The activities are extracted from the "Things to do" section of the website. The dataset contains 2765 entries and each entry contains text data related to 7 attributes, including an activity's title, description, popular mentions, price, duration, ratings, and itinerary.

AirBnB Touristic Experiences (ATE): We collected a dataset from AirBnB which consists of data related to touristic experiences mentioned on the Airbnb website. The data is mined from the "Experiences" module of the web portal for the region of Rome, Italy. This dataset is based on 737 records and each record is about a touristic experience published on Airbnb. Each record holds textual data related to 8 attributes: title, description, price, ratings, number of pictures, location, number of reviews, and video availability.

KuriU (KU): To explore the plurilingual aspect of the models, we have devised a distinctive dataset based on the Italian Language. It has 5724 entries, each having 30 attributes such as id, document type, title, description, locations, duration, images, distance, publishing date, and more. The dataset consists of data related to tourist services and POIs, for the Italian touristic experiences. The dataset is obtained from the beta testing phase of the KuriU application. KuriU is a research project oriented for touristic recommendations for experiences, integrating this study as one of its modules.

4.1.1. Data preprocessing and preparation

Data preprocessing is an important phase of many topic models [69]. Some topic models work on the principle of "Garbage in garbage out", so it is significantly crucial to learn what a model feeds on. Suitably preprocessed data will get the best out of a topic model while inappropriately preprocessed data may fail the performance of even a highly well-performing topic model. Hence, in this subsection, we mention the stages of data pre-processing applied to the datasets for each model as per its requirements. Table 4 shows a summary of the data preparation steps of each technique.

Note that the context of our study requires nouns as topics rather than adjectives or verbs. For instance, a topic such as "Museum" or "Cuisine" is a more insightful topic for tourist interest detection rather than a topic such as "Beautiful" or "Walking". Hence, data is processed in such a way for the models which require pre-processing. Moreover, since some methods included in the study such Top2Vec, BERTopic, RoBERTa, and CTM are recommended to be used without data preprocessing, no pre-processing is applied to datasets for these models.

For the purpose of experimentation, we considered English language documents for 4 out of 5 datasets. Hence, from the devised datasets, AirBnB Touristic Experiences (ATE), we considered 611 documents that are in the English language, and from TripAdvisor Tourist Activities (TAT), we considered 1860 documents that are in the English language. To analyze the behavior of the models on a multilingual aspect, all 5724 documents from the Italian Language dataset, KuriU (KU), are considered. On the other hand, all documents are considered from the benchmark datasets, that is, 18,846 documents from 20 Newsgroup (20NG) and 8000 documents from TP. We have considered the description text of all the documents for the purpose of analyzing the topic models.

4.2. Evaluation parameters

4.2.1. Topic diversity

Topic diversity (TD): It is a significantly impactful evaluation parameter to evaluate the generated topics by a TM [70]. It estimates the uniqueness of the document clusters generated by the TMs. TD has been used in multiple studies to support the evaluation, including those by Azarbonyad et al. [71] and Hashimoto et al. [72]. It estimates the percentage of constituent unique words in given K top words for all topics, as illustrated in Equation (14). The value of TD score spans from 0 to 1, reflecting that a score approaching 1 indicates greater topic diversity while a score near 0 shows a lesser topic diversity. A topic model generating greater topic diversity is preferred for a considered dataset.

$$\mathrm{TD} = \frac{n(U)}{K \cdot n(T)} \tag{14}$$

Equation (14) shows $n(U)$ as cardinality of the set where U represents unique words. K represents Top K words from all topics. T represents the set containing all topics produced by a model having $n(T)$ as its cardinality.

Inverted RBO (IRBO): Another interesting parameter used to evaluate the diversity of topics is IRBO. It is a recently introduced metric that has already been used in several works to estimate the quality of topics, as demonstrated in Murakami and Chakraborty [42], Carbone and Sarti [73] as well as Terragni and Fersini [74]. It estimated the degree of variation in topics [75]. Its value varies from 0 to 1, where 0 represents entirely similar and 1 shows fully distinct topics. It uses Ranked-Based Overlap measure [76] and computes how disjoint are the topics based on word ranking for top K words. We decided to use this metric because, differently from the standard topic diversity measure, here topics having common words at different ranks are assigned lesser penalty than the ones having common words at the top ranks [77].

Table 4
Preprocessing on datasets

Models	Stopwords removal	Lemmatize	Removal of punctuations, special characters, hashtags, emojis, URLs, numbers	Part of speech
LDA	Yes	Yes	Yes	Nouns
Top2Vec	No	No	No	All
NMF	Yes	Yes	Yes	Nouns
BERTopic	No	No	No	All
RoBERTa	No	No	No	All
CTM	No	No	No	All
ETM	Yes	Yes	Yes	Nouns

4.2.2. Topic coherence

Topic coherence evaluates how interpretable and coherent are the topics generated by a model in relation to the considered data [78, 79]. The idea is based on the distributional hypothesis of linguistics. Unlike perplexity and predictive likelihood, which can be contrary to experts' judgment [80], the versions of topic coherence we are using are considered the best approximation for human rating [78] and have been practiced in many studies, including those by Syed and Spruit [81], O'Callaghan et al. [82] and Bellaouar et al. [83].

Note that a greater reading of topic coherence exhibits better outcomes of a topic model in regard to generating interpretable topics. We have used the following variants of the topic coherence, for the purpose of evaluation. For each N top words from a topic cluster, $P(w_i, w_j)$ illustrates the probability of appearing together of words w_i and w_j, while $P(w_i)$ and $P(w_j)$ indicate the probability of these words occurring individually. The details of these measures can be referred from Röder et al. [78].

1) C_{uci} uses sliding window and the pointwise mutual information (PMI) of all word pairs for top words as shown in Equation (15).

$$c_{uci} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) P(w_j)} \tag{15}$$

2) C_v uses sliding window, top words' one-set segmentation with an indirect confirmation measure, using cosine similarity with normalized pointwise mutual information (NPMI) using Equations (16) and (17) set of equations:

$$\vec{v}(w') = \left\{ \sum_{w_i \in W'} \left(\frac{\log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) P(w_j)} \right)}{-\log(P(w_i, w_j) + \epsilon)} \right)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{16}$$

$$\Phi_{s_i}\left(\vec{u}, \vec{w}\right) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\left|\vec{u}\right|_2 \cdot \left|\vec{w}\right|_2} \tag{17}$$

In Equation (16), the context vector $\vec{v}(w')$ uses NPMI for all the word pairs. γ places more weight on larger NPMI values. In Equation (17), Φ is the confirmation measure that measures the vector cosine similarity of all the context vectors.

3) C_{umass} uses the count of document co-occurrences, one-preceding segmentation, and confirmation measure (logarithmic conditional probability), following the computation from Equation (18).

$$C_{umass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log\left(\frac{P(w_i, w_j)+\epsilon}{P(w_j)}\right) \tag{18}$$

4) C_{npmi} is an enhancement of the C_{uci} measure that utilizes NPMI.

4.3. Method

This article subsection documents the setup and methodology used to conduct experimental exploration. We used Python version 3.9.7 on Google Colab and Jupyter Notebook for experiment implementation. The evaluation of coherence parameters is conducted using Gensim toolkit, while Octis toolkit is used to estimate topic diversity parameters. The experiments comprise ten iterative runs for each model and the average recorded for each experiment shows the results illustrated in this section.

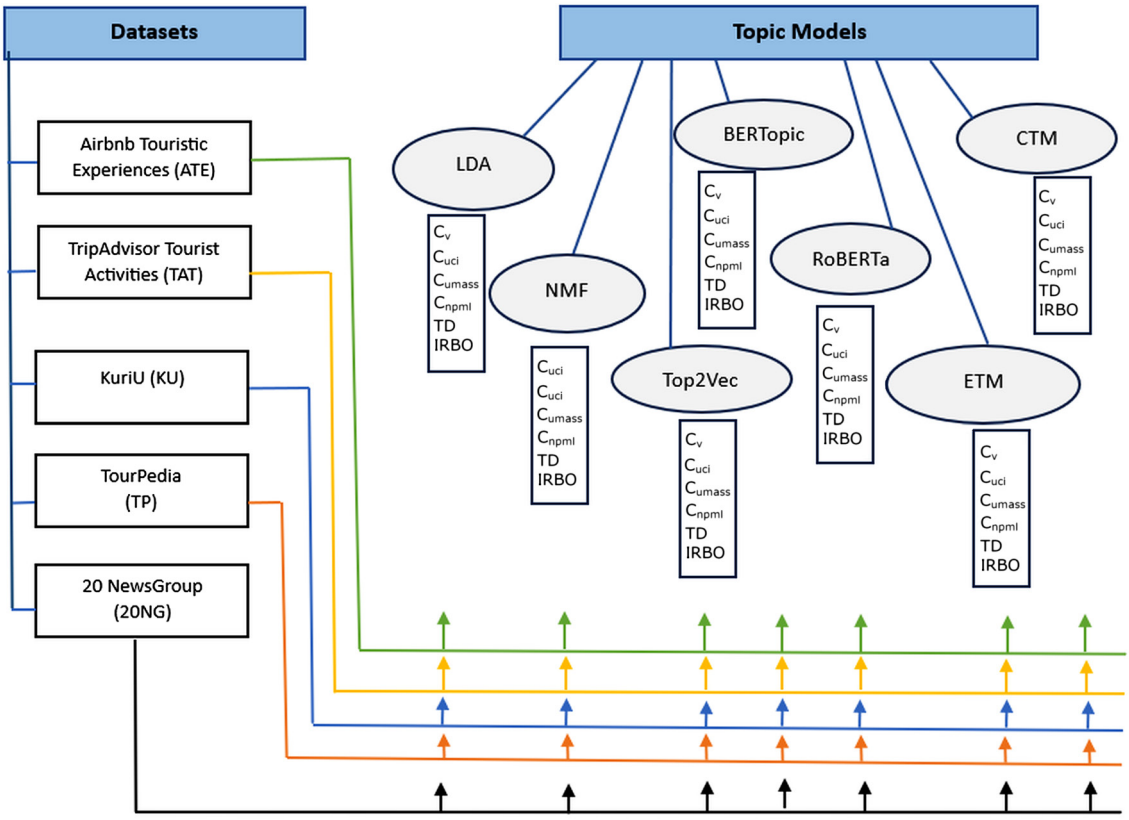
The workstation is equipped with Intel(R) Core(TM) i5-10210U CPU functioning at 1.60GHz having boost 2.11 GHz and 20GB of RAM. Default text embedding models have been used for experimentation respectively for each TM: roberta-base-nli-stsb-mean-tokens used for RoBERTa, Doc2Vec used for Top2Vec, and *all-MiniLM-L6-v2* used for BERTopic (English datasets). Additionally, *paraphrase-multilingual-MiniLM-L12-v2* is used to process Italian language dataset. Moreover, elbow method is opted to pre-decide the number of topics for LDA, NMF, CTM, and ETM as used in Vijayan [84] and Kirilenko et al. [10], while Top2Vec, BERTopic, and RoBERTa topics use embeddings-based clustering to define the number of topics, automatically. Note that while reflecting their design philosophy, there may be a reduced number of topics due to attempted balance of coherence and graduality by Top2Vec, BERTopic, and RoBERTa. The procedural flow of our experimentation for comparative analysis is illustrated in Figure 2.

4.4. Results

4.4.1. Topic diversity

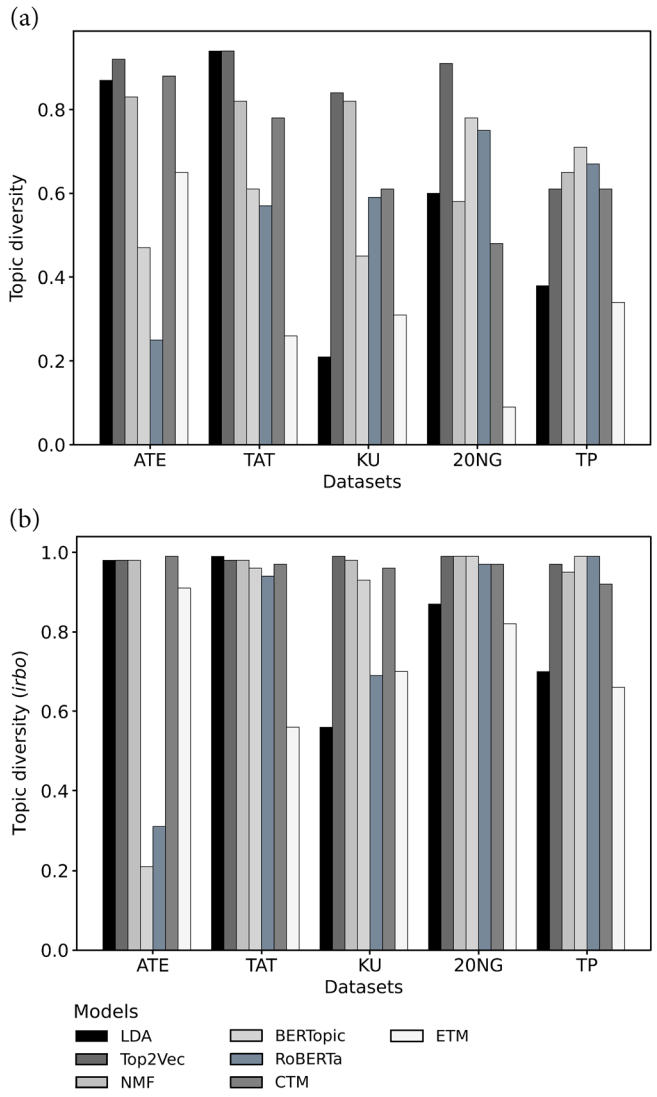
An interesting quality indicator analyzed in this study is diversity of the topics. A well-regarded model is required to generate greater topic diversity for a reasonable number of topics. Figure 3 presents the results acquired in this context, where Figure 3(a) shows models' comparison considering average TD, while Figure 3(b) presents IRBO readings on average obtained respective to each dataset. Notably, on average, a greater diversity is recorded by Top2Vec, for all datasets, considering both cases. Additionally, Figure 3(a) shows a notable finding regarding the TP dataset, which demonstrates a lower variability of topic diversity across models, while BERTopic exhibiting better performance in this instance. Similarly, from Figure 3(b), it is interesting to note that RoBERTa and BERTopic generate reasonably lower IRBO when implemented for a small-sized dataset with shorter document length such as ATE. Although Top2Vec has shown higher topic diversity on average, however, it is important to note that the number of topics (clusters) it produces is also considerably reduced for most of the datasets (Figure 3(a)). This shows a greater diversity within a cluster of a topic which is not preferred for a good topic model.

Figure 2
Concept model of comparative analysis



Note: *C_v, C_{uci}, C_{umass}, C_{npmi} are four Coherence Evaluators. TD and IRBO are Diversity Evaluators.

Figure 3
Topic model evaluation based on diversity metrics



4.4.2. Topic coherence

We analyzed coherence evaluators including C_{uci}, C_v, C_{umass} and C_{npmi}, to determine how well semantically coherent topics are produced by each topic model for the datasets under consideration. Figure 4 depicts a comparative evaluation of all the models for the given datasets for each coherence parameter. Notably, higher values of coherence score indicate more coherent topics, except for C_{umass}, where a better coherence is signified by lesser score value, as per Gensim implementation [85].

Notice from Figure 4(a), for relatively smaller-sized datasets (ATE and TAT), NMF exhibits preferable C_{uci}. However, as datasets grow in size, ETM initiates to present improved outcomes. On average ETM delivers better coherence as compared to other models, for C_{uci}. Figure 4(b) shows the case of C_v coherence, where, on average, NMF generates more coherent topics for 3 datasets out of 5. However, its performance deteriorates when applied on the largest considered dataset, 20NG, where Top2Vec outperforms others for C_v. This indicates sensitivity of NMF for dataset size, where it is considered suitable for small to medium-sized datasets in regard to C_v.

An interesting observation can be made from Figure 4(c) for C_{umass}, where, on average, LDA performs better than other models, while better results are exhibited by Top2Vec explicitly for KU — the Italian language dataset. Even though better results for C_{umass} on average are shown by LDA, it is observed that BERTopic outperforms others for TP — the medium-sized English dataset. This recommends LDA when dealing with small and large-sized English datasets while referring C_{umass} for coherence. Top2Vec may be selected when processing medium-sized datasets with multi-lingual nature, such as KU, while suggesting BERTopic for English based medium-sized datasets when C_{umass} is under consideration. Observably, another intriguing finding from Figure 4(d) shows better performance of NMF for C_{npmi} for nearly all datasets (except for TP) and as a whole on average. While ETM outperforms other models for TP, in regard to C_{npmi}. Furthermore, ETM and NMF show similar readings for the dataset 20NG. Hence, it is deducible that when C_{npmi} is focused, NMF is found to perform better for small-sized to medium-sized datasets, while for medium-sized to large-sized datasets, ETM is well-suited. Since C_v is regarded as the best approximated coherence measure to human judgment [78, 81], we can deduce that topics generated by NMF show more human interpretability in comparison to other models under consideration.

Additionally, the diverse shortcomings of this study reveal implicit insightful findings depicting coherence of topic modeling techniques to be prominently dependent on the size and type of the datasets as well as the number of topics utilized by a model. This trend is demonstrated across Tables 5–9, where results are presented thoroughly.

4.5. Validation of analysis

In this subsection, we aim to validate the findings of the study by relating to behaviors of models from previous studies or providing a rationale for unexpected behavior.

The shortcomings of the study reveal that Top2Vec generates topics with better diversity for the majority of the datasets under consideration. This has been found for both parameters: topic diversity and IRBO. Such behavior for Top2Vec generating better topic diversity has been found in the studies of Hendry et al. [45], Alenezi and Hirtle [85] as well as Vianna and Silva de Moura [86]. At this point, it is important to justify the use of Doc2Vec embedding for Top2Vec instead of other variants. Note that we conducted a sub-analysis among the other embedding variants for Top2Vec, and found Doc2Vec to be performing better than the others on average for our datasets. We compared Doc2Vec, universal-sentence-encoder-multilingual, and distiluse-base-multilingual-cased for two variants of documents: chunked and not chunked, to analyze the impact of length of documents also. Figure 5 shows a partial visualization of results for C_v and C_{npmi} obtained for the KU dataset. Since KU is a unique Italian language dataset, multilingual settings have been used for it.

For the C_{uci} parameter, which measures point-wise mutual information, we observed that ETM depicts better results for the majority of the datasets in our study. The appreciable results by ETM for C_{uci} can also be found in Huynh et al. [87] and Meng et al. [88]. Note that as mentioned earlier, ETM is a devised strategy from LDA with Word2Vec improvement. The LDA component in ETM allows it to identify coherent latent topics probabilistically, while the Word2Vec component provides semantic context and associations of the words, thus possibly providing better C_{uci}, which evaluates topics based on individual probabilities, co-occurrences, and semantic relatedness of words. Also, as LDA is already a well-established strategy delivering considerable C_{uci} coherence [89], an improved version of it is expected to perform even better.

Considering the mean value of C_v coherence parameter for all topic models, NMF was analyzed to generate significantly better results. Such behavior of NMF has been supported by multiple studies, including those by O'Callaghan et al. [82] as well as George and

Figure 4
Topic model evaluation based on coherence metrics

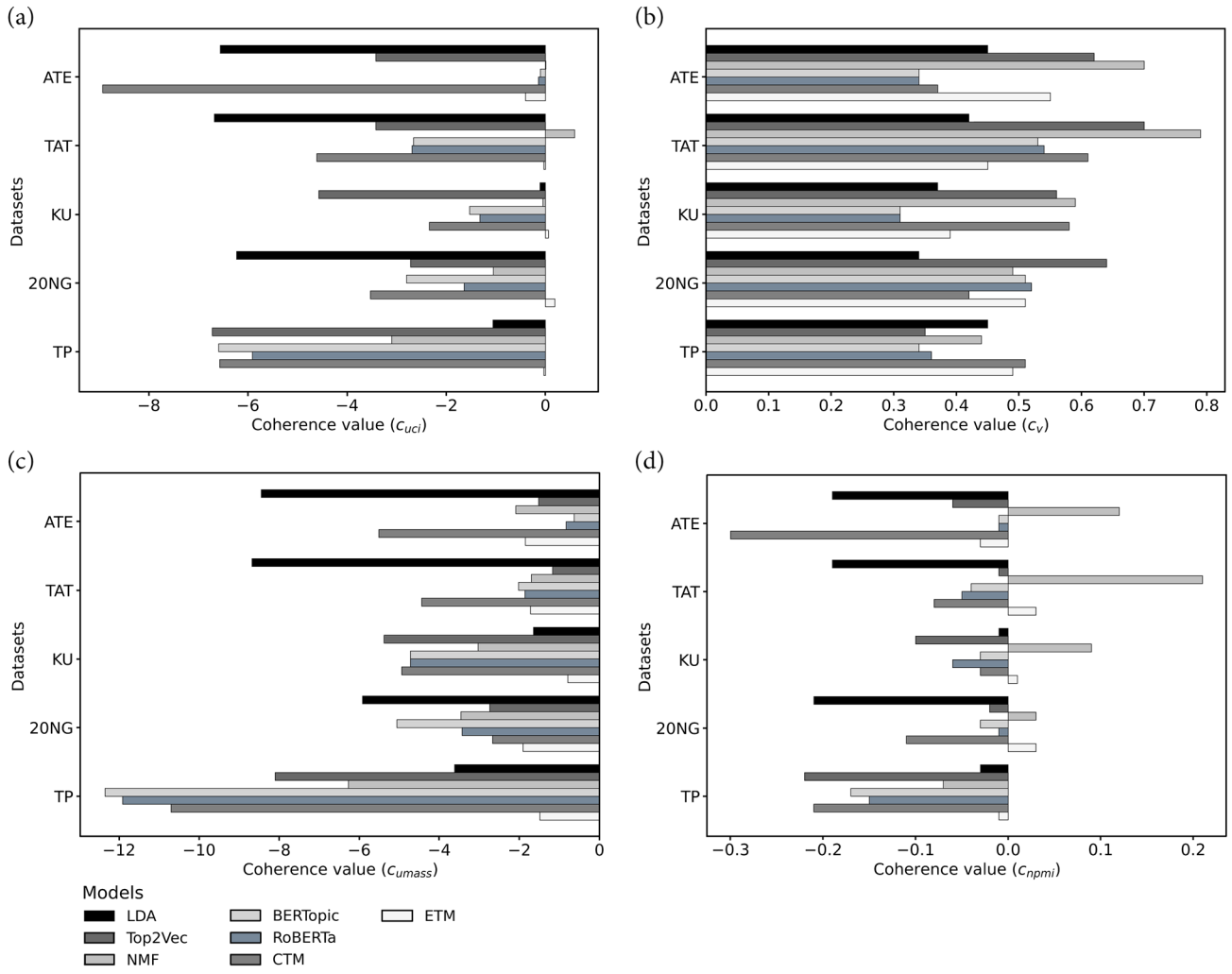


Table 5
Comparison of topic models on ATE dataset

Models	Coherence C_{uci}	Coherence C_v	Coherence C_{umass}	Coherence C_{npmi}	Topic diversity (TD)	IRBO	Number of topics the model uses
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

Srividhya [90]. NMF outperform others solely in 3 out of 5 datasets: ATE, TAT, and KU, while in TP it preceded with a marginal variation in readings. One possible reason of such behaviour is the non-negativity constraint and additive vector approach of NMF that assigns clear and non-overlapping word contributions to topics, enhancing interpretability that is the core of C_v. An interesting observation can be made for the 20NG dataset, where NMF was outperformed by others with a considerable variation. As the size of the 20NG dataset exhibits reasonable expansion in size, we can relate that NMF may not be suitable for larger datasets, as also supported by Guan et al. [91]. A potential reason for it is that NMF extracts dominant patterns that reduce reconstruction errors, prominently for small to medium-sized data, causing consistency with semantic similarity. Hence, it is found more suitable for small to medium-sized data for C_v.

Table 6
Comparison of topic models on TAT dataset

Models	Coherence C_{uci}	Coherence C_v	Coherence C_{umass}	Coherence C_{npmi}	Topic diversity (TD)	IRBO	Number of topics the model uses
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

Note: Values in bold in each column denote the best-performing model.

Table 7
Comparison of topic models on KU dataset

Models	Coherence C_{uci}	Coherence C_v	Coherence C_{umass}	Coherence C_{npmi}	Topic diversity (TD)	IRBO	Number of topics the model uses
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

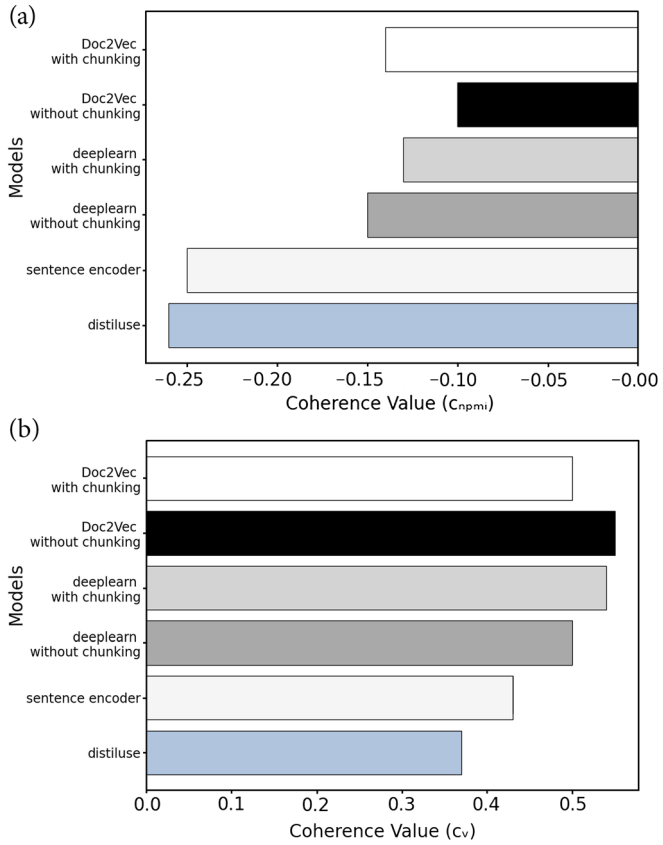
Table 8
Comparison of topic models on TP dataset

Models	Coherence C_{uci}	Coherence C_v	Coherence C_{umass}	Coherence C_{npmi}	Topic diversity (TD)	IRBO	Number of topics the model uses
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 9
Comparison of topic models on 20NG dataset

Models	Coherence C_{uci}	Coherence C_v	Coherence C_{umass}	Coherence C_{npmi}	Topic diversity (TD)	IRBO	Number of topics the model uses
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

Figure 5
Comparisons of variations for Top2Vec



Furthermore, we observed that C_{umass} is rather a different parameter where a lower value signifies better coherence [85]. Note that LDA outperforms others on average in this regard, as similarly found by Tijare and Jhansi Rani [79]. The probable reason for this could be that LDA considers a document as a mixture of topics and words, occurring together with considerable probability, while C_{umass} involves counting of co-document appearance [92] which are more likely to be supported by topics and words occurring together in a document, that is a mixture of topics and words. Hence, the topics produced by LDA are likely to have better C_{umass} scores.

Finally, the study finds that NMF delivers a better C_{npmi} score for the majority of the datasets compared to all other models. The additive vector policy of NMF produces additive topic-word associations that capture strong co-occurrence patterns aligning with NPMI's emphasis on positive word pair relationships. Also its deterministic nature avoids randomness and models words co-occuring frequently with consistency, thus providing better C_{npmi}. Also, C_{npmi} uses a normalized version of the PMI score (known as NPMI) and C_v is also estimated based on the NPMI score, along with cosine similarity. It is most likely for a technique performing better on C_v to also perform better on C_{npmi}, which is observable in the case of NMF for the majority of the datasets.

4.6. Discussion

This section presents a thorough discussion about the working mechanisms and correspondingly achieved findings of topic models of the considered domain context. Notice that in addition to the quantitative results, this section also includes some implicit qualitative findings.

4.6.1. Probabilistic distribution models

As illustrated in Section 4.3, the results obtained are diverse, and our study does not indicate one model to be better than all others, rather it suggests the suitability of models based on the size and type of dataset. Notice that LDA performs visibly better on average in the case of the TAT dataset (Figure 6(b)), considering one coherence and both diversity parameters, C_{umass}, TD and IRBO, recall that TAT is a medium-sized English dataset. While NMF performs better on average considering the coherence parameters, C_{uci}, C_v and C_{npmi} for TAT as well as for ATE, both being small to medium-sized English datasets. Due to the visible difference obtained in coherence readings and a marginal difference in the diversity readings, we suggest that NMF outperforms LDA, which in turn performs better than all others for small to medium English datasets.

4.6.2. Matrix factorization-based models

The qualitative implications find NMF to be faster, more consistent, and producing more human-interpretable topics for small and medium sized English datasets: ATE (Figure 6(a)) and TAT (Figure 6(b)) as compared to others. Furthermore, Figure 6(c) shows that NMF performs appreciably for the KU dataset, a medium-sized Italian language dataset, in terms of C_v and C_{npmi}. Similarly, NMF outperforms others for C_{npmi} and IRBO parameters in the large-sized 20NG dataset (Figure 6(e)). These findings suggest that NMF is suitable for datasets requiring high coherence with human interpretability, particularly small to medium-sized datasets.

4.6.3. Neural embedding-based models

Figure 6(c) illustrates that Top2Vec outperforms others for the KU dataset in terms of C_{umass}, TD, and IRBO, making it suitable for multilingual medium-sized datasets. Similar behavior is observed in Figure 6(e) for the 20NG dataset, where Top2Vec performs better for C_v, TD, and IRBO parameters, followed by NMF. While BERTopic is derived from the Top2Vec architecture, it marginally underperforms compared to Top2Vec in the 20NG dataset. BERTopic shows considerable results only for the TP dataset (Figure 6(d)), particularly for C_{umass}, TD, and IRBO parameters. Notice that there is a marginal difference between BERTopic and RoBERTa for these parameters. The qualitative analysis found BERTopic to be much stochastic in nature for small to medium-sized datasets, producing an insufficient number of topics over multiple runs often illustrating the inclusion of stopwords in the topic words for short-lengthened documents. Hence, we suggest that Top2Vec is suitable for large English datasets while RoBERTa may be suitable for medium-sized English datasets instead of BERTopic due to its better stability, consistency, and efficiency.

For large datasets, Top2Vec demonstrates less variability and consistent performance, making it the most suitable neural embedding-based model.

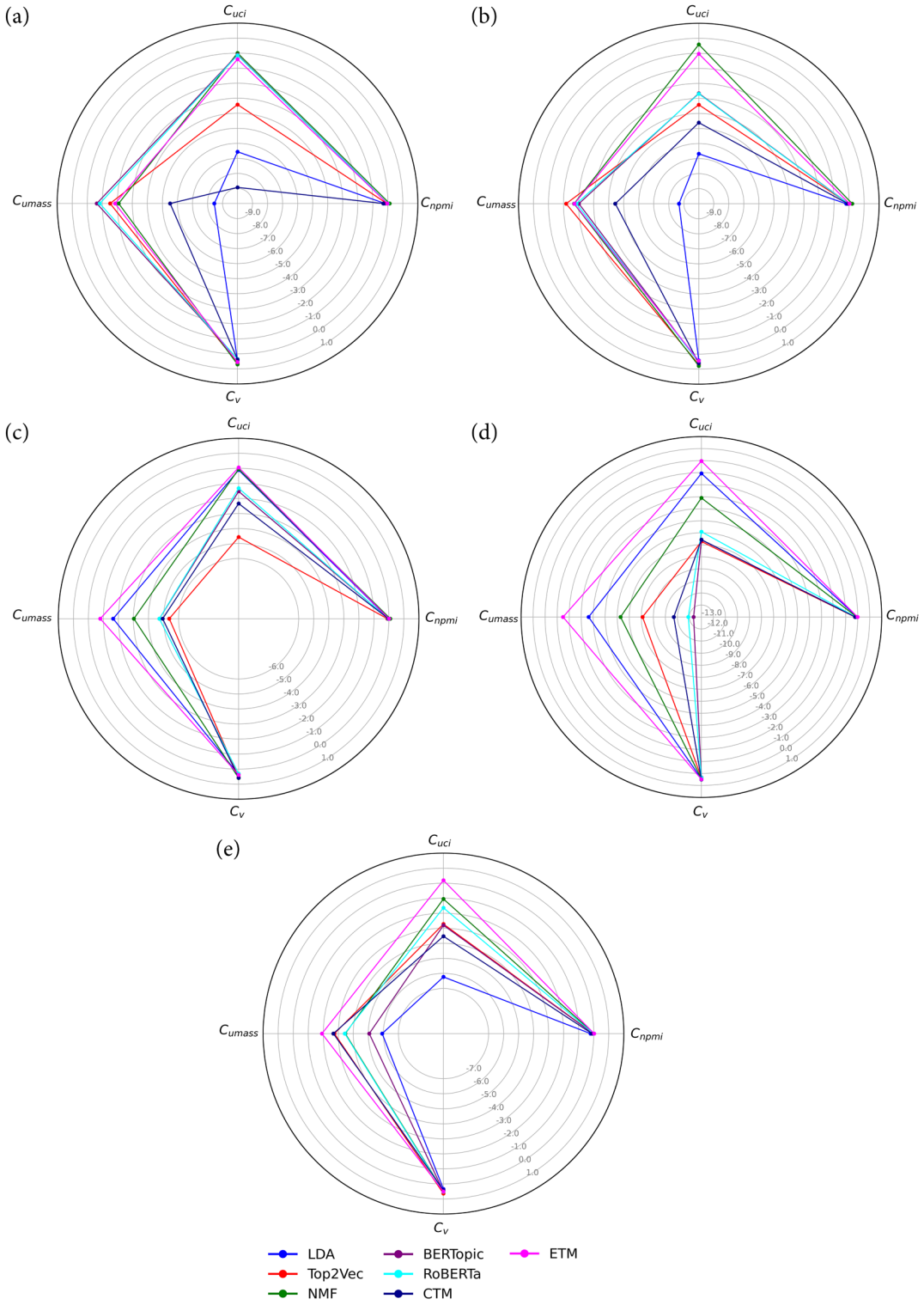
Hence, Top2Vec is suitable for a multilingual medium-sized dataset, followed by NMF which may be adopted if only coherence is under consideration.

4.6.4. General observations

Another interesting implicit finding of this study indicates an observable relationship between the number of topics and TD. We noticed increased TD when the number of topics generated by a model is comparatively lesser. This may be because of the fact that a lesser number of topics indicate a lesser number of clusters, whereas if a dataset is clustered with a comparatively smaller number, the chances of obtaining better inter-cluster distance are increased compared to when clusters are overlapping. Hence the more disjoint the topic clusters are, the higher the TD.

Furthermore, the study finds transformer-based models, Top2Vec, BERTopic, and RoBERTa, stochastic in nature. This is because of the

Figure 6
Comparison of variation for Top2Vec



utilization of UMAP that produces variations in results for repetition of the same experiment [93]. However, out of these three models, Top2Vec shows comparatively lesser variation in results followed by RoBERTa, where BERTopic requires several iterations to produce stable results, in the case of our datasets which comprises usually short-lengthened documents and are small to medium size. An inferred rationale for such behavior of BERTopic can be that a lesser number of documents in the dataset might have resulted in not much distinguishable cluster formation, and since it uses cluster level TF-IDF (cTF−IDF), it may have resulted in the same words in multiple topics (in the case of overlapping clusters) or much lesser number of clusters (topics) which have degraded the performance overall. BERTopic often lacks the accurate identification of all the topics present in our medium-sized datasets, as also mentioned in Thielmann et al. [94].

5. Conclusion, Open Issues, and Future Directions

5.1. Conclusion

Our study delineates a comprehensive review, explaining systemic architectures, principles, and operational mechanisms of promising topic models from the categories of probabilistic models, matrix factorization-based models, and neural embedding-based models. The models are LDA, NMF, Top2Vec, BERTopic, RoBERTa, CTM, and ETM. The study propagates with illustration of a comprehensive performance oriented comparative analysis of the considered topic models based on a series of experimental evaluations, considering the under-explored and unique domain of digital tourism platforms. We analyzed performance of models considering multiple and diverse categorical parameters of topic coherence and diversity. Precisely, the four parameters of coherence this study considers include C_{uci}, C_v, C_{umass}, and C_{npmi}, while TD and IRBO are the two diversity parameters observed in this study. The series of experiments, evaluations and observations is implemented on five diverse and variant datasets where one is generic while the other four are tourism-related, having three exclusive datasets developed for this research. The study provides prominent conclusive quantitative findings along with multiple insightful latent inferences.

The wide-ranging shortcomings and analysis of this study reveal no definitive optimality of one of the models under consideration and that the suitability and performance of the models are dependent to the type and size of data. Therefore, we deduct the suitability of the models in terms of mentioned attributes for the datasets. Considering Table 5, NMF outperforms other models for 3 out of 6 parameters: C_{uci}, C_v, and C_{npmi}, while for each of the parameters: TD, C_{umass} and IRBO, the models Top2Vec, LDA, and CTM performed better than others. Similarly, Table 6 presents the TAT dataset result summary where NMF and LDA outperforms others for 3 out of 6 parameters. NMF outperforms for C_{uci}, C_v and C_{npmi}, while LDA outperforms for C_{umass}, TD, and IRBO. Here LDA exhibits higher diversity while NMF accounts for higher coherence on average. It is notable that Top2Vec generates same TD score as LDA for TAT and even higher for ATE. Therefore, we conclude that NMF is the preferred model for moderately shorter document length based small-sized to medium-sized datasets for better coherence. On the other hand, LDA or Top2Vec might be adopted in similar cases when diversity is concerned.

Furthermore, Table 7 reveals better performance of Top2Vec on average for the multi-lingual medium-sized dataset, KU. Outperforming for 3 out of 6 parameters, Top2Vec generates appreciable outputs for C_{umass}, TD, and IRBO. This trend is followed by NMF performing better for C_v and C_{npmi}. Hence, in such cases, we suggest preference of Top2Vec given the fact that moderately reasonable coherence is satisfactory coupled with high diversity. Otherwise, NMF is well-suited in such cases if higher coherence is the concern irrespective of diversity. Interestingly, Table 8 depicts quantitative outperformance of BERTopic for the medium-sized English dataset, TP, for the parameters TD, C_{umass}, and IRBO. In contrast, qualitative analysis reveals RoBERTa to be exhibiting prominently better qualitative results than BERTopic with marginal difference in quantitative readings for the same parameters. Moreover, Table 9 signifies, on average, the outperformance of Top2Vec for the English large-sized dataset, 20NG, for 3 parameters out of 6 including C_v, TD, and IRBO. Hence, we recommend adopting RoBERTa for the medium-sized datasets and Top2Vec for the large-sized datasets in the English language. ETM might be preferred as well for such cases if only given attention since it illustrates reasonably high coherence in terms of C_{uci} and C_{npmi} for medium to large size datasets.

5.2. Open issues and future directions

The diverse study field of tourism through digital platforms comprises heterogeneous related issues in regard to topic modeling as reported by several studies including the work of Vu et al. [13]. Firstly, the context of tourism lacks a comparative standard for datasets compared to other fields. Secondly, new topic modeling approaches based on deep learning need large and often labelled data, which are often not available for this field. Furthermore, the text or documents describing touristic experiences, tourism products, or tourist reviews are often particularly short in length, which can be challenging for topic models. Although there exist promising attempts to cater to this concern [95], still the issue persists.

Moreover, as neural network-based topic models are often stochastic black boxes, their use may lead to a loss of interpretability of the results or unexpected behavior for different iterative runs, as we experienced in our study. Another important issue in the context of touristic experience is the lack of availability of diverse and versatile benchmark or publicly accessible datasets, which can be used to establish a judgment for topic models.

Additionally, the domain to tourism encounters unique limiting practical challenges including changing preferences of tourists over time or through a single travel lifecycle, restricted access to tourists' reviews, data incompleteness and multimodality, situational events such as pandemics and natural hazards at tourism destinations, availability of experience in seasonality, and impact of social media on tourists' behaviour. Further research in topic models for digital tourism significantly influences real-world digital travel systems offering aiding in intelligent personalized recommender systems, dynamic itinerary planning, and guest satisfaction prediction based on sentiment analysis.

Also, these current limitations and challenges may serve as potential future orientations of this study, where other methods can help cater these problems, such as the usage of knowledge graphs [96] and transfer learning [97] for the approaches based on deep learning and the usage of side information [98] or multimodal data [99].

Another interesting future direction in this particular context of the study is the consideration of the connection between data consulted by the tourists and the period in which such content is consulted. Here, the continuation of our work can consider the dynamic aspect of the data to detect which topics are important in a determined period of time and forecast the topics potentially important for similar future events.

Acknowledgment

Results presented in this article are part of the KuriU Project, supported by Amarena Srl to which the authors are highly grateful.

Funding Support

This work was supported by the Italian Ministry for Economic Development under Grant EASYTOUR: Experience – oriented Search Engine for Touristic Products.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data used in this study includes public and private datasets. The public datasets, Tourpedia and 20 Newsgroups, are available in Tourpedia datasets at http://tour-pedia.org/about/datasets.html and in scikit-learn real world datasets at https://scikit-learn.org/stable/datasets/real_world.html, respectively. Other datasets used in this study are not publicly available but are available from the corresponding author upon reasonable request.

Author Contribution Statement

Maryam Kamal: Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Gianfranco Romani:** Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Visualization. **Giuseppe Ricciuti:** Funding acquisition. **Aris Anagnostopoulos:** Conceptualization, Methodology. **Ioannis Chatzigiannakis:** Conceptualization, Methodology, Writing – original draft, Visualization, Supervision, Project administration.

References

[1] Naab, T. K., & Sehl, A. (2017). Studies of user-generated content: A systematic review. *Journalism*, *18*(10), 1256–1273. https://doi.org/10.1177/1464884916673557

[2] Chatzigiannakis, I., Mylonas, G., & Vitaletti, A. (2011). Urban pervasive applications: Challenges, scenarios and case studies. *Computer Science Review*, *5*(1), 103–118. https://doi.org/10.1016/j.cosrev.2010.09.003

[3] Yan, Q., Jiang, T., Zhou, S., & Zhang, X. (2024). Exploring tourist interaction from user-generated content: Topic analysis and content analysis. *Journal of Vacation Marketing*, *30*(2), 327–344. https://doi.org/10.1177/13567667221135196

[4] Xu, H., Cheung, L. T. O., Lovett, J., Duan, X., Pei, Q., & Liang, D. (2023). Understanding the influence of user-generated content on tourist loyalty behavior in a cultural World Heritage Site. *Tourism Recreation Research*, *48*(2), 173–187. https://doi.org/10.1080/02508281.2021.1913022

[5] Kaosiri, Y. N., Fiol, L. J. C., Tena, M. Á. M., Artola, R. M. R., & García, J. S. (2019). User-generated content sources in social media: A new approach to explore tourist satisfaction. *Journal of Travel Research*, *58*(2), 253–265. https://doi.org/10.1177/0047287517746014

[6] Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2022). A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3549–3568. https://doi.org/10.1109/TKDE.2020.3028705

[7] Kamal, M., & Chatzigiannakis, I. (2021). Influential factors for tourist profiling for personalized tourism recommendation systems—A compact survey. In *2021 International Conference on Innovative Computing*, 1–6. https://doi.org/10.1109/ICIC53490.2021.9693063

[8] Bourg, L., Chatzidimitris, T., Chatzigiannakis, I., Gavalas, D., Giannakopoulou, K., Kasapakis, V., ..., & Zaroliagis, C. (2023). Enhancing shopping experiences in smart retailing. *Journal of Ambient Intelligence and Humanized Computing*, *14*(12), 15705–15723. https://doi.org/10.1007/s12652-020-02774-6

[9] Chatzidimitris, T., Gavalas, D., Kasapakis, V., Konstantopoulos, C., Kypriadis, D., Pantziou, G., & Zaroliagis, C. (2020). A location history-aware recommender system for smart retail environments. *Personal and Ubiquitous Computing*, *24*(5), 683–694. https://doi.org/10.1007/s00779-020-01374-7

[10] Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021). Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management*, *83*, 104241. https://doi.org/10.1016/j.tourman.2020.104241

[11] Korenčić, D., Ristov, S., Repar, J., & Šnajder, J. (2021). A topic coverage approach to evaluation of topic models. *IEEE Access*, *9*, 123280–123312. https://doi.org/10.1109/ACCESS.2021.3109425

[12] Zhang, P., Wang, S., Li, D., Li, X., & Xu, Z. (2020). Combine topic modeling with semantic embedding: Embedding enhanced topic model. *IEEE Transactions on Knowledge and Data Engineering*, *32*(12), 2322–2335. https://doi.org/10.1109/TKDE.2019.2922179

[13] Vu, H. Q., Li, G., & Law, R. (2019). Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management*, *75*, 435–446. https://doi.org/10.1016/j.tourman.2019.06.011

[14] Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, *124*, 389–404. https://doi.org/10.1016/j.jbusres.2020.10.044

[15] Pröllochs, N., & Feuerriegel, S. (2020). Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management*, *57*(1), 103070. https://doi.org/10.1016/j.im.2018.05.003

[16] Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., ..., & Liu, W. (2023). Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology*, *13*, 1219326. https://doi.org/10.3389/fonc.2023.1219326

[17] Asuncion, H. U., Asuncion, A. U., & Taylor, R. N. (2010). Software traceability with topic modeling. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering*, *1*, 95–104. https://doi.org/10.1145/1806799.1806817

[18] Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, *2*, 27–40. https://doi.org/10.1162/tacl_a_00163

[19] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

[20] Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Proceedings of the 14th International Conference on Neural Information Processing Systems*, 535–541.

[21] Angelov, D. (2020). *Top2Vec: Distributed representations of topics*. arXiv. https://doi.org/10.48550/ARXIV.2008.09470

[22] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://doi.org/10.48550/ARXIV.2203.05794

[23] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ..., & Stoyanov, V. (2019). *RoBERTa: A Robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/ARXIV.1907.11692

[24] Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 759–766. https://doi.org/10.18653/v1/2021.acl-short.96

[25] Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. https://doi.org/10.1162/tacl_a_00325

[26] Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. In *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, 267–281. https://doi.org/10.1007/978-981-15-9651-3_23

[27] Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, *83*(13), 37979–38007. https://doi.org/10.1007/s11042-023-17007-z

[28] Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162. https://doi.org/10.1017/S1351324916000334

[29] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, *32*(2), 1188–1196.

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ..., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

[31] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[32] Deshmukh, A., & Raut, A. (2025). Applying BERT-based NLP for automated resume screening and candidate ranking. *Annals of Data Science*, *12*(2), 591–603. https://doi.org/10.1007/s40745-024-00524-5

[33] Wang, Y., Gong, C., Ji, X., & Yuan, Q. (2025). Text classification for evaluating digital technology adoption maturity based on BERT: An evidence of Industrial AI from China. *Technological Forecasting and Social Change*, *211*, 123903. https://doi.org/10.1016/j.techfore.2024.123903

[34] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ..., & Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[35] Uddin, M. S., Haque, M. A., Rifat, R. H., Kamal, M., Gupta, K. D., & George, R. (2024). Bangla SBERT - Sentence embedding using multilingual knowledge distillation. In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, 495–500. https://doi.org/10.1109/UEMCON62879.2024.10754765

[36] Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

[37] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[38] Yu, D., Xu, D., Wang, D., & Ni, Z. (2019). Hierarchical topic modeling of twitter data for online analytical processing. *IEEE Access*, *7*, 12373–12385. https://doi.org/10.1109/ACCESS.2019.2891902

[39] EkiNci, E., & Omurca, S. I. (2020). NET-LDA: A novel topic modeling method based on semantic document similarity. *Turkish Journal of Electrical Engineering & Computer Sciences*, *28*(4), 2244–2260. https://doi.org/10.3906/elk-1912-62

[40] Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 4713–4720. https://doi.org/10.24963/ijcai.2021/638

[41] Zhang, W., Li, Y., & Wang, S. (2019). Learning document representation via topic-enhanced LSTM model. *Knowledge-Based Systems*, *174*, 194–204. https://doi.org/10.1016/j.knosys.2019.03.007

[42] Murakami, R., & Chakraborty, B. (2022). Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts. *Sensors*, *22*(3), 852. https://doi.org/10.3390/s22030852

[43] Liessens, O. (2021). *Unsupervised topic modeling for short documents*. Master's Thesis, Université catholique de Louvain.

[44] Jung, Y. L. (2024). Market intelligence applications leveraging a product-specific Sentence-RoBERTa model. *Applied Soft Computing*, *165*, 112077. https://doi.org/10.1016/j.asoc.2024.112077

[45] Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021). Topic modeling for customer service chats. In *2021 International Conference on Advanced Computer Science and Information Systems*, 1–6. https://doi.org/10.1109/ICACSIS53237.2021.9631322

[46] Kumar, N., & Hanji, B. R. (2023). Normalized category travel personality by considering explicit and implicit feedback (NCTP): Approach for improving travel recommender systems search result. *International Journal of Information Technology*, *15*(7), 3689–3708. https://doi.org/10.1007/s41870-023-01403-7

[47] Kurashima, T., Iwata, T., Hoshide, T., Takaya, N., & Fujimura, K. (2013). Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 375–384. https://doi.org/10.1145/2433396.2433444

[48] Jiang, S., Qian, X., Shen, J., & Mei, T. (2015). Travel recommendation via author topic model based collaborative filtering. In *MultiMedia Modeling: 21st International Conference*, 392–402. https://doi.org/10.1007/978-3-319-14442-9_45

[49] Rossetti, M., Stella, F., & Zanker, M. (2016). Analyzing user reviews in tourism with topic models. *Information Technology & Tourism*, *16*(1), 5–21. https://doi.org/10.1007/s40558-015-0035-y

[50] Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

[51] Irawan, H., Akmalia, G., & Masrury, R. A. (2019). Mining tourist's perception toward Indonesia tourism destination using sentiment analysis and topic modelling. In *Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things*, 7–12. https://doi.org/10.1145/3361821.3361829

[52] Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020). Topic modeling of online accommodation reviews via latent Dirichlet allocation. *Sustainability*, *12*(5), 1821. https://doi.org/10.3390/su12051821

[53] Park, E., Park, J., & Hu, M. (2021). Tourism demand forecasting with online news data mining. *Annals of Tourism Research*, *90*, 103273. https://doi.org/10.1016/j.annals.2021.103273

[54] Egger, R. (Ed.). (2022). *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*. Switzerland: Springer. https://doi.org/10.1007/978-3-030-88389-8

[55] Mishra, R. K., Jothi, J. A. A., Urolagin, S., & Irani, K. (2023). Knowledge based topic retrieval for recommendations and tourism promotions. *International Journal of Information Management Data Insights*, *3*(1), 100145. https://doi.org/10.1016/j.jjimei.2022.100145

[56] Babina, O. I. (2024). Topic modeling for mining opinion aspects from a customer feedback corpus. *Automatic Documentation and Mathematical Linguistics*, *58*(1), 63–79. https://doi.org/10.3103/S0005105524010060

[57] Karamouzi, E., Pontiki, M., & Krasonikolakis, Y. (2024). Historical portrayal of Greek tourism through topic modeling on international newspapers. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 121–132. https://doi.org/10.18653/v1/2024.latechclfl-1.13

[58] Guerrero-Rodríguez, R., Álvarez-Carmona, M. Á., Aranda, R., & Díaz-Pacheco, Á. (2024). Big data analytics of online news to explore destination image using a comprehensive deep-learning approach: A case from Mexico. *Information Technology & Tourism*, *26*(1), 147–182. https://doi.org/10.1007/s40558-023-00278-5

[59] Rabadán-Martín, I., Barcos-Redín, L., Pereira-Delgado, J., Aguado-Correa, F., & Padilla-Garrido, N. (2025). Topic-based engagement analysis: Focusing on hotel industry Twitter accounts. *Tourism Management*, *106*, 104981. https://doi.org/10.1016/j.tourman.2024.104981

[60] Gan, J., Liu, T., Li, L., & Zhang, J. (2021). Non-negative matrix factorization: A survey. *The Computer Journal*, *64*(7), 1080–1092. https://doi.org/10.1093/comjnl/bxab103

[61] Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: An experimental study on BERTopic technique. *Procedia Computer Science*, *189*, 191–194. https://doi.org/10.1016/j.procs.2021.05.096

[62] Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing*, *39*(2), 441–459. https://doi.org/10.1002/mar.21608

[63] Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., ..., & Costa, A. H. R. (2022). ZeroBERTo: Leveraging zero-shot text classification by topic modeling. In *Computational Processing of the Portuguese Language: 15th International Conference*, 125–136. https://doi.org/10.1007/978-3-030-98305-5_12

[64] Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1676–1683. https://doi.org/10.18653/v1/2021.eacl-main.143

[65] Churchill, R., & Singh, L. (2021). Topic-noise models: Modeling topic and noise distributions in social media post collections. In *2021 IEEE International Conference on Data Mining*, 71–80. https://doi.org/10.1109/ICDM51629.2021.00017

[66] Taylor, R. M. C., & du Preez, J. A. (2023). SimLDA: A tool for topic model evaluation. In *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 3*, 534–554. https://doi.org/10.1007/978-3-031-18344-7_38

[67] Bechini, A., Gazzè, D., Marchetti, A., & Tesconi, M. (2016). Towards a general architecture for social media data capture from a multi-domain perspective. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications*, 1093–1100. https://doi.org/10.1109/AINA.2016.75

[68] Patel, J., & Urolagin, S. (2021). Sentiment analysis and prediction of point of interest-based visitors' review. In *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*, 393–401. https://doi.org/10.1007/978-981-15-5243-4_36

[69] Churchill, R., & Singh, L. (2021). textPrep: A text preprocessing toolkit for topic modeling on social media data. In *Proceedings of the 10th International Conference on Data Science, Technology and Applications DATA - Volume 1*, 60–70. https://doi.org/10.5220/0010559000600070

[70] Wu, X., Nguyen, T., & Luu, A. T. (2024). A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, *57*(2), 18. https://doi.org/10.1007/s10462-023-10661-7

[71] Azarbonyad, H., Dehghani, M., Kenter, T., Marx, M., Kamps, J., & de Rijke, M. (2019). HiTR: Hierarchical topic model re-estimation for measuring topical diversity of documents. *IEEE Transactions on Knowledge and Data Engineering*, *31*(11), 2124–2137. https://doi.org/10.1109/TKDE.2018.2874246

[72] Hashimoto, T., Shepard, D. L., Kuboyama, T., Shin, K., Kobayashi, R., & Uno, T. (2021). Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, *77*(5), 4375–4388. https://doi.org/10.1007/s11227-020-03433-5

[73] Carbone, G., & Sarti, G. (2020). ETC-NLG: End-to-end topic-conditioned natural language generation. *Italian Journal of Computational Linguistics*, *6*(2), 61–77. https://doi.org/10.4000/ijcol.728

[74] Terragni, S., & Fersini, E. (2022). OCTIS 2.0: Optimizing and comparing topic models in Italian is even simpler! In *Proceedings of the Eighth Italian Conference on Computational Linguistics*, 1–7.

[75] Terragni, S., Fersini, E., & Messina, E. (2021). Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems*, 33–45. https://doi.org/10.1007/978-3-030-80599-9_4

[76] Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), 20. https://doi.org/10.1145/1852102.1852106

[77] Khodeir, N., & Elghannam, F. (2025). Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques. *Education and Information Technologies*, *30*(5), 5501–5527. https://doi.org/10.1007/s10639-024-13003-4

[78] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

[79] Tijare, P., & Jhansi Rani, P. (2020). Exploring popular topic models. *Journal of Physics: Conference Series*, *1706*(1), 012171. https://doi.org/10.1088/1742-6596/1706/1/012171

[80] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 288–296.

[81] Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics*, 165–174. https://doi.org/10.1109/DSAA.2017.61

[82] O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657. https://doi.org/10.1016/j.eswa.2015.02.055

[83] Bellaouar, S., Bellaouar, M. M., & Ghada, I. E. (2021). Topic modeling: Comparison of LSA and LDA on scientific publications. In *Proceedings of the 2021 4th International Conference on Data Storage and Data Engineering*, 59–64. https://doi.org/10.1145/3456146.3456156

[84] Vijayan, R. (2021). Teaching and learning during the COVID-19 pandemic: A topic modeling study. *Education Sciences*, *11*(7), 347. https://doi.org/10.3390/educsci11070347

[85] Alenezi, T., & Hirtle, S. (2022). Normalized attraction travel personality representation for improving travel recommender systems. *IEEE Access*, *10*, 56493–56503. https://doi.org/10.1109/ACCESS.2022.3178439

[86] Vianna, D., & Silva de Moura, E. (2022). Organizing Portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3388–3392. https://doi.org/10.1145/3477495.3536329

[87] Huynh, V., Zhao, H., & Phung, D. (2020). OTLDA: A geometry-aware optimal transport approach for topic modeling. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

[88] Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022). Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, 3143–3152. https://doi.org/10.1145/3485447.3512034

[89] Mohammed, S. H., & Al-augby, S. (2020). LSA & LDA topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 353–362. https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

[90] George, S., & Srividhya, V. (2020). Comparison of LDA and NMF topic modeling techniques for restaurant reviews. *Indian Journal of Natural Sciences*, *10*(62), 28210–28216.

[91] Guan, N., Tao, D., Luo, Z., & Yuan, B. (2012). Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(7), 1087–1099. https://doi.org/10.1109/TNNLS.2012.2197827

[92] Zhou, Z., & Wakabayashi, K. (2022). Topic modeling using jointly fine-tuned BERT for phrases and sentences. In *The 14th Forum on Data Engineering and Information Management*, 1–8.

[93] Lázaro, C., & Angulo, C. (2024). Using UMAP for partially synthetic healthcare tabular data generation and validation. *Sensors*, *24*(23), 7843. https://doi.org/10.3390/s24237843

[94] Thielmann, A., Weisser, C., Kneib, T., & Säfken, B. (2023). Coherence based document clustering. In *2023 IEEE 17th International Conference on Semantic Computing*, 9–16. https://doi.org/10.1109/ICSC56153.2023.00009

[95] Wu, X., Li, C., Zhu, Y., & Miao, Y. (2020). Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1772–1782. https://doi.org/10.18653/v1/2020.emnlp-main.138

[96] Li, D., Zamani, S., Zhang, J., & Li, P. (2019). Integration of knowledge graph embedding into topic modeling with hierarchical Dirichlet process. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 940–950. https://doi.org/10.18653/v1/N19-1099

[97] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ..., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

[98] Hu, C., Rai, P., & Carin, L. (2016). Non-negative matrix factorization for discrete data with hierarchical side-information. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, *51*, 1124-1132.

[99] Li, M. (2021). Research on extraction of useful tourism online reviews based on multimodal feature fusion. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 82. https://doi.org/10.1145/3453694

How to Cite: Kamal, M., Romani, G., Ricciuti, G., Anagnostopoulos, A., & Chatzigiannakis, I. (2026). Exploring Digital Tourism Through Topic Models: A Review and Experimental Study. *Journal of Data Science and Intelligent Systems*, *4*(2), 137–154. https://doi.org/10.47852/bonviewJDSIS62024472