Leveraging Topic Models to Extract Tourist Preference from Points of Interest Descriptions

Francesca Pulerà, Enza Messina, Elisabetta Fersini

Abstract Tourist preferences play a crucial role in designing personalized travel recommendations and enhancing destination marketing strategies. This paper presents a comparative analysis and a novel approach for modelling tourist preferences by leveraging topic modeling. The main goal of the paper is to extract Topics Of Interest (TOIs) from textual descriptions of points of interest (POIs) to be subsequently used to express tourist preferences when adopting a tourism recommender system. Descriptions of POIs, which encapsulate their key features and attractions, serve as input to advanced topic models to uncover latent topics that represent key themes or categories of interest. The inferred topics can be used to characterize the preferences of tourists to subsequently recommend relevant itineraries.

Keywords: Topic modeling, Tourism, Statistical models, Neural Models.

1 Introduction

In recent decades, digitalization and the growing access to online platforms have radically transformed the way people interact with cultural and historical sites. Data related to tourism and cultural heritage are now available in enormous amounts in various digital channels, including websites, social media, and online databases. This phenomenon presents a unique opportunity to understand how people discover, visit, and learn from Points of Interest (POIs), but it also creates a major challenge: how to organize, analyze, and extract meaningful insights from such vast amounts

Elisabeta Fersini
Universiy of Milano-Bicocca, Viale Sarca 336, Milano e-mail: elisabetta.fersini@unimib.it,
Enza Messina
Universiy of Milano-Bicocca, Viale Sarca 336, Milano e-mail: enza.messina@unimib.it,
Francesca Pulerà
Universiy of Milano-Bicocca, Viale Sarca 336, Milano e-mail: f.pulera@campus.unimib.it

of data to subsequently collect preferences in tourism recommender systems and subsequently suggest relevant itineraries. The analysis of textual content, such as POIs descriptions, has emerged as a powerful solution. Advances in Natural Language Processing (NLP) and machine learning enable the identification of recurring themes in descriptions such as monuments, museums, or natural parks.

In the context of cultural heritage, POIs play a crucial role in attracting tourists and preserving cultural and historical significance. Automating the identification and categorization of Topics of Interest (TOIs) from POI descriptions can enhance tourism management, improve visitor experiences, and inform policies by uncovering trends such as growing interest in immersive experiences or sustainable tourism. Moreover, the identification of TOIs from descriptions of POIs can serve to elicit tourist preferences in the context of a tourism recommender system. Advanced topic modeling techniques offer an effective solution to overcome the challenges of manually analyzing large datasets, allowing a clearer organization of POIs and contributing to the value of cultural heritage. For example, if an automated system identifies that a particular geographic area has a high number of POIs related to contemporary art or medieval history, this could be useful for planning thematic tourist routes or organizing events focused on specific topics.

Our first contribution is the use of a custom dataset to study POIs from the Italian regions of Umbria and Abruzzo, which includes both real data extracted from open-source sources like OpenStreetMap and Wikipedia, as well as synthetic data generated through AI tools. The dataset includes around 5000 POI descriptions in the Italian language. This hybrid approach enriches the dataset and enables the creation of a neural topic model grounded on the touristic attractions of interest. As a second contribution, we compare and optimize three state-of-the-art topic extraction models, Latent Dirichlet Allocation (LDA), Latent Dirichlet Allocation With Products Of Experts (ProdLDA), and Contextualized Topic Models (CTM), to explore thematic distribution of the corpus. The primary goal of this work is to test the effectiveness of these advanced techniques in the context of POI descriptions by comparing both frequentist approaches and neural models. To achieve this goal, we used the OCTIS framework [14], which allowed us to train, optimize, and compare various models for topic distribution analysis in our data set. We conducted both quantitative and qualitative evaluations of the models based on two main metrics: topic coherence and topic diversity. As a final contribution, we propose a Tourism Contextualized Topic Model (T-CTM), which integrates contextualized embeddings from the tourism domain to cover the specificity of the language and to improve the extraction of topics in the considered domain.

The paper is structured as follows: Section 2 provides an overview of related work, focusing on previous applications of topic modeling and advances in the field. Section 3 describes the methodology used to create the dataset, detailing the data sources and enrichment processes. Section 4 outlines the models employed for topic analysis, while Section 5 presents the results obtained, including both quantitative and qualitative evaluations. Finally, Section 6 discusses the implications of the findings, highlights the limitations of the proposed approach, and suggests potential directions for future research.

2 Related Work

Topic models are a class of unsupervised machine learning techniques designed to uncover the underlying themes in a collection of documents. These models are particularly useful for analyzing large corpora where the goal is to extract meaningful topics without requiring prior labeling of the data.

One of the most widely used statistical models for topic discovery is Latent Dirichlet Allocation (LDA) [3]. LDA assumes that documents are mixtures of topics, and topics are distributions over words. However, it generally treats words as being generated independently, which limits its ability to capture complex relationships between terms. To address this limitation, more advanced topic modeling techniques have been proposed. One such method is the Product-of-Experts LDA (ProdLDA), introduced by Srivastava and Sutton [13]. ProdLDA builds on LDA by replacing the traditional mixture model with a product of experts, allowing for a larger number of interpretable topics. This approach is particularly useful for identifying nuanced patterns in large datasets and provides more flexibility in topic discovery compared to standard LDA. Another relevant approach is the Contextualized Topic Model (CTM) [2], which is a family of topic models that use pre-trained representations of language (eg, BERT) to support topic modeling. CTM extends LDA by replacing the input BoW document representations with multilingual contextualized embeddings. This is particularly advantageous when analyzing datasets characterized by different languages where topics are described by different terms but representing the same semantic concept.

Regarding the tourism domain [9], topic modelling has been extensively adopted for several purposes among which suggesting destinations based on user-generated contents [15, 12, 7], or analysing tourism reviews for sentiment analysis purposes [4, 6, 5]. Most of the available works in the state of the art exploit the capabilities of existing models to uncover latent semantic themes underlying tourism related contents. In this paper, we not only compare the most promising approaches but we also propose a Tourism Contextualized Topic Model grounded on Italian destinations.

3 Dataset Creation and Enrichment

3.1 Data Gathering from OpenStreetMap and Wikipedia

The dataset of POI's descriptions was constructed by integrating multiple data sources to include cultural heritage sites located in the Italian regions of Umbria and Abruzzo. The initial step involved retrieving geolocated information necessary for the dataset. OpenStreetMap (OSM)[1], an open-source global mapping service, was chosen due to its accessibility and collaborative nature, which allows users to freely access, modify, and contribute geographical data.

In our case study, focused on the cultural heritage of Umbria and Abruzzo regions, the set of POI have been downloaded from OSM under the amenity, historic,

[1] `https://www.openstreetmap.org/`

Table 1 Example of Italian textual POI's desctiptions

id	title	Lat	Lon	Description
1000	Museo delle Tradizioni Umbre	42.92	12.62	[generated] Il Museo Tradizioni Umbre è un'affascinante struttura situata nel cuore dell'Umbria...
1001	Parco Naturale del Gran Sasso	42.55	13.56	[generated] Il Parco Naturale del Gran Sasso è una vasta area protetta che si estende nel cuore dell'Abruzzo...
...	...	...	...	...
2808	Cascata delle Marmore	42.54	12.72	[real] La cascata delle Marmore è formata dal Velino e dal Nera, affluenti del Tevere,nei pressi di Terni...

and tourism categories. An initial analysis revealed that only a small fraction of POIs was accompanied by the corresponding description. To address this limitation, additional information was sourced from Wikipedia, which provides rich and detailed descriptions of cultural heritage sites. In particular, using longitude and latitude of each POI gathered from OSM we collected the corresponding descriptions from the Wikipedia pages [2] The dataset of real POI's descriptions is composed of 3141 texts of Italian (Umbria and Abruzzo) locations.

3.2 Data Enrichment

To further enrich the dataset and enhance the performance of the topic models, additional synthetic POI descriptions have been created. The goal was to increase the variety and volume of the available data, thus creating a more complete and heterogeneous dataset, which is essential for training the topic models. The process of generating the synthetic POIs was carried out using two different tools that cover the Italian language. In particular, we adopted iGenius[3], a 9 billion parameters language model that has been trained from scratch in Italian on trillions of tokens, using a heterogeneous mix of data as public sources, synthetic data, and domain-specific content. Additionally, we used the GPT3.5[4], which is a 175 billion parameter model that covers multiple languages, to ensure a greater variety in the generated information. Both models have been used to create coherent and realistic descriptions in Italian of various points of interest, including museums, archaeological parks, stadiums, and other relevant categories for the Umbria and Abruzzo regions. The synthetic POIs were then added to the final dataset, which was used to train the topic modeling models to extract potential categories of Topics of Interest. The final dataset, which includes both real and synthetic POIs, consists of a total of 5000 points of interest (POIs). Table 1 shows an extract of the dataset, which includes both real and synthetic points of interest descriptions.

[2] This process was implemented using the Python library wikipediaapi.

[3] `https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1`

[4] `https://chatgpt.com/g/g-F00faAwkE-open-a-i-gpt-3-5`

4 Extracting TOIs from POI's desctiptions

4.1 State-of-the-art topic models

In this study, we analyzed the distribution of topics within our corpus of touristic desctiptions using three existing models: Latent Dirichlet Allocation (LDA), Product of Experts LDA (ProdLDA), and Contextualized Topic Model (CTM). LDA serves as a baseline, offering a probabilistic approach to topic modeling by assuming a Dirichlet prior for topic distributions. ProdLDA extends LDA by incorporating a Variational Autoencoder (VAE) architecture, replacing the mixture model in LDA with a product of experts to enhance topic interpretability. Finally, CTM utilizes pre-trained embeddings to leverage semantic relationships between terms to improve the quality of the extracted topics. In order to exploit the three above-mentioned models, several pre-processing steps have been performed:

- Punctuation removal, to eliminate characters irrelevant for analysis.
- Lowercasing, to ensure uniform word representation.
- Stopword removal, eliminating frequently used words with poor semantic value, such as articles and conjunctions.
- Lemmatization, to reduce inflected word forms to their base form, using tools optimized for the Italian language.
- For CTM, long documents were split into sub-documents using a sentence splitter to meet the architectural requirements of the model.

These steps ensures a suitable corpus that is optimized for topic model issues.

4.2 Tourism Contextualized Topic Model (T-CTM)

To tailor the extraction of TOIs to the specific language nuances of the touristic domain, we extended CTM by integrating a fine-tuned version of the pre-trained embeddings. To this purpose used OCTIS [14], a Python framework that supports the integration of external embedding models. Although the default model for CTM is BERT, we chose to fine-tune an Italian variant of BERT (dbmdz/bert-base-italian-uncased) [11] to obtain embeddings more relevant to the domain of interest. Given the dataset, the Hugging Face transformers API's have been exploited to to initialize the Italian BERT model and the corresponding tokenizer. The text corpus has been tokenized, considering the limit of 512 tokens per sequence, according to the BERT model. During fine-tuning, we applied the Masked Language Modeling (MLM) technique, which randomly masks 15% of the tokens to promote contextual learning. The dataset was split into training (80%) and validation (20%) sets. Fine-tuning was performed using the Hugging Face framework with a learning rate of 2e-5 and a batch size of 16 instances. Once fine-tuning was completed, the model was saved to ensure the continuity of the tensors for proper loading. T-CTM has trained for 50 epochs and employing a zero-shot inference to avoid the need for labelled data. The source code of all the investigated models is available at `https://github.com/MIND-Lab/Tourist-Topic-Model`.

4.3 Hyper-parameters Optimization

In order to determine the optimal configuration of hyperparameters of the considered topic models, we adopted a Bayesian Optimization (BO) strategy. In particular, we adopt the single-objective optimization, focusing on two different objective functions as target performance measures to optimize. In particular, we considered:

- **Topic Coherence**, which measures how much the top-n words representing each topic are related. Specifically, Normalized Pointwise Mutual Information (NPMI) [10] has been adopted. NPMI takes values between [-1; +1], where the higher is the coefficient and the better is the model to denote a topic solution of good quality.
- **Topic diversity**, which quantifies how much the top-n words of a topic differ from the others. In this case, we used the recent IRBO measure proposed in [1]. IRBO takes values between [0; 1], where the higher is the coefficient and the better is the model to denote a topic solution of non-overlapping topics.

The details about the considered hyper-parameters used to optimize Topic Coherence and Topic Diversity are reported in Table 2.

Table 2 Hyper-parameters and their search spaces for the models.

Model	**Hyper-parameter**	**Values/[Range]**
LDA	Num. of topics (k)	[5, 12]
	α	[0.001, 5.0]
	η	[0.001, 5.0]
ProdLDA	Num. of layers	{1, 2, 3}
	Num. of neurons	{100, 200, 300}
	Activation function	{sigmoid, relu, softplus}
	Dropout	[0.0, 0.95]
	Num. of topics (k)	[5, 15]
CTM	Num. of layers	{1, 2, 3}
	Num. of neurons	{100, 200, 300}
	Activation function	{sigmoid, relu, softplus}
	Dropout	[0.0, 0.95]
	Num. of topics (k)	[5, 15]
T-CTM	Num. of layers	{1, 2, 3}
	Num. of neurons	{100, 200, 300}
	Activation function	{sigmoid, relu, softplus}
	Dropout	[0.0, 0.95]
	Num. of topics (k)	[5, 15]

5 Experimental Results

In order to optimize NPMI and IRBO, we used the *Dragonfly* library [8]. To obtain robust evaluations, we train each model 30 times and consider the median of the 30 evaluations as the evaluation of the function to be optimized. A number n of initial

configurations is randomly sampled. The total number of Bayesian Optimization iterations for each model is 125. We use Gaussian Process as the probabilistic surrogate model and the Upper Confidence Bound (UCB) as the acquisition function. As performance measures, we consequently used Topic Coherence, which evaluates the semantic coherence of words within each topic, and Topic Diversity, which measure the non-overlapping between topics. The results of the considered models related to such performance measures are reported in Table 3.

Table 3 Experimental results in terms of Topic Coherence (NPMI) and Topic Diversity (IRBO).

Model	Topic Coherence (NPMI)	Topic Diversity (IRBO)
LDA	0.12	0.14
ProdLDA	- 0.07	0.95
CTM	0.09	0.97
T-CTM	0.05	0.99

We report here a few observations based on the obtained results:

- **LDA**: the LDA model shows a relatively high value of NPMI (0.12) and very low IRBO (0.14). This suggests that, despite LDA is the most widely used topic model, it suffers from generating overlapping topics that can not be used to acquire user preferences for a tourism experience.
- **ProdLDA**: although the NPMI value (-0.07) is negative, the TD score (0.95) is considerably higher. This suggests that the ProdLDA model, while not improving coherence (as shown by the negative NPMI), is much better at maintaining high topic diversity. This suggests that ProdLDA is suitable at managing a variety of heterogeneous and distinct topics, even if they are not as coherent as those produced by other models.
- **CTM**: the CTM model shows significant improvements, with a Topic Coherence that is similar to LDA, but with a higher Topic Diversit. This suggests that CTM strikes a good balance between topic coherence and diversity, likely due to its use of state-of-the-art contextualized embeddings (i.e. BERT), which capture the semantic relationships among words.
- **T-CTM**: the proposed model has a similar performance in terms of Topic Coherence and a slightly better Topic Diversity with respect to the original CTM. This may suggest that while the optimization process increased topic diversity, it might have slightly compromised coherence in favour of a broader range of semantic topics. The results of T-CTM indicate that fine-tuning the underlying language models using tourism-related data can lead to an improvement of the performance with respect to other general models, particularly in terms of Topic Diversity, which is essential for ensuring a distinct separation between topics. The quality of the extracted TOIs can therefore be used to model specific POIs of any geographical area of interest and to acquire user preferences (e.g. ratings) to suggest itineraries.

In Table 4, the top-three topics extracted by each topic model are reported by listing the top-five most probable words that characterize each theme. We can note

Table 4 Top-3 topics extracted by the considered models.

Model	Top-5 most probable words	Candidate Label
LDA	1) art work history museum ancient	1) Museums and Art Galleries
	2) art italian library history place	2) Historical Libraries
	3) history art work ancient italian	3) Museums and Art Galleries
ProdLDA	1) church century two chapel work	1) Churches and Religious Places
	2) reserve flora fauna peak bird	2) Natural Parks and Wildlife Reserves
	3) match pompeo daily stadium spectator	3) Stadiums and Arenas
CTM	1) child canvas Madonna depict fresco	1) Religious and artistic works
	2) church place convent dedicate worship	2) Churches and places of worship
	3) xii medieval 19th century fort	3) Medieval architecture
T-CTM	1) depict Madonna child canvas angel	1) Religious and artistic works
	2) xii xix 13th century medieval	2) Medieval architecture
	3) roman center archaeological area era	3) Roman Archaeological Sites

that T-CTM results in more specific TOIs than the other considered approaches. This suggests that, even if the neural layer of T-CTM has been tuned with a small number of POI's descriptions, it helped to generate more targeted representations, which are particularly valuable when TOIs should be extracted from POI's descriptions and therefore to subsequently acquire user preferences in a tourism recommender systems. The results reflect the model's ability to capture specific themes related to cultural and artistic heritage, providing a good balance between customization and overall performance.

6 Conclusions and future work

In this study, we introduced a novel topic model suitable for the tourism domain. The proposed T-CTM can enhance the thematic analysis in the cultural and artistic domains. The results indicate that (i) T-CTM demonstrated a good ability to provide specific and targeted insights into topic distribution, outperforming the state-of-the-art models and (ii) the results are relevant from the perspective of scholars and professionals interested in cultural heritage, highlighting the effectiveness of neural topic modeling as an advanced analytical tool. T-CTM enables not only the systematic comparison of thematic distributions across different cultural artifacts but also the discovery of relevant and non-overlapping Topics of Interest in a given geographical area that can be subsequently modeled by a tourism recommender system.

Regarding future work, we aim to extend this methodology to larger and more diverse datasets, encompassing a wide range of cultural domains and temporal spans. Additionally, exploring the applicability of transfer learning approaches to adapt topic modeling models to smaller and specialized datasets represents an intriguing direction for further research.

Acknowledgments

This work was supported by the research project *RASTA: Realtá Aumentata e Story Telling Automatizzato per la valorizzazione di Beni Culturali ed Itinerari* funded by the Italian Ministry of University and Research (MUR) under the PON Project ARS01-00540.

References

1. Federico Bianchi, Silvia Terragni, and Dirk Hovy. 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, ACL/IJCNLP 2021*, pages 759–766. Association for Computational Linguistics, 2021.
2. Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1676–1683. Association for Computational Linguistics (ACL), 2021.
3. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
4. Amit Chauhan and Rajni Mohana. Implementing lda topic modelling technique to study user reviews in tourism. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 357–360. IEEE, 2022.
5. Benjamin Garner, Corliss Thornton, Anita Luo Pawluk, Roberto Mora Cortez, Wesley Johnston, and Cesar Ayala. Utilizing text-mining to explore consumer happiness within tourism destinations. *Journal of Business Research*, 139:1366–1377, 2022.
6. Nassera Habbat and Hicham Nouri. Unlocking travel narratives: a fusion of stacking ensemble deep learning and neural topic modeling for enhanced tourism comment analysis. *Social Network Analysis and Mining*, 14(1):82, 2024.
7. Maryam Kamal, Gianfranco Romani, Giuseppe Ricciuti, Aris Anagnostopoulos, and Ioannis Chatzigiannakis. Analyzing topic models: A tourism recommender system perspective. In *International Conference on Advanced Information Networking and Applications*, pages 250–262. Springer, 2024.
8. Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R. Collins, Jeff Schneider, Barnabás Póczos, and Eric P. Xing. Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly. *Journal of Machine Learning Research*, 21:81:1–81:27, 2020.
9. Andrei P Kirilenko and Svetlana Stepchenkova. Facilitating topic modeling in tourism research: Comprehensive comparison of new ai technologies. *Tourism Management*, 106:105007, 2025.
10. Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 530–539, 2014.
11. Antonio Lyding, Marco Sanguinetti, and Roberto Basili. Bertino: A distilled bert for italian. In *Proceedings of the 16th International Conference on Computational Linguistics*, 2021.
12. Seong-Taek Park and Chang Liu. A study on topic models using lda and word2vec in travel route recommendation: focus on convergence travel and tours reviews. *Personal and Ubiquitous Computing*, pages 1–17, 2022.
13. Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, pages 1–12, 2017.

14. Silvia Terragni, Elisabetta Fersini, Bruno Galuzzi, Pietro Tropeano, and Antonio Candelieri. Octis: Comparing and optimizing topic models is simple! *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the System Demonstrations*, page 263–270, 2021.
15. Jun Wang, Yunpeng Li, Bihu Wu, and Yao Wang. Tourism destination image based on tourism user generated content on internet. *Tourism Review*, 76(1):125–137, 2021.