Lingue e Linguaggi
Lingue Linguaggi 73 (2025), 141-161
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v73p141
http://siba-ese.unisalento.it, © 2025 Università del Salento

PROMOTING PUGLIA
A comparative analysis of the destination image and the tourist gaze through BERTopic[1]

ANGELA D'EGIDIO
UNIVERSITY OF SALENTO

Abstract - Destination image has drawn great interest in tourism-related research for several years, with a range of studies approaching the topic from different perspectives. The main objective of this study is to introduce an innovative automated methodology, using the state-of-the-art machine learning model BERTopic (Grootendorst 2022), for analysing the online destination image in institutional tourist communication. To test this approach, Puglia, known as the heel of Italy's boot, was selected as a case study. More specifically, a textual content analysis of two official tourism promotion websites was carried out with the aim of determining whether the regional Destination Marketing Organization (DMO) promotes a coherent destination image and how it directs the tourist gaze (Larsen, Urry 2011) across different digital platforms. Findings reveal discrepancies in the projected destination image, particularly in terms of thematic focus and the language used. In turn, these discrepancies raise discussions about the strategic alignment and coherence of Puglia's destination branding efforts. Therefore, this study not only represents a methodological advancement in the content analysis of destination images but also provides data-driven practical insights for DMOs to refine their strategies for more effective engagement with potential visitors.

Keywords: destination image; DMOs; topic modelling; tourism discourse; promotion; Puglia.

1. Introduction

The exploration of destination image has become a common research area within tourism, due to its significant impact on tourism development (e.g. Marchi and Raschi 2022). Accordingly, a vast number of empirical and theoretical studies have emerged with the aim of conceptualizing destination image, its attributes, measurement and influences (Stepchenkova and Zhan 2013). Previous studies have also largely focused on linguistic aspects of the destination image with the aim of exploring the intersection of cultural representations and the effectiveness of online institutional communication (e.g. Katan 2012, pp. 83-95).

Traditionally, Destination Marketing Organizations (DMOs) play a key role in building the destination image and, by extension, the tourist gaze. Undoubtedly, tourism websites and digital platforms remain a powerful source of institutional communication. They describe destination attractions, influence tourist choices and engage with tourists. Despite this acknowledgment, there is limited research comparing the communicative efforts of a specific DMO across their controlled sources of travel information. In response to this gap, our study employs a state-of-the-art content mining approach grounded in content analysis methodology, defined by Weber (1990, p. 7) as a technique where "many words of a text can be classified into much fewer content categories". Specifically, the

[1] This paper was written as part of a Research Fellowship within the National PRIN project "Destination Italy in English Translation and Language over the Years (DIETALY)" (CUP F83C22000640001).

contribution of this research is twofold: firstly, it presents a new computer-assisted content analysis methodology through topic modelling to automatically investigate how digital platforms strategically direct the tourist gaze; secondly, the proposed methodology allows us to find out whether the Puglia DMO has communicated a consistent and coherent destination image to an international audience over the years.

Rooted in machine learning and natural language processing (NLP), topic modelling is a method that attempts to efficiently structure large amounts of text, based on co-occurrences of terms in similar texts (Daenekindt and Huisman 2020). In other words, topic modelling refers to a group of methods that attempts to identify topics and their prevalence (hidden semantic structures) within a corpus (a collection of documents or a large volume of unstructured text data) in an automated way (Silva, Galster and Gilson 2021). Although it can be described as an inductive approach with quantitative measurements of latent topics from a group of documents, it is also suitable for descriptive and explorative analyses. However, as discussed by Egger (2022) in his overview of the most recent studies in the field of tourism using topic modelling, research using this approach is still quite limited, with Latent Dirichlet Allocation (LDA) being the best known and most widely used algorithm among the topic modelling approaches developed in recent years. Some of the most recent studies with LDA methods in the tourism domain mainly analysed user-generated content from various social media channels (Cai *et al.* 2018), perceptions of tourism experiences on online reviews (Bi *et al.* 2019; Kim *et al.* 2019), tourist activity preferences in the context of travel itineraries (Vu *et al.* 2019) or sentiment analysis to recommend under-emphasised locations (Shafqat and Byun 2020). Nonetheless, recent developments in topic modelling demonstrate that BERTopic can provide clearer and more relevant topic representations compared to the traditional probability-based approach of LDA, thus encouraging broader application within tourism research (Jin 2022). The limited number of studies can be attributed to a knowledge gap in effectively applying topic modelling, especially regarding statistical and mathematical basics as well as programming skills in machine learning techniques (e.g. Python language). Only recently has topic modelling seen broader integration into research, most likely because more user-friendly toolkits – designed for use without requiring any programming expertise – are now available on the market[2].

On the other hand, research in tourism discourse from an applied linguistic perspective has been extensive, tracing both the evolution of theoretical frameworks and the application of methodological tools for analysing the language of tourism. Early studies, such as those by Gotti (2006) and Cappelli (2006; 2013), examined the linguistic features and persuasive strategies within tourism discourse. Francesconi (2012; 2014) further advanced the field by analysing promotional texts, while Manca (2018) provided a detailed analysis of the lexical and semantic dimensions of tourism language. These studies have significantly contributed to understanding how tourism discourse functions as a tool for promoting specific images of places, particularly on websites.

The application of corpus linguistics has enriched tourism discourse analysis, enabling the exploration of large textual corpora to identify patterns and strategies in the verbal promotion of destinations (Maci 2020; Manca 2016; 2018). This approach, combined with manual content analysis, has helped evaluate the quality of tourism discourse and how effectively destinations are marketed. Additionally, research from a translation and cross-cultural perspective (e.g., Denti and Fodde 2017) has shed light on

[2] This paper was written with the assistance of Gianluca Lorenzo, an informatics engineer at the University of Salento, who helped with the data processing.

the complexities of linguistic choices in global tourism communication, emphasising the cultural mediation involved in conveying a destination's image across languages.

What is innovative in this paper is the integration of computer science approaches with tourism language research. To the best of our knowledge, topic modelling techniques and artificial intelligence (AI) methods have not been used so far to explore the semantic relationship between words and extract meaningful information (e.g. keywords and topics) especially in official tourism destination websites. Moreover, while traditional corpus analysis and qualitative approaches provide valuable insights, they may have limited capacity for handling large-scale datasets and identifying deeper semantic patterns. Topic modelling, as implemented in this research, is useful for revealing hidden topics and semantic structures that traditional manual analysis might overlook.

2. Institutional tourism destination websites

In the pre-decision phase, travelers look for as much information as possible to narrow the gap between their expectations and their actual experiences (Maci 2020; D'Egidio 2019), particularly on the Web. Institutional tourism destination websites, which are maintained by DMOs, are the main medium through which info-promotional tourism content is disseminated. They provide potential tourists with detailed information about a destination and, at the same time, create and project positive destination image. Hence, they are both institutional and editorial in terms of genre, in line with Francesconi's classification (2014) of genres in tourism discourse, and serve both persuasive and informative functions. Moreover, as Manca (2016) suggests, "official tourist destination websites are the 'official' representation of a place and act as mediators in the relationship between tourists and destinations at a pre-trip stage". Consequently, they can be considered an official version of the local culture and history.

Accessible with a single click, regularly updated and interactive (Sulaiman *et al.* 2019), tourism websites allow potential tourists to explore holiday experiences virtually thanks to their hypertextual nature. Hyperlinks, or clickable icons and words, give rise to their three key features: multimodality, nonlinearity and interactivity. Most importantly, interactivity lets users control what information they engage actively with, e.g. downloading brochures. Undoubtedly, recent technologies have influenced the way content is presented, without altering the fundamental socio-psychological needs of potential tourists for accurate and up-to-date information regarding a destination's attractions, accommodations, activities and traditions, as well as the search for authenticity and strangeness. However, creating effective tourism websites poses considerable challenges. These websites must not only provide valuable information but also convey it in a way that captures the unique essence of the destination. This task is now commissioned to skilled content creators who expertly write the website's textual content to highlight the distinctiveness of a destination. Their work ensures that the destination's identity and values are communicated effectively, distinguishing it from others and appealing directly to the target audience's desires for unique and authentic travel experiences.

Moreover, most scholars claim that language critically influences tourists' expectations about the sights and experiences and, consequently, becomes the most powerful driving force in tourism promotion. As claimed by Dann (1996, p. 2), in the tourism process 'phrase precedes gaze'. Sulaiman et al. (2019, pp. 17-35) also state that successfully transforming web users into tourists requires leveraging the conventional strengths of the language of tourism promotion. Dann (1996) widely discussed the relationship between language and tourism from four perspectives: authenticity, strangerhood, play and conflict. The perspective of authenticity allows tourists to experience the *typical, real, original, authentic, actual*. The strangerhood perspective emphasizes the *primitive, simple, unsophisticated, natural, different*. The play perspective sees tourism as a game or a form of leisure where tourists engage in activities that allow them to 'play' different roles or temporarily escape from their everyday identities. The conflict perspective explores how language can be used to negotiate and sometimes intensify differences within the tourism context. All the four perspectives are used as a pull factor to attract tourists and manipulate the destination image through textual and communicative strategies.

As a result, destination websites must be well designed, provide relevant information and effectively fulfil the complex needs of modern travelers. However, the creation and management of such institutional websites do not occur in a vacuum.

The following section explores the broader role of DMOs in shaping both the destination image and the tourist gaze, thus providing the theoretical foundation for the comparative analysis developed in this study.

3. The role of destination marketing organisations in shaping the destination image and the tourist gaze

As introduced in the previous section, institutional websites function as primary tools through which DMOs communicate destination narratives. Yet, these platforms are part of a larger strategic framework that governs how destinations are imagined, branded, and promoted.

The concept of the *tourist gaze*, as explained by Urry and Larsen (2011), significantly informs how DMOs shape and project the image of a destination. Since the early 1970s, numerous studies have underlined the importance of creating an attractive and unique destination image, which prove the extensive efforts of DMOs in researching, building, promoting, evaluating and maintaining these images through various promotional materials such as brochures, guidebooks, postcards and, more recently, digital platforms and websites (e.g. Mayo 1973; Pike 2002; Stepchenkova and Zhan 2012; Tasci *et al*. 2007). As Kolcun and Grabara (2014, p. 3) note, "the first contact between a tourist and a touristic location isn't the location itself but a representation of it", underlining the important role of DMOs in creating these images. Through medium- and long-term strategy planning, DMOs try to create an attractive and authentic brand image that highlights the destination's unique attributes, such as culture, nature, food and architecture. Indeed, over the years, DMOs have intensively worked on defining a persuasive destination image, extending beyond mere promotion to adopt strategies aimed at differentiating a destination's identity (Blain et al. 2005).

Importantly, the dissemination of such images through marketing communications strongly influences how tourism products and experiences are perceived (Choi *et al*. 2007). In this regard, the content featured on official tourist websites becomes essential for conveying what Marchi and Raschi (2022) describe as "ideas and impressions about a place available for potential visitors' consideration." The external construction of destination images, as explored by Manca (2016), is manipulated as part of a pre-trip promotional strategy. These representations not only influence tourist expectations but also actively shape their perceptions of host cultures, guiding their gaze during their travels. Tourists arrive at destinations with preconceived ideas, shaped not only by official communications but also by user-generated content such as travel blogs and social media, which often replicate and reinforce dominant perceptions (Francesconi 2012). As a result, tourists are both consumers and reproducers of destination images.

Accordingly, the tourist gaze remains a powerful tool in the hands of the tourism institutions that strategically shape and structure it externally, through selective presentation of information made available to consumers. As MacCannell (1973, p. 589) observes, the industry constructs images of places as “worth seeing,” thereby reinforcing the distinction between the ordinary and the extraordinary. In doing so, certain elements are amplified, while others are omitted or silenced.

Additionally, as demonstrated by Echtner and Prasad (2003) and Stone and Nyaupane (2019), in their role as both creators and controllers of promotional images that direct the tourist gaze, DMOs construct images of destinations corresponding to the needs of generating markets, thereby creating tailored destination representations. In this light, by managing tourist experiences DMOs shape not only what is seen but also how it is interpreted. This underlines the importance of authorship and translation in the production of promotional content, as these linguistic and cultural mediations significantly influence how destinations are perceived and interpreted by international audiences.

Moving beyond the notion of tourism as a merely passive visual consumption of carefully constructed images and sites, Larsen and Urry’s (2011) notion of “performance turn” provides a more complex view of tourism as involving embodied practices, performances and experiences. This perspective views tourism as an active engagement where tourists immerse themselves in destinations, for example, through well-designed thematic itineraries, published by DMOs on official destination websites. These itineraries not only facilitate logistical planning but also enhance the performative aspect of tourism, deepen tourists’ connections with local cultures and environments. Thus, DMOs play a key role in enabling tourists to engage more meaningfully with the destination from their time before arriving at the destination through to the end of their tour.

4. The case study of Puglia

Puglia, known for its mixture of nature, tradition, architecture, history and culinary heritage, was identified as a case study destination for the current research. Before the establishment of a dedicated promotional entity, there was no real systematic promotion of Puglia, leaving its potential underexploited in both national and international tourism markets. To address this gap and enhance its market position, the Regional Tourism Agency, Pugliapromozione, was founded in 2011. Since then, this entity has focused on promoting the region through digital initiatives and strategic planning, thanks to the implementation of the unified Regional Strategic Plan. This shift towards structured promotion has aimed to exploit new market opportunities and elevate Puglia’s profile on the global stage.

The significant impact of these efforts was evident in the positive press exposure Puglia received in 2014, highlighting the effectiveness of the agency’s strategies. The region was awarded the “Best Value Travel Destination” by the National Geographic Travel magazine for its cultural heritage, stunning landscapes and off-the-beaten-path experiences; the Lonely Planet put Puglia at second place in the “Top ten best value travel destinations”, for its beautiful beaches, hilltop towns, ancient sights, simple and tasty cuisine, and relaxed pace of life; while the New York Post praised its rustic charm and lower tourist traffic compared to other Italian regions.

As part of the promotion and digitalization strategies, Pugliapromozione re-designed the official tourism portal Viaggiareinpuglia.it in terms of content, editorial choices and customization in 2015, created the Destination Management System platform to integrate all tourism-related businesses with Pugliapromozione and organised the BuyPuglia business program to facilitate connections between global buyers and local sellers.

Among others, these efforts contributed to a significant upward trend in the tourism growth: in 2013, the region registered a positive growth of +5% in foreign arrivals for the consecutive year; in 2014, it maintained its strong appeal in foreign markets, with a +5% increase in arrivals compared to 2013; in 2015, the results of internationalization efforts in Puglia became increasingly evident, as the region registered a positive growth of +10% increase in foreign visitor arrivals compared to the previous year (Regione Puglia 2015). The rise in international tourist arrivals from 29% in 2022 to 34% in 2023 (Regione Puglia 2023) further highlights Puglia’s successful strategies in attracting more tourists from abroad and its growing appeal as a tourist destination.

5. Data collection and methods

In this section, we will outline the processes and methodologies employed in gathering and analysing the data for this study. We will provide a detailed description of the data sources and the techniques used to ensure a comprehensive understanding of the destination brand and tourist gaze projected by Pugliapromozione.

5.1. Data collection

To explore the destination brand and tourist gaze projected by Pugliapromozione, we analysed two distinct textual datasets, each considered as a separate corpus: the official website *Viaggiareinpuglia* and the section *Puglia Routes & Experiences*, available within the BuyPuglia digital platform.

More specifically, the first corpus includes the English translation of the official tourism website *Viaggiareinpuglia*, designed in 2015 and updated in March 2023 mainly in terms of layout, not content, except for the addition of thematic travel guides. *Viaggiareinpuglia* is the main tourism touchpoint for the region and offers tourist information and practical solutions for all visitors to Puglia, covering places and attractions, but also activities, accommodation and travel ideas also in the form of downloadable itineraries. Since the content on this website is both editorial and written by local operators, only editorial content was taken into account for the purpose of this research, as this reflects the institutional communication. In detail, the corpus includes content from the section *Where to go*, with the description of the six geographical areas of Puglia; the subcategories *Treasures to discover*, *Villages and cities*, *UNESCO heritage* and *Travel ideas* along with the suggested itineraries for discovering Puglia; and the section *Guides and media*, which provides downloadable thematic travel guides. The second corpus consists of content from the digital platform *Puglia Routes & Experiences*, created exclusively in English in April 2021 and entirely downloadable. Commissioned by Pugliapromozione, this content was written by Duncan Garwood, an expert travel writer known for his contributions to Lonely Planet guides on Rome and Italy. Designed initially as a resource for international tourism professionals, buyers and media stakeholders, the guide offers practical planning information, detailed itineraries and authentic travel experiences. Recently, acknowledging its broader potential appeal, Pugliapromozione integrated the content into *Viaggiareinpuglia* as a downloadable PDF, thus making it directly accessible to all visitors interested in exploring the region.

Although they were originally tailored to suit different targets, the DMO now employs both resources to promote the region's international visibility.

For the textual content analysis, web scraping was employed to collect data from the above-mentioned sections of *Viaggiareinpuglia* and the platform *Routes & Experiences*. Specifically, Selenium (version 4.25.0, available at https://www.selenium.de v/) with a Chrome driver was used, and the scraping process was programmed using Python. Web scraping is particularly advantageous as it allows to automatically extract large amounts of textual data, significantly reducing the time and effort compared to manual data collection, and helps navigate the complex structure of websites characterised by multimodality, nonlinearity and interactivity.

5.2. Methods

The topic modelling technique was used to identify the main similarities and differences between the two DMO-controlled sources. More specifically, topic modelling helped find hidden semantic patterns and uncover latent themes or "topics" within a group of documents, usually in the form of a bag of the most important words (tokens). This process is simplified by using BERTopic approach (Grootendorst 2022), amongst others, to capture the main topics in the two corpora. More specifically, it allowed to find hidden semantic patterns in texts (usually called documents) and uncover latent themes or "topics" within a group of documents, usually in the form of a bag of the most important words (tokens), as it leverages the power of pre-trained transformer-based language models, in order to create semantically rich document embeddings. It then uses a novel class-based variation of TF-IDF (Term Frequency-Inverse Document Frequency) to extract coherent topic representations from these embeddings. The process can be broken down into three main steps: generating document embeddings using a pre-trained language model, clustering these embeddings to form topics, and finally, creating topic representations using the class-based TF-IDF procedure.

In detail, we embedded each document of our corpora using the pre-trained transformer-based language model Sentence-BERT (Reimers and Gurevych 2019). During this embedding step, sentences and paragraphs are converted to dense vector representations through the sentence-transformer package Hugging Face in Python 3.10 allowing for the semantic comparison of documents discussing similar topics based on their contextual relationship.

After generating document embeddings, we reduced their dimensionality to enhance the clustering process and overcome the "curse of dimensionality" (Pandove et al. 2018). UMAP (Uniform Manifold Approximation and Projection) was applied to project the high-dimensional embeddings into a lower-dimensional space of document vectors and to facilitate the clustering step that groups semantically similar documents. We reduced the vectors to 15 neighbours to emphasize the data local structures and to see more of the overall structure of the data. The minimum distance between embedded points was set at 0.00, which determines the density with which UMAP arranges the points. The dimensionality-reduced embeddings were clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), a clustering method refined by McInnes *et al*. (2017). This step groups semantically similar documents together into clusters, each delineating a distinct topic. HDBSCAN was chosen for its ability to identify clusters of varying densities and its soft-clustering approach, which helps in treating noise as outliers, thereby improving the purity of clusters. Interestingly, HDBSCAN has a parameter to control the number of resulting topics. A higher *min_cluster_size* generates fewer, but broader, topics with larger cluster of data points (documents). For

Viaggiareinpuglia, a high *min_cluster_size*=16 was set, as it comprised 5110 documents. Originally there were 6549 documents, including 1923 web pages and eight travel guides in PDF format. This adjustment was needed as some website pages were not translated into English. Elements containing Italian text were excluded by using a language detector. With large document datasets, a higher *min_cluster_size* ensures the formation of clusters that represent significant and substantial topics across the documents. Conversely, a lower *min_cluster_size* is preferred in order to obtain a larger number of topics in smaller clusters of documents, simplifying their analysis and interpretation. Accordingly, a low *min_cluster_size*=5 was set for *Routes & Experiences*, which contained only 1115 documents. This different setting across the two corpora led to a simplified topic landscape in *Viaggiareinpuglia* by merging closely related topics, and a more detailed exploration of the topics within *Routes & Experiences*.

After clustering, CountVectorizer was used to create a document-term matrix, quantifying the occurrence of each term across documents within the same topic cluster. This matrix of token counts supports the calculation of class-based TF-IDF (Term Frequency-Inverse Document Frequency) scores and enables BERTopic to identify terms that are frequent and distinctive within each topic cluster (and not within individual documents). In order to have a set of keywords that concisely represent each topic, a list of stop words was also added. Through this process, CountVectorizer facilitates the extraction of explicit, keyword-based topic descriptions from the nuanced, contextual relationships captured by document embeddings, underscoring its crucial role in making BERTopic a powerful tool for extracting meaningful and interpretable topics from text data. This approach has the advantage of linking the deep semantic understanding provided by embeddings and the need for interpretable, keyword-based topic representations, highlighting the innovative integration of deep learning embeddings with traditional NLP techniques within BERTopic.

Many different topic representations are possible through the BERTopic, each of which providing insightful perspectives of topic descriptions. Among the topic representations, we tested KeyBERT Inspired, a method that draws inspiration from KeyBERT's keyword extraction strategy (Sharma and Li 2019) and leverages BERT embeddings to identify semantically significant phrases that enclose the core themes of topics, offering a keyword-centric view of the topics discovered. We also tested the Part-Of-Speech representation, which uses SpaCy's POS tagging. By analysing the grammatical structure of the text, Part-Of-Speech focuses on extracting nouns, adjectives or other relevant parts of speech that significantly contribute to the topic's thematic content. While KeyBERT is useful for identifying contextually rich and semantically nuanced keywords and phrases, POS analysis complements this by highlighting the frequency and distribution of specific grammatical categories, which play a key role in tourism discourse. Comparing POS and KeyBERT analyses thus provides a more comprehensive understanding of how the destination image is linguistically constructed, revealing whether the marketing strategies emphasize thematic specificity (KeyBERT) or broader, descriptive conceptual categories (POS). This comparative approach enabled a richer interpretation of how language strategically shapes the tourist gaze.

Lastly, and more importantly, the integration of OpenAI's technology, using ChatGPT, introduces a novel approach to topic labelling. By inputting topic descriptors into ChatGPT with a carefully crafted prompt, it generated more natural and understandable labels for each topic, bridging the gap between computational topic modelling and intuitive topic interpretation.

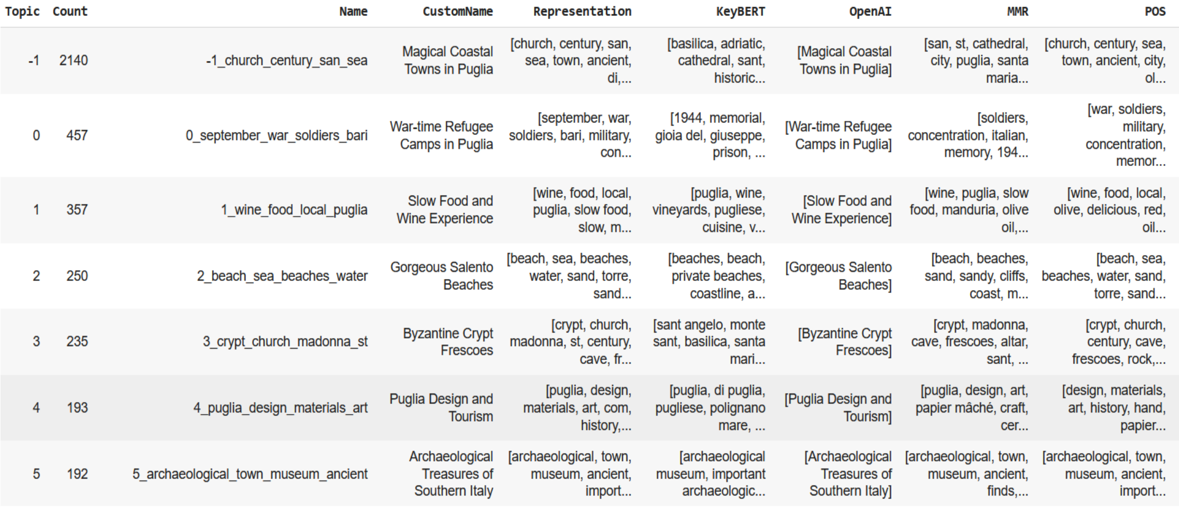
Finally, BERTopic's versatility was enhanced by setting the hyperparameter *top_n_words=10*, ensuring that each topic representation focuses on the ten most representative words.

6. Findings

In this section, we present the results of our analysis, starting with the output from BERTopic and an examination of outliers in the two websites (section 6.1). We then provide a thematic overview using descriptions generated by OpenAI's technology, highlighting key differences and similarities (section 6.2). This is followed by a comparison of the KeyBERT and POS tagging representations (section 6.3). Finally, we analyse the top words identified by KeyBERT and POS tagging to highlight the thematic focus and language used by the Puglia DMO in shaping the tourist gaze (section 6.4).

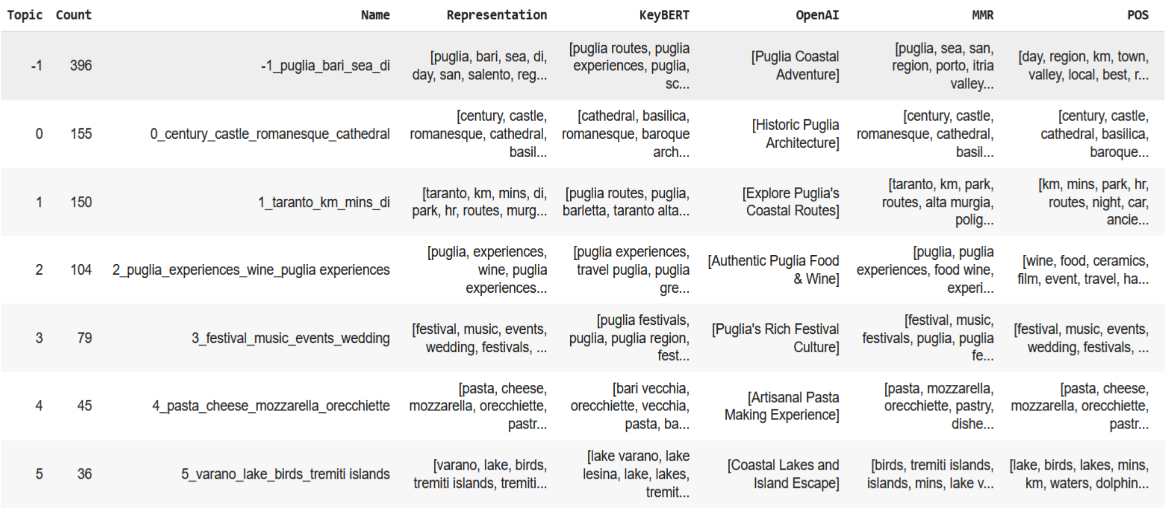
6.1. BERTopic output and analysis of outliers

The topic model returned 28 topics for *Viaggiareinpuglia* and 17 topics for *Routes & Experiences*, as shown in Figures 1 and 2 below, which display an extraction of the output from BERTopic of the two datasets.



Topic	Count	Name	CustomName	Representation	KeyBERT	OpenAI	MMR	POS
-1	2140	-1_church_century_san_sea	Magical Coastal Towns in Puglia	[church, century, san, sea, town, ancient, di,...	[basilica, adriatic, cathedral, sant, historic...	[Magical Coastal Towns in Puglia]	[san, st, cathedral, city, puglia, santa maria...	[church, century, sea, town, ancient, city, ol...
0	457	0_september_war_soldiers_bari	War-time Refugee Camps in Puglia	[september, war, soldiers, bari, military, con...	[1944, memorial, gioia del, giuseppe, prison, ...	[War-time Refugee Camps in Puglia]	[soldiers, concentration, italian, memory, 194...	[war, soldiers, military, concentration, memor...
1	357	1_wine_food_local_puglia	Slow Food and Wine Experience	[wine, food, local, puglia, slow food, slow, m...	[puglia, wine, vineyards, pugliese, cuisine, v...	[Slow Food and Wine Experience]	[wine, puglia, slow food, manduria, olive oil,...	[wine, food, local, olive, delicious, red, oil...
2	250	2_beach_sea_beaches_water	Gorgeous Salento Beaches	[beach, sea, beaches, water, sand, torre, sand...	[beaches, beach, private beaches, coastline, a...	[Gorgeous Salento Beaches]	[beach, beaches, sand, sandy, cliffs, coast, m...	[beach, sea, beaches, water, sand, torre, sand...
3	235	3_crypt_church_madonna_st	Byzantine Crypt Frescoes	[crypt, church, madonna, st, century, cave, fr...	[sant angelo, monte sant, basilica, santa mari...	[Byzantine Crypt Frescoes]	[crypt, madonna, cave, frescoes, altar, sant, ...	[crypt, church, century, cave, frescoes, rock,...
4	193	4_puglia_design_materials_art	Puglia Design and Tourism	[puglia, design, materials, art, com, history,...	[puglia, di puglia, pugliese, polignano mare, ...	[Puglia Design and Tourism]	[puglia, design, art, papier mâché, craft, cer...	[design, materials, art, history, hand, papier...
5	192	5_archaeological_town_museum_ancient	Archaeological Treasures of Southern Italy	[archaeological, town, museum, ancient, import...	[archaeological museum, important archaeologic...	[Archaeological Treasures of Southern Italy]	[archaeological, town, museum, ancient, finds,...	[archaeological, town, museum, ancient, import...

Figure 1
Extraction of the output from BERTopic of *Viaggiareinpuglia*.



Topic	Count	Name	Representation	KeyBERT	OpenAI	MMR	POS
-1	396	-1_puglia_bari_sea_di	[puglia, bari, sea, di, day, san, salento, reg...	[puglia routes, puglia experiences, puglia, sc...	[Puglia Coastal Adventure]	[puglia, sea, san, region, porto, itria valley...	[day, region, km, town, valley, local, best, r...
0	155	0_century_castle_romanesque_cathedral	[century, castle, romanesque, cathedral, basil...	[cathedral, basilica, romanesque, baroque arch...	[Historic Puglia Architecture]	[century, castle, romanesque, cathedral, basil...	[century, castle, cathedral, basilica, baroque...
1	150	1_taranto_km_mins_di	[taranto, km, mins, di, park, hr, routes, murg...	[puglia routes, puglia, barletta, taranto alta...	[Explore Puglia's Coastal Routes]	[taranto, km, park, routes, alta murgia, polig...	[km, mins, park, hr, routes, night, car, ancie...
2	104	2_puglia_experiences_wine_puglia experiences	[puglia, experiences, wine, puglia experiences...	[puglia experiences, travel puglia, puglia gre...	[Authentic Puglia Food & Wine]	[puglia, puglia experiences, food wine, experi...	[wine, food, ceramics, film, event, travel, ha...
3	79	3_festival_music_events_wedding	[festival, music, events, wedding, festivals, ...	[puglia festivals, puglia, puglia region, fest...	[Puglia's Rich Festival Culture]	[festival, music, festivals, puglia, puglia fe...	[festival, music, events, wedding, festivals, ...
4	45	4_pasta_cheese_mozzarella_orecchiette	[pasta, cheese, mozzarella, orecchiette, pastr...	[bari vecchia, orecchiette, vecchia, pasta, ba...	[Artisanal Pasta Making Experience]	[pasta, mozzarella, orecchiette, pastry, dishe...	[pasta, cheese, mozzarella, orecchiette, pastr...
5	36	5_varano_lake_birds_tremiti islands	[varano, lake, birds, tremiti islands, tremiti...	[lake varano, lake lesina, lake, lakes, tremit...	[Coastal Lakes and Island Escape]	[birds, tremiti islands, islands, mins, lake v...	[lake, birds, lakes, mins, km, waters, dolphin...

Figure 2
Extraction of the output from BERTopic of *Routes & Experiences*.

As a way of quantifying the importance of each topic, the tool provides the number of documents falling within each topic.

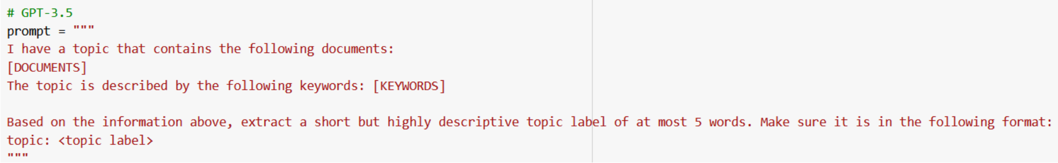
Interestingly, the outcome of BERTopic is a set of the top most representative words for each topic in the corpus (see column *Representation*). The first four keywords form the name of the topic label (see column *Name*). Although they may seem to have a recognizable semantic relationship, such keywords require further inferring of the general theme to understand the underlying topics. To overcome this problem and have easily interpretable topic labels, the use of the language model GPT-4 was tested. The topics obtained with the AI approaches are presented in column *OpenAI*.

What is worth noting is the presence of the largest groups in both corpora *Topic -1*, which correspond to outliers. By default, using HDBSCAN for clustering, BERTopic does not force all data points to be part of clusters. Outliers are often referred to as “noise points” or documents that do not belong to any of the identified topic clusters due to their insufficient similarity to the core samples of any cluster. The semantic content of these documents suggests the presence of too broad or too distinctive topics within the data. By keeping track of outliers, BERTopic ensures that the analysis does not overlook potentially significant or less common themes. In our case, the *Viaggiareinpuglia* corpus, with its dominant outlier topic “*Magical coastal town in Puglia*”, highlighted the region’s coastal charm and a significant emphasis on descriptions of a set of locations along Puglia’s coast. The presence of such a dominant outlier topic can be indicative of the dataset’s rich coverage of scenic coastal towns. Conversely, the *Routes & Experiences* corpus presented “*Puglia Coastal Adventure*” as its outlier topic. While also emphasising the coastal aspect of Puglia, this topic seems to suggest a more activity-oriented perspective. This indicates a thematic overlap with the *Viaggiareinpuglia* corpus but from a different angle: emphasis on activities over scenic beauty or historical significance. The number of documents within outliers raises an interesting point regarding their representativeness. By definition, outliers differ significantly from the core clusters, suggesting that these topics may not align closely with the main themes of each dataset. However, the relatively high number of documents within these outliers suggests that these are not mere anomalies but rather significant themes within their respective corpora. Their substantial count indicates a considerable amount of text focussing on these themes. As a result, they seem to capture essential aspects of Puglia’s destination image not covered by more narrowly defined topics.

Interpreting both the outliers already reveals a different gaze in the sources. The Puglia DMO highlights the natural beauty and historical depth of its coastal towns in *Viaggiareinpuglia*, whereas they invite exploration and adventure along the region's shores in *Routes & Experiences*. In conclusion, these outliers, far from being peripheral, provide insightful insights into the thematic diversity of the DMO-controlled sources.

6.2. Thematic overview based on OpenAI descriptions

OpenAI's technology was used to fine-tune topics and generate summarizing labels for each identified cluster. Specifically, writing the custom prompt in Figure 3, we asked ChatGPT within BERTopic to extract a short description of the topics.

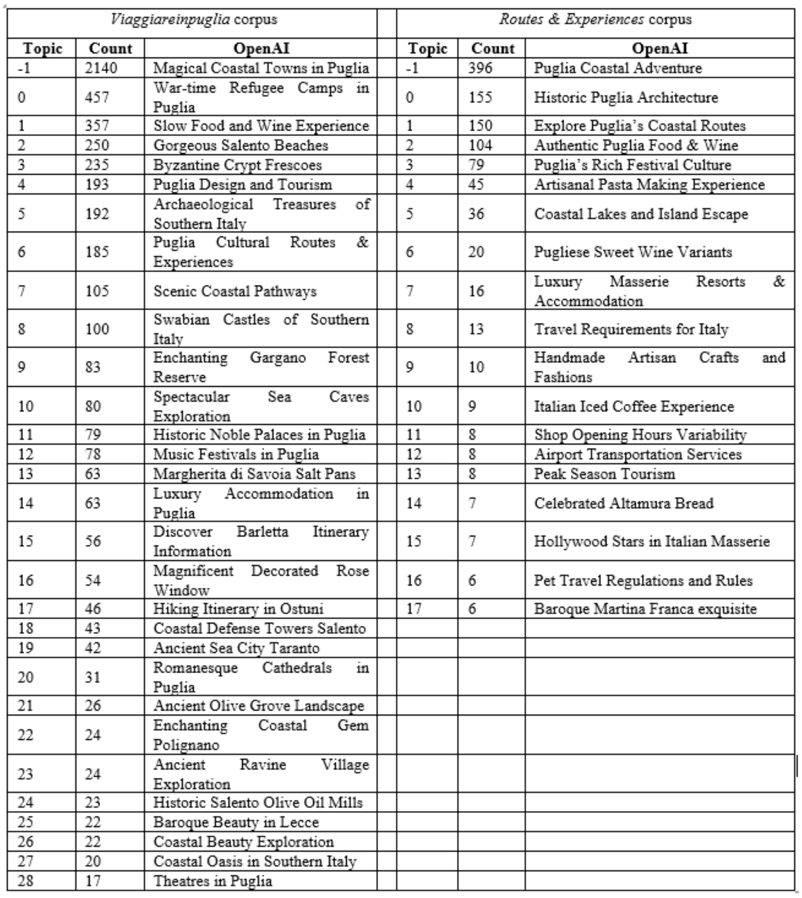


```
# GPT-3.5
prompt = """
I have a topic that contains the following documents:
[DOCUMENTS]
The topic is described by the following keywords: [KEYWORDS]

Based on the information above, extract a short but highly descriptive topic label of at most 5 words. Make sure it is in the following format:
topic: <topic label>
"""
```

Figure 3
Custom prompt for ChatGPT within BERTopic.

Evidently, OpenAI descriptions provide a high-level overview of what each topic is about (see Figure 4).



Viaggiareinpuglia corpus				*Routes & Experiences* corpus		
Topic	**Count**	**OpenAI**		**Topic**	**Count**	**OpenAI**
-1	2140	Magical Coastal Towns in Puglia		-1	396	Puglia Coastal Adventure
0	457	War-time Refugee Camps in Puglia		0	155	Historic Puglia Architecture
1	357	Slow Food and Wine Experience		1	150	Explore Puglia's Coastal Routes
2	250	Gorgeous Salento Beaches		2	104	Authentic Puglia Food & Wine
3	235	Byzantine Crypt Frescoes		3	79	Puglia's Rich Festival Culture
4	193	Puglia Design and Tourism		4	45	Artisanal Pasta Making Experience
5	192	Archaeological Treasures of Southern Italy		5	36	Coastal Lakes and Island Escape
6	185	Puglia Cultural Routes & Experiences		6	20	Pugliese Sweet Wine Variants
7	105	Scenic Coastal Pathways		7	16	Luxury Masserie Resorts & Accommodation
8	100	Swabian Castles of Southern Italy		8	13	Travel Requirements for Italy
9	83	Enchanting Gargano Forest Reserve		9	10	Handmade Artisan Crafts and Fashions
10	80	Spectacular Sea Caves Exploration		10	9	Italian Iced Coffee Experience
11	79	Historic Noble Palaces in Puglia		11	8	Shop Opening Hours Variability
12	78	Music Festivals in Puglia		12	8	Airport Transportation Services
13	63	Margherita di Savoia Salt Pans		13	8	Peak Season Tourism
14	63	Luxury Accommodation in Puglia		14	7	Celebrated Altamura Bread
15	56	Discover Barletta Itinerary Information		15	7	Hollywood Stars in Italian Masserie
16	54	Magnificent Decorated Rose Window		16	6	Pet Travel Regulations and Rules
17	46	Hiking Itinerary in Ostuni		17	6	Baroque Martina Franca exquisite
18	43	Coastal Defense Towers Salento				
19	42	Ancient Sea City Taranto				
20	31	Romanesque Cathedrals in Puglia				
21	26	Ancient Olive Grove Landscape				
22	24	Enchanting Coastal Gem Polignano				
23	24	Ancient Ravine Village Exploration				
24	23	Historic Salento Olive Oil Mills				
25	22	Baroque Beauty in Lecce				
26	22	Coastal Beauty Exploration				
27	20	Coastal Oasis in Southern Italy				
28	17	Theatres in Puglia				

Figure 4
OpenAI descriptions of topics in both corpora.

Their analysis revealed the multifaceted tourist gaze as constructed by the DMO in their websites to promote the visit to Puglia. In reading all the topic labels, it emerged that both corpora offer a rich overview of the region's cultural and experiential offerings. However, a different thematic emphasis is also evident.

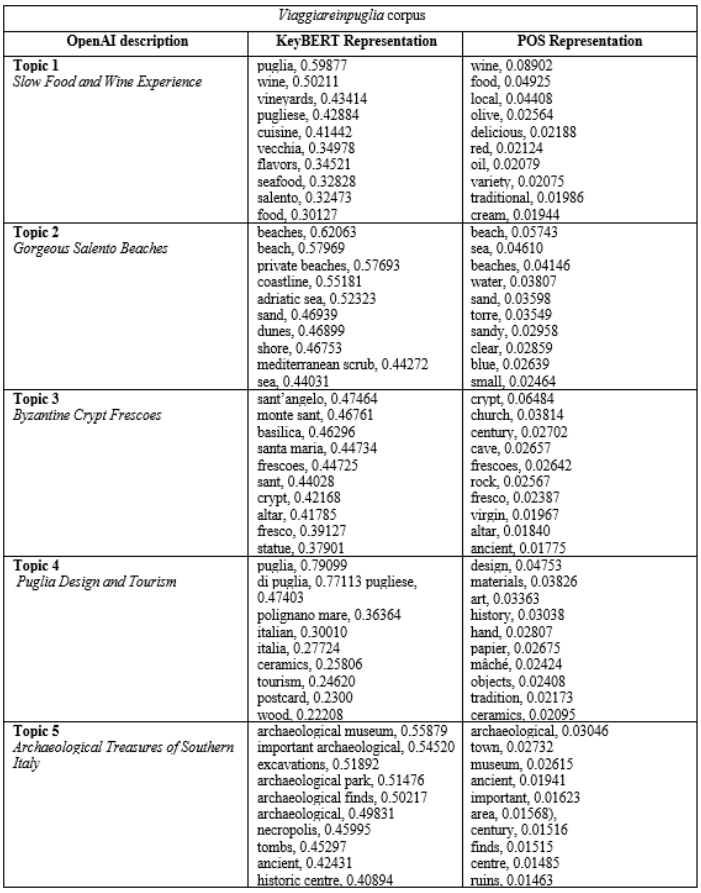
The topic labels in *Viaggiareinpuglia* unearthed themes deeply rooted in the region's history, architecture, culture and natural beauty and underline the persuasive language employed to depict Puglia. In particular, Topic 0 "*War-time Refugee Camps in Puglia*", Topic 3 "*Byzantine Crypt Frescoes*", Topic 5 "*Archaeological Treasures of Southern Italy*", Topic 8 "*Swabian Castles of Southern Italy*" or Topic 11 "*Historic Noble Palaces in Puglia*" highlight the region's deep historical roots, reflecting on its significant past events and architectural, artistic and cultural heritage. Similarly, Topic 2 "*Gorgeous Salento Beaches*", Topic 7 "*Scenic Coastal Pathways*", Topic 9 "*Enchanting Gargano Forest Reserve*", Topic 10 "*Spectacular Sea Caves Exploration*" or Topic 21 "*Ancient Olive Grove Landscape*" reveal the tourist gaze on the natural attractions of Puglia. Local traditions and cultural events also emerge as significant topics in Topic 1 "*Slow Food and Wine Experience*" or Topic 12 "*Music Festivals in Puglia*". Moreover, OpenAI descriptions included several towns and places as points of interest. One of the most mentioned areas is Salento with its beaches, coastal towers, olive oil mills and its Baroque capital Lecce. This reflects Salento's status as a key cultural and architectural hub in Puglia. Furthermore, the frequent mentions of natural parks and reserves reflect Puglia's diverse landscapes (Topic 9 and 13). The thematic diversity of *Viaggiareinpuglia* demonstrates the Puglia DMO's efforts to describe the region as a destination rich in culture, history and natural beauty. Additionally, each topic label's linguistic representation, with its carefully chosen descriptors and evocative language, shows the DMO's role in constructing a persuasive and not merely informative narrative. If the frequently mentioned adjectives *archaeological*, *historic* and *ancient* highlight the region's rich historical and cultural heritage, the inclusion of adjectives such as *scenic*, *enchanting*, *spectacular*, *magical* in topic labels invoke a sense of wonder. Also, the nouns *treasures*, *gem*, *beauty*, *oasis* convey a sense of peace, rarity and aesthetic pleasure.

Conversely, among the most representative topics in *Routes & Experiences*, only one topic, namely Topic 0 "*Historic Puglia Architecture*" captures the essence of the region's past and cultural heritage. Here, the projected destination image shifts towards a more experiential perspective, engaging activities and immersive cultural experiences. This is evident in Topic 1 "*Explore Puglia's Coastal Routes*", Topic 3 "*Puglia's Rich Festival Culture*", Topic 4 "*Artisanal Pasta Making Experience*" and Topic 11 "*Italian Iced Coffee Experience*". Moreover, this corpus well describes the nuances of Puglia's gastronomy (Topic 2, 7, 15 and 18) and artisanal crafts (Topic 10). More importantly, *Routes & Experiences* provides future tourists with a number of practical travel information, from logistical details about transportation and accommodations to insights into local customs and regulations (Topic 8, 9, 12, 13, 14, 17). Only one destination is mentioned in the OpenAI descriptions, namely Martina Franca in the Itria Valley (Topic 18). In addition, each topic label's linguistic representation in this corpus shows a strategic use of evocative and experiential language. Adjectives such as *authentic* and *artisanal* evoke experiences or products deeply rooted in local traditions and skills, whereas *handmade* and *Italian* emphasise the uniqueness and cultural origin of these experiences. The noun *escape* implies a departure from the ordinary and daily routines. The verb *explore* is action-oriented, implies adventure and suggests an active engagement with the environment and culture.

In conclusion, the analysis of the OpenAI descriptions showed that Puglia is distinctively presented as a destination of historical significance and natural beauty within *Viaggiareinpuglia*. Conversely, *Routes & Experiences* portrays Puglia as a dynamic playground that should encourage potential tourists to immerse themselves in the culture, traditions and landscape of Puglia.

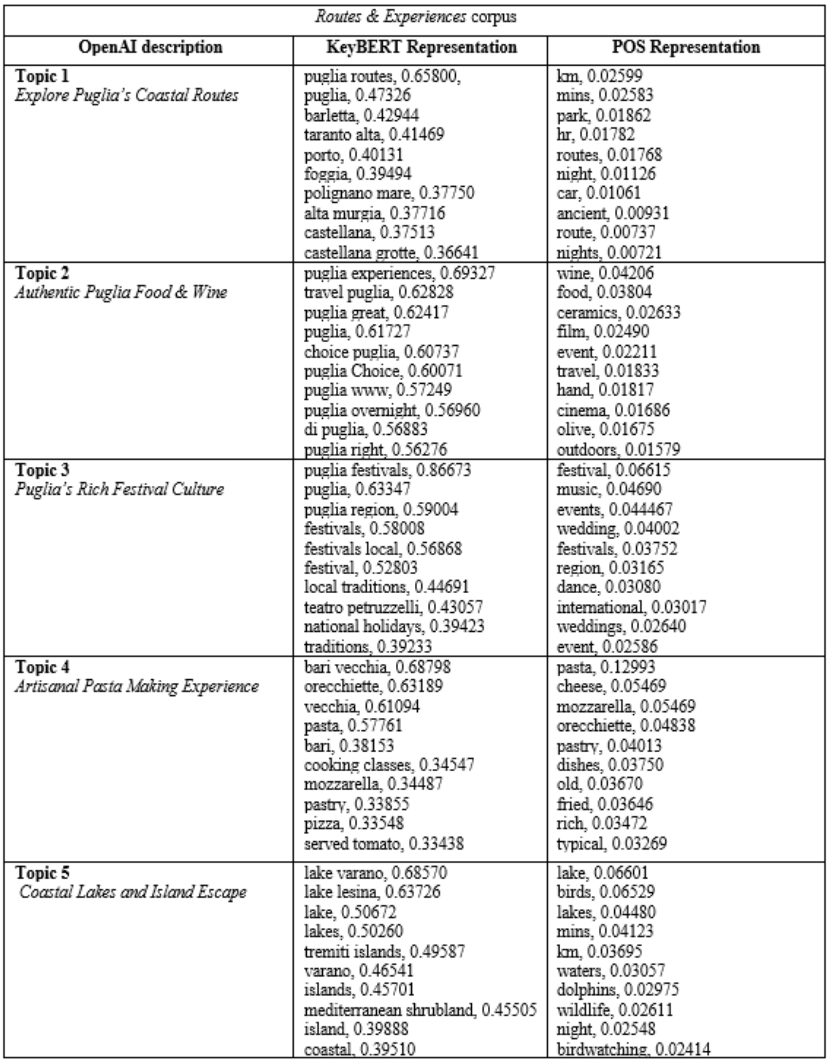
6.3. Comparison of KeyBERT and POS tagging representations

The results of the topic representation models, specifically KeyBERT and Part-Of-Speech tagging include two lists of the most representative words for each topic across both corpora, ranked according to their c-TF-IDF score. These scores quantitatively measure the words' relevance by assessing their frequency across the documents. Thanks to its use of semantic embeddings, KeyBERT extracts contextually rich keywords and phrases that are likely to provide more detailed insights into the text content and takes advantage of BERT's deep understanding of context and semantics. In contrast, POS tagging involves analysing the grammatical structure of text, categorizing words into nouns, verbs, adjectives. Each term is weighted by its significance in the topic, but without the same level of contextual depth as KeyBERT. This approach is more about identifying individual words that are statistically significant for each topic rather than the contextually rich phrases found with KeyBERT. To gain a deeper understanding of the difference between the two topic representation models, the top ten words derived from the KeyBERT and POS representations for the first five topics (Figures 5 and 6) were analysed.



Viaggiareinpuglia corpus		
OpenAI description	**KeyBERT Representation**	**POS Representation**
Topic 1 *Slow Food and Wine Experience*	puglia, 0.59877 wine, 0.50211 vineyards, 0.43414 pugliese, 0.42884 cuisine, 0.41442 vecchia, 0.34978 flavors, 0.34521 seafood, 0.32828 salento, 0.32473 food, 0.30127	wine, 0.08902 food, 0.04925 local, 0.04408 olive, 0.02564 delicious, 0.02188 red, 0.02124 oil, 0.02079 variety, 0.02075 traditional, 0.01986 cream, 0.01944
Topic 2 *Gorgeous Salento Beaches*	beaches, 0.62063 beach, 0.57969 private beaches, 0.57693 coastline, 0.55181 adriatic sea, 0.52323 sand, 0.46939 dunes, 0.46899 shore, 0.46753 mediterranean scrub, 0.44272 sea, 0.44031	beach, 0.05743 sea, 0.04610 beaches, 0.04146 water, 0.03807 sand, 0.03598 torre, 0.03549 sandy, 0.02958 clear, 0.02859 blue, 0.02639 small, 0.02464
Topic 3 *Byzantine Crypt Frescoes*	sant'angelo, 0.47464 monte sant, 0.46761 basilica, 0.46296 santa maria, 0.44734 frescoes, 0.44725 sant, 0.44028 crypt, 0.42168 altar, 0.41785 fresco, 0.39127 statue, 0.37901	crypt, 0.06484 church, 0.03814 century, 0.02702 cave, 0.02657 frescoes, 0.02642 rock, 0.02567 fresco, 0.02387 virgin, 0.01967 altar, 0.01840 ancient, 0.01775
Topic 4 *Puglia Design and Tourism*	puglia, 0.79099 di puglia, 0.77113 pugliese, 0.47403 polignano mare, 0.36364 italian, 0.30010 italia, 0.27724 ceramics, 0.25806 tourism, 0.24620 postcard, 0.2300 wood, 0.22208	design, 0.04753 materials, 0.03826 art, 0.03363 history, 0.03038 hand, 0.02807 papier, 0.02675 mâché, 0.02424 objects, 0.02408 tradition, 0.02173 ceramics, 0.02095
Topic 5 *Archaeological Treasures of Southern Italy*	archaeological museum, 0.55879 important archaeological, 0.54520 excavations, 0.51892 archaeological park, 0.51476 archaeological finds, 0.50217 archaeological, 0.49831 necropolis, 0.45995 tombs, 0.45297 ancient, 0.42431 historic centre, 0.40894	archaeological, 0.03046 town, 0.02732 museum, 0.02615 ancient, 0.01941 important, 0.01623 area, 0.01568), century, 0.01516 finds, 0.01515 centre, 0.01485 ruins, 0.01463

Figure 5
KeyBERT and POS Representations in the *Viaggiareinpuglia* corpus.



Routes & Experiences corpus		
OpenAI description	**KeyBERT Representation**	**POS Representation**
Topic 1 *Explore Puglia's Coastal Routes*	puglia routes, 0.65800, puglia, 0.47326 barletta, 0.42944 taranto alta, 0.41469 porto, 0.40131 foggia, 0.39494 polignano mare, 0.37750 alta murgia, 0.37716 castellana, 0.37513 castellana grotte, 0.36641	km, 0.02599 mins, 0.02583 park, 0.01862 hr, 0.01782 routes, 0.01768 night, 0.01126 car, 0.01061 ancient, 0.00931 route, 0.00737 nights, 0.00721
Topic 2 *Authentic Puglia Food & Wine*	puglia experiences, 0.69327 travel puglia, 0.62828 puglia great, 0.62417 puglia, 0.61727 choice puglia, 0.60737 puglia Choice, 0.60071 puglia www, 0.57249 puglia overnight, 0.56960 di puglia, 0.56883 puglia right, 0.56276	wine, 0.04206 food, 0.03804 ceramics, 0.02633 film, 0.02490 event, 0.02211 travel, 0.01833 hand, 0.01817 cinema, 0.01686 olive, 0.01675 outdoors, 0.01579
Topic 3 *Puglia's Rich Festival Culture*	puglia festivals, 0.86673 puglia, 0.63347 puglia region, 0.59004 festivals, 0.58008 festivals local, 0.56868 festival, 0.52803 local traditions, 0.44691 teatro petruzzelli, 0.43057 national holidays, 0.39423 traditions, 0.39233	festival, 0.06615 music, 0.04690 events, 0.044467 wedding, 0.04002 festivals, 0.03752 region, 0.03165 dance, 0.03080 international, 0.03017 weddings, 0.02640 event, 0.02586
Topic 4 *Artisanal Pasta Making Experience*	bari vecchia, 0.68798 orecchiette, 0.63189 vecchia, 0.61094 pasta, 0.57761 bari, 0.38153 cooking classes, 0.34547 mozzarella, 0.34487 pastry, 0.33855 pizza, 0.33548 served tomato, 0.33438	pasta, 0.12993 cheese, 0.05469 mozzarella, 0.05469 orecchiette, 0.04838 pastry, 0.04013 dishes, 0.03750 old, 0.03670 fried, 0.03646 rich, 0.03472 typical, 0.03269
Topic 5 *Coastal Lakes and Island Escape*	lake varano, 0.68570 lake lesina, 0.63726 lake, 0.50672 lakes, 0.50260 tremiti islands, 0.49587 varano, 0.46541 islands, 0.45701 mediterranean shrubland, 0.45505 island, 0.39888 coastal, 0.39510	lake, 0.06601 birds, 0.06529 lakes, 0.04480 mins, 0.04123 km, 0.03695 waters, 0.03057 dolphins, 0.02975 wildlife, 0.02611 night, 0.02548 birdwatching, 0.02414

Figure 6
KeyBERT and POS Representations in the *Routes & Experiences* corpus.

The comparison of the KeyBERT and POS tagging representations of the first five topics underlined some similarities and differences of the nuances of language processing techniques. Both methods effectively identify key aspects of each topic. For instance, in Topic "*Gorgeous Salento Beaches*" both representations list the terms *beach*, *sea* and *sand*, thus showing their ability to capture keywords relevant to the topic's core subject.

However, the two approaches provide distinct insights into text analysis. KeyBERT tends to extract specific keywords and named entities closely related to the topic, and to offer a focused perspective. For instance, it identified specific geographic locations such as *lake Varano* and *lake Lesina, Bari vecchia* or *teatro Petruzzelli*. It also highlighted thematic keywords such as *mediterranean shrubland, local traditions, Puglia experiences, cooking classes* and *national holidays*. On the other hand, POS tagging tends to highlight more conceptual or descriptive terms, such as *delicious, typical* and *traditional* or descriptive terms such as *birds* and *lakes*. It also provides a broader context and a general descriptive quality, as noticed with terms *clear* and *blue* for beaches, *pasta* and *cheese* for culinary experiences, *festival* and *music* when discussing the region's festival culture.

The difference in approach between KeyBERT's detailed specificity and POS tagging's thematic aggregation highlights the complementary nature of the two methods in extracting meaningful insights from text, each from its different perspective.

6.4. Analysis of the words of KeyBERT and POS tagging in the two corpora

The top ten words, ranked by their c-TF-IDF scores, from the KeyBERT and POS tagging representations in each corpus (Table 7) provided a comprehensive overview of the different thematic focus and language used by the Puglia DMO as marketing strategies to shape the tourist gaze.

	Viaggiareinpuglia corpus	*Routes & Experiences* corpus
KeyBERT Representation	basilica, adriatic, cathedral, sant, historic centre, santa maria, palazzo, piazza, monte, mediterranean.	puglia routes, puglia experiences, scenery, itria valley, porto cesareo, porto, pugliese, adriatic, beaches.
POS Representation	church, century, sea, town, ancient, city, old, small, centre area.	day, region, km, town, valley, local, best, routes, coast, experiences.

Table 7
The top ten words in each corpus.

Evidently, *Viaggiareinpuglia* predominantly features words that reflect a strong emphasis on cultural and historical sites, with references to specific attractions such as *basilica*, *cathedral* and *historic centre*. This emphasis suggests a marketing strategy focused on the region's historical and architectural assets. On the other hand, *Routes & Experiences* highlights the experiential and geographical aspects of tourism in Puglia, with frequent mentions of terms related to routes, scenery and outdoor activities. This approach appears to aim at promoting Puglia as a destination for experiential travel and natural exploration.

The linguistic analysis based on the top ten words from both the KeyBERT and POS tagging representations already well summarises the different strategies employed by the Puglia DMO to influence potential visitors' perceptions and expectations of Puglia as a travel destination. By examining all the terms identified in each topic, shown in bar charts (Charts 1-4), we have a nuanced understanding of how Puglia destination brand communicates through language.

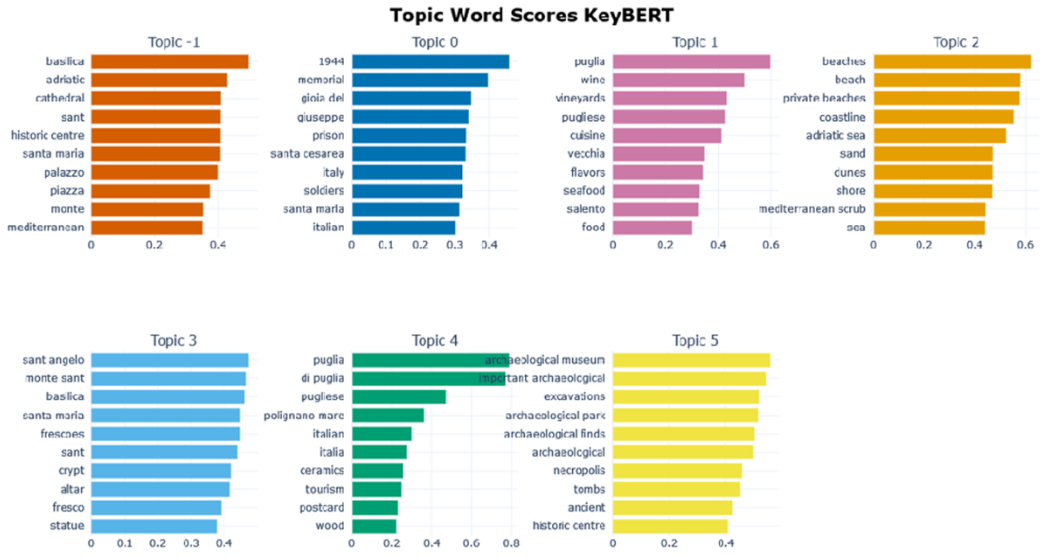


Chart 1
Topic word scores KeyBERT in *Viaggiareinpuglia*.

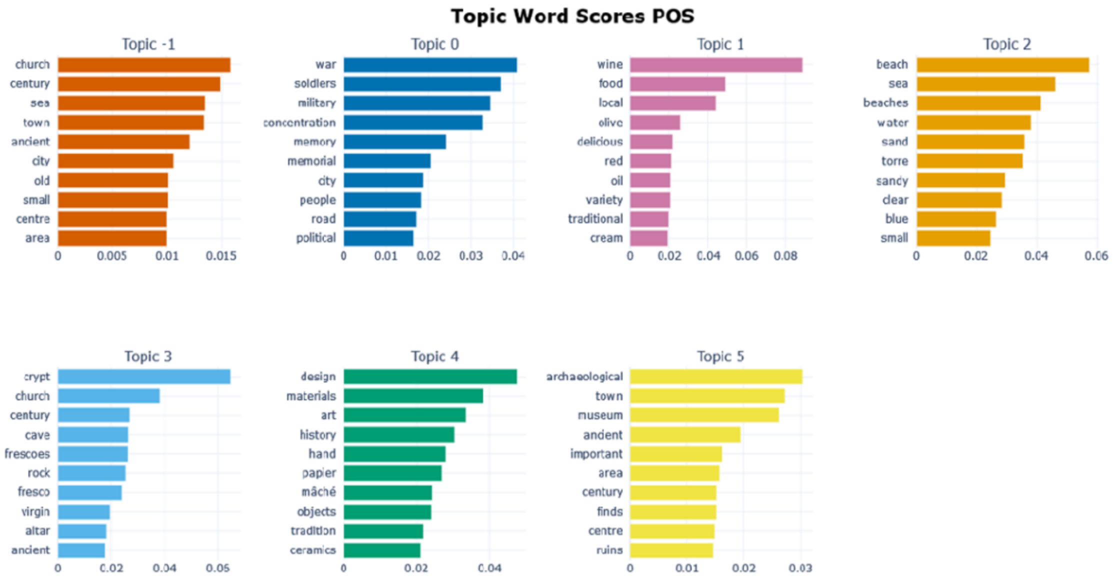


Chart 2
Topic word scores POS in *Viaggiareinpuglia*.

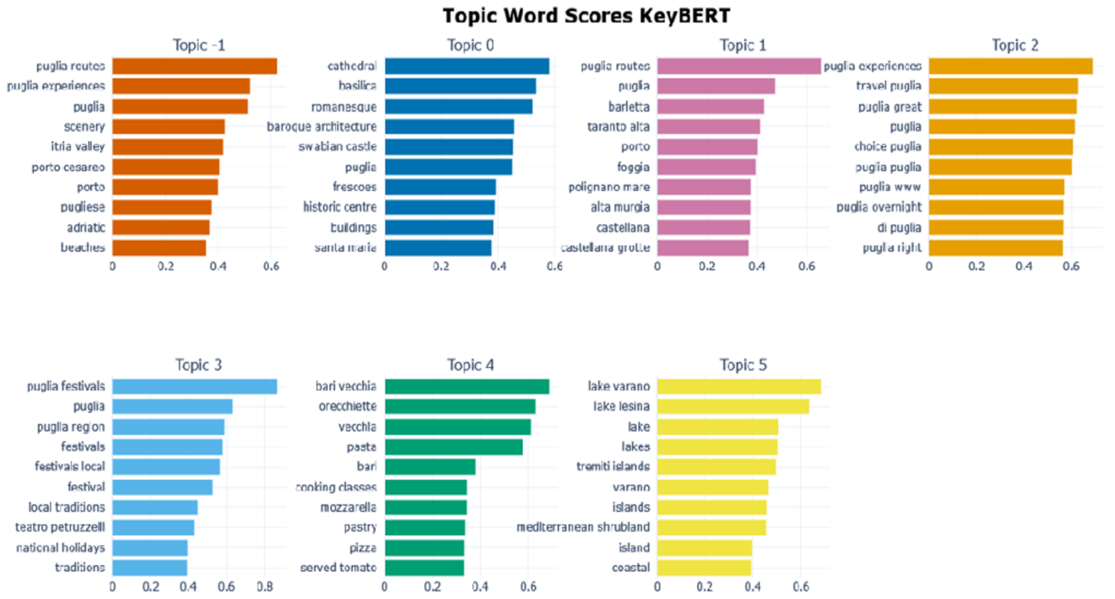


Chart 3
Topic word scores KeyBERT in *Routes & Experiences*.

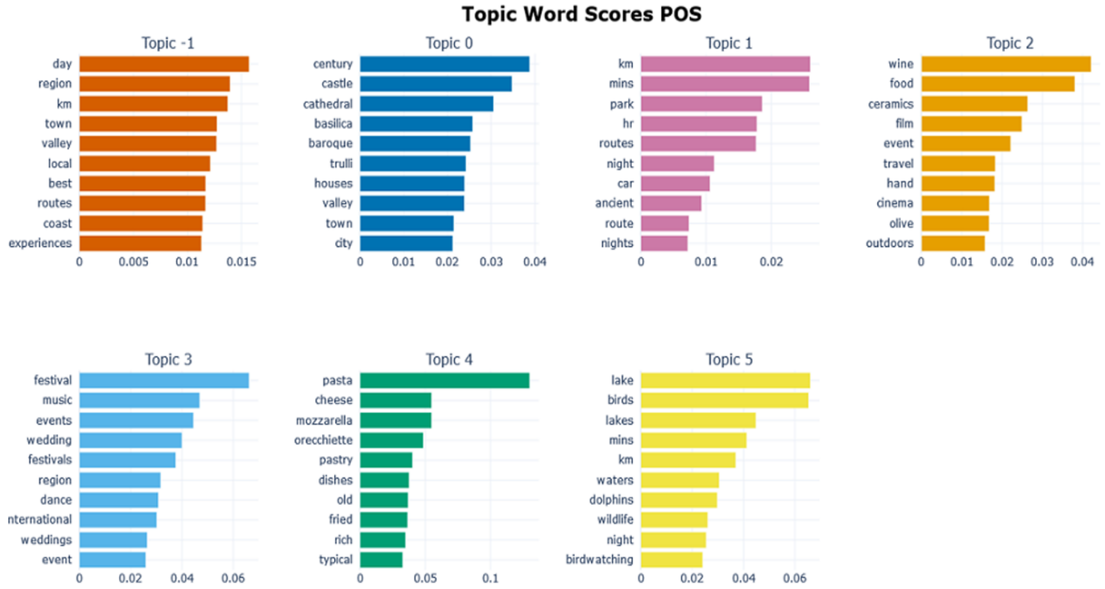


Chart 4
Topic word scores POS in *Routes & Experiences*.

In particular, we noticed that *Viaggiareinpuglia* seems to include a wide range of attractions from cultural and historical sites to natural landscapes, as seen in terms such as *basilica*, *Adriatic*, *cathedral*, *Santa Maria*, *palazzo*, *Baroque architecture*, *sea caves*, *archaeological museum*, *Swabian castle* from the KeyBERT findings and *church*, *century*, *sea*, *town*, *ancient*, *city*, *archaeological*, *museum*, *crypt* from the POS tagging findings. Moreover, KeyBERT's extraction of specific terms such as *basilica*, *vineyards*, *archaeological museum* and *Mediterranean scrub* and the adjectives *typical*, *real*, *original*, *ancient* and *authentic* describe Puglia's authentic aspects, appealing to tourists seeking experiences away from their routine. Similarly, the strangerhood perspective's emphasis on seeking different experiences is reflected in the qualifiers *ancient*, *traditional* and *baroque*, which echo the search for experiences that are *simple*, *remote* and *unchanging*. These terms, rooted in the language of tourism, illustrate the pull factor designed to attract tourists by highlighting the distinctive and unspoilt nature of the destination. The strangerhood dimension of Puglia is also created by mentioning nouns such as *grottoes* and *cave*, which serve as nouns that exoticize the region, highlighting elements that are unusual or extraordinary to foreign visitors.

The analysis of the most frequent terms in *Routes & Experiences*, meanwhile, confirmed a significant emphasis on the natural landscape and experiential travel and exploration, highlighted by terms such as *Puglia routes*, *Puglia experiences*, *Itria valley*, *Porto Cesareo*, *Adriatic beaches*, *artisanal pasta making*, *coastal lakes*, *island escape*, *lake Lesina* from the KeyBERT findings and *day*, *km*, *town*, *valley*, *routes*, *coast*, *experiences*, *lake*, *beaches* from the POS tagging findings.

Furthermore, the words *cathedral*, *Baroque architecture*, *Swabian castle*, *Itria valley* and *mediterranean* convey the geographical and historical authenticity of Puglia. The verbs *travel* and *explore* imply the act of discovering authentic places and the search for different and new experiences reflective of the strangerhood perspective. They also suggest a more dynamic interaction with the destination. Authentic aspects of the attractions are also conveyed through the adjectives *real*, *actual*, *primitive*, *simple* and *traditional*, while the adjectives *local* and *sweet* suggest a dive into less familiar but traditional aspects of the destination. In addition, *artisan crafts*, *handmade*, *artisanal*, *artisans* and *orecchiette* highlight unique local crafts, foods and tradition. Actions related to cultural and culinary experiences through the verbs *crafting* and *cooking classes* emphasize out-of-the-ordinary experiences and further exoticize the destination.

The KeyBERT and POS findings confirm the strategic use of language to promote the different attributes of Puglia and the differences in thematic focus.

7. Conclusions

This study showed the effectiveness of using the topic modeling technique BERTopic to examine the online destination image and tourist gaze as communicated by the Puglia DMO. By combining advanced computational methods such as OpenAI-generated descriptions, KeyBERT and POS tagging, the research highlighted differences in thematic focus, the language and strategy used to promote the region across two distinct institutional corpora. In particular, *Viaggiareinpuglia* mainly portrays Puglia as a historical landmark for its cultural and historical richness. In contrast, *Routes & Experiences* presents Puglia as a dynamic destination promoting experiential travel and the natural beauty of the region and emphasizing outdoor activities and immersive cultural experiences. However, while these thematic divergences are partly attributable to the original intended audiences – *Viaggiareinpuglia* for general international users and *Routes & Experiences* for a specialised international market – it is increasingly important for the DMO to project a coherent destination image across platforms. This is especially true now that *Routes & Experiences* has been incorporated into *Viaggiareinpuglia* to enhance the region's global visibility. Coherence in narrative and tone becomes not just a matter of stylistic consistency, but a strategic necessity for effective branding.

An important factor contributing to the differences identified is also the fact that *Viaggiareinpuglia* was originally written in Italian and subsequently translated, whereas *Routes & Experiences* was created directly in English by a native-speaking travel writer. Therefore, these differences may reflect deeper cultural perspectives: Italian and international writers may, indeed, "gaze" at the same place differently, as shaped by their socio-cultural lenses. This reinforces the importance of authorship and translation as mediators of the tourist gaze.

Another significant consideration is the temporal dimension of the content. *Viaggiareinpuglia* was created in 2015, whereas *Routes & Experiences* dates from 2021. The temporal gap between these materials coincides with shifts in tourism trends, from traditional sightseeing to more participatory and experiential travel, explaining the stronger emphasis on immersive and sensorial engagement with place in *Routes & Experiences*.

From a methodological point of view, this study offers important contributions. BERTopic proved to be a powerful tool for uncovering and interpreting latent themes, and extracting meaningful information in large collections of text data, thanks to its flexibility and adaptability to various datasets and preprocessing needs. More specifically, OpenAI descriptions within BERTopic helped the interpretability of topics through human-readable labels and the exploration of the corpora. Both the KeyBERT and POS tagging representations were useful to show the use of adjectives, nouns and verbs employed by the Puglia DMO to attract tourists seeking authentic experiences.

This innovative approach contributes to methodological advancements through integrating advanced computational techniques in the field of tourism and language of tourism studies. It also provides the Puglia DMO with data-driven insights to refine their marketing strategies, for example, harmonizing content strategies across platforms and languages while staying responsive to evolving tourist expectations.

However, the present methodology is not without limitations. One of the primary challenges is inherent in the topic modelling process itself. The number of topics to be extracted must be determined in advance, a task requiring both sensitivity and knowledge of appropriate hyperparameter tuning. This reliance on a priori parameter setting renders the outcome sensitive to the chosen hyperparameters. Such sensitivity may lead to variations in topic coherence and interpretability, particularly when the approach is applied across datasets with different linguistic features.

In conclusion, future studies might include fine-tuning pre-trained transformer models on tourism-specific corpora, enhancing the extraction of domain-specific features and improving the interpretability of the resulting topics.

Bionote: Angela D'Egidio holds a PhD in Linguistic, Literary and Intercultural Studies from the University of Salento, where she is currently a contract teacher of English and a research fellow in Translation and English Language (SSD L-LIN/12). Her research interests include corpus linguistics, intercultural studies, translation studies and semantic analysis, with a particular focus on tourism discourse and translation. She has taken part in international research projects such as DIETALY (Destination Italy in English Translation and Language over the Years). She is the author of articles in international journals analysing tourist language in English and Italian.

Author's address: angela.degidio@unisalento.it

References

Bi J.-W., Liu Y., Fan Z.-P. and Cambria E. 2019, *Modelling Customer Satisfaction from Online Reviews Using Ensemble Neural Network and Effect-Based Kano Model*, in "International Journal of Production Research" 57 [22], pp. 7068-7088. https://doi.org/10.1080/00207543.2019.1574989.

Blain C., Levy S.E. and Ritchie J.B. 2005, *Destination branding: Insights and practices from destination management organizations*, in "Journal of Travel Research" 43 [4], pp. 328-338. https://doi.org/10.1177/0047287505274646.

Cai G., Sun F. and Sha Y. 2018, *Interactive Visualization for Topic Model Curation*, in "CEUR workshop proceedings of the IUI workshops" 2068.

Cappelli G. 2006, *Sun, Sea, Sex and the Unspoilt Countryside. How the English Language makes Tourists out of Readers*, Pari Publishing, Pari.

Cappelli G. 2013, *Travelling words: Languaging in English tourism discourse*, in Yarrington A., Villani S. and Kelly J. (ed.), *Travels and translations. Anglo-Italian cultural transactions*, Rodopi, New York and Amsterdam, pp. 353-374.

Choi S., Lehto X.Y. and Morrison A.M. 2007, *Destination image representation on the web: content analysis of Macau travel related websites*, in "Tourism Management" 28 [1], pp. 118-129. https://doi.org/10.1016/j.tourman.2006.03.002.

D'Egidio A. 2019, *Investigating the Tourist Gaze through Automated Semantic Analysis of Corpora*, Atena Libri, Lecce.

Daenekindt S. and Huisman J. 2020, *Mapping the Scattered Field of Research on Higher Education. A Correlated Topic Model of 17,000 Articles, 1991-2018*, in "Higher Education" 80 [3], pp. 571-587. https://doi.org/10.1007/s10734-020-00500-x.

Dann G.M.S. 1996, *The language of tourism: A sociolinguistic perspective*, Cab International, Oxon.

Denti L. and Fodde A. 2017, *What is Sardinia's Destiny? Cultural Heritage and Cross-cultural Tourist Marketing Constraints in Institutional Communication*, in "Letterature Straniere" 17 [XVII], pp. 101-119.

Echtner C.M. and Prasad P. 2003, *The Context of Third World Tourism Marketing*, in "Annals of Tourism Research" 30 [3], pp. 660-682.

Egger R. 2022, *Topic modelling: modelling hidden semantic structures in textual data*, in Egger R. (ed.), *Applied data science in tourism. Interdisciplinary approaches, methodologies, and applications*, Springer, Cham, pp. 375-403.

Francesconi S. 2012, *Generic integrity and innovation in tourism texts in English*, Tangram Edizioni Scientifiche, Trento.

Francesconi S. 2014, *Reading tourism texts: A multimodal analysis*, Channel view publications, Bristol.

Gotti M. 2006, *The Language of Tourism as Specialised Discourse*, in Palusci O. and Francesconi S. (ed.), *Translating Tourism. Linguistic/Cultural Representations*, Editrice Università degli Studi di Trento, Trento, pp. 15-34.

Grootendorst M. 2022, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*, pp. 1-10. https://doi.org/10.48550/arXiv.2203.05794. (01.03.2024).

Jin Y. 2022, *Travel Guide Using Text Mining and BERTopic*, Master's thesis, University of California, Los Angeles. https://escholarship.org/uc/item/69m2f87d.

Katan D. 2012, *Translating the tourist gaze: from heritage and 'culture' to actual encounter*, in "PASOS Revista de Turismo y Patrimonio Cultural" 10 [4], pp. 83-95.

Kim K., Park O., Barr J. and Yun H. 2019, *Tourists' Shifting Perceptions of UNESCO Heritage Sites: Lessons from Jeju Island-South Korea*, in "Tourism Review" 74 [1], pp. 20-29. https://doi.org/10.1108/TR-09-2017-0140.

Kolcun M. and Grabara J. 2014, *Use of elements of semiotic language in tourism marketing*, in "International Letters of Social and Humanistic Sciences" 15 [1], pp. 1-6.

Larsen J. and Urry J. 2011, *Gazing and performing*, in "Environment and Planning D: Society and space" 29 [6], pp. 1110-1125.

Maci S.M. 2020, *English Tourism Discourse: Insights into the Professional, Promotional and Digital Language of Tourism*, Hoepli Editore, Milan.

Manca E. 2016, *Official Tourist Websites and the Cultural Communication Grammar Model: Analysing Language, Visuals, and Cultural Features*, in "Cultus" 9 [1], pp. 2-22.

Manca E. 2018, *Verbal Techniques of the Language of Tourism across Cultures: An Analysis of Five Official Tourist Websites*, in Garzone G. and Catenaccio M. (eds.), *Innovative perspectives on tourism discourse*, IGI Global, Hershey (PA), pp. 91-110.

Marchi V. and Raschi A. 2022, *Measuring destination image of an Italian island: An analysis of online *content generated by local operators and tourists*, in "Island Studies Journal" 17 [1], pp. 259-279.

Mayo E.J. 1973, *Regional images and regional travel behavior. Research for changing travel patterns: Interpretation and utilization*. Paper presented at the Travel Research Association 4th Annual Conference, Sun Valley, Idaho.

McInnes L., Healy J. and Astels S. 2017, *HDBSCAN: Hierarchical Density Based Clustering*, in "The Journal of Open Source Software" 2 [11], article 205.

Pandove D., Goel S. and Rani R. 2018, *Correlation clustering methodologies and their fundamental results*, in "Expert Systems" 35 [1], e12229.

Pike S. 2002, *Destination Image Analysis: A Review of 142 Papers from 1973 to 2000*, in "Tourism Management" 23 [5], pp. 541-549.

Regione Puglia, *I Dati Turistici della Puglia nel 2015*. https://aret.regione.puglia.it/en/dati-e-ricerche/rapporti-e-statistiche/dettaglio/-/asset_publisher/c2UOAprz5aIk/content/id/153399/i-dati-turistici-della-puglia-nel-2015. (15.03.2024)

Regione Puglia, *I Trend del Turismo in Puglia nel 2023*. https://aret.regione.puglia.it/en/dati-e-ricerche/rapporti-e-statistiche/dettaglio/-/asset_publisher/c2UOAprz5aIk/content/id/3564085/i-trend-del-turismo-in-puglia-nel-2023. (15.03.2024)

Reimers N. and Gurevych I. 2019, *SentenceBERT: Sentence Embeddings Using Siamese BERT-Networks*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.

Shafqat W. and Byun Y.C. 2020, *A Recommendation Mechanism for Under-Emphasized Tourist Spots Using Topic Modeling and Sentiment Analysis*, in "Sustainability" 12 [1], article 320. https://doi.org/10.3390/su12010320.

Sharma P. and Li Y. 2019, *Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling*, Preprints, https://doi.org/10.20944/preprints201908.0073.v1

Silva C.C., Galster M. and Gilson F. 2021, *Topic modeling in software engineering research*, in "Empir Software Eng" 26 [6], pp. 1-62. https://doi.org/10.1007/s10664-021-10026-0.

Stepchenkova S. and Zhan F. 2013, *Visual destination images of Peru: Comparative content analysis of DMO and user-generated photography*, in "Tourism management" 36 [1], pp. 590-601.

Stone L.S. and Nyaupane G.P. 2019, *The Tourist Gaze: Domestic versus International Tourists*, in "Journal of Travel Research" 58 [5], pp. 877-891.

Sulaiman M.Z. and Wilson R. 2019, *Tourism Promotional Materials*, in Sulaiman M.Z. and Wilson R. (ed.), *Translation and Tourism: Strategies for Effective Cross-Cultural Promotion*, pp. 17-35.

Tasci A.D.A., Gartner W.C. and Cavusgil S.T. 2007, *Conceptualization and Operationalization of Destination Image*, in "Journal of Hospitality & Tourism Research" 31 [2], pp. 194-223.

Urry J. and Larsen J. 2011, *The tourist gaze 3.0*, Sage Publications, Los Angeles.

Vu H.Q., Li G. and Law R. 2019, *Discovering Implicit Activity Preferences in Travel Itineraries by Topic Modeling*, in "Tourism Management" 75, pp. 435-446. https://doi.org/10.1016/j.tourman.2019.06.011.

Weber R.P. 1990, *Basic content analysis*, Sage Publications, London.