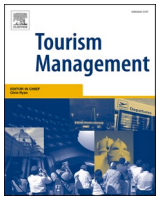
Tourism Management 106 (2025) 105007

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman



Facilitating topic modeling in tourism research:Comprehensive comparison of new AI technologies

Andrei P. Kirilenko *, Svetlana Stepchenkova

Department of Tourism, Hospitality and Event Management, University of Florida, Gainesville, FL, 32611-8209, USA

A B S T R A C T

In the past few years, a new crop of transformer-based language models such as Google's BERT and OpenAI's ChatGPT has become increasingly popular in text analysis, owing their success to their ability to capture the entire document's context. These new methods, however, have yet to percolate into tourism academic literature. This paper aims to fill in this gap by providing a comparative analysis of these instruments against the commonly used Latent Dirichlet Allocation for topic extraction of contrasting tourism-related data: coherent vs. noisy, short vs. long, and small vs. large corpus size. The data are typical of tourism literature and include comments of followers of a popular blogger, TripAdvisor reviews, and review titles. We provide recommendations of data domains where the review methods demonstrate the best performance, consider success dimensions, and discuss each method's strong and weak sides. In general, GPT tends to return comprehensive, highly interpretable, and relevant to the real-world topics for all datasets, including the noisy ones, and at all scales. Meanwhile, ChatGPT is the most vulnerable to the issue of trust common to the "black box" model, which we explore in detail.

1. Introduction

Tourism and hospitality methodologies have always been centered around systematic analysis of texts, images, or other media. Di Maggio et al. (2013) highlight three main analytic approaches. The first one is to produce "virtuoso interpretations" based on a simple text reading, image observation, etc. This approach is severely limited (DiMaggio et al., 2013) in its ability to address the problem of the "replication crisis" in social science (Open Science Collaboration, 2015). The second approach, championed by Holsti (1969), is based on a systematic text reading, producing a list of common topics (e.g., based on theoretical insights or research questions), creating a coding table, text coding, and results validation. There are two major limitations to this approach (DiMaggio et al., 2013). First, it is unlikely to generate topics outside the researcher's expertise. Second, processing large volumes of data is beyond its' capability. The latter limitation is especially important in the current era of media digitization producing corpora that are unreasonable for manual processing. Thus, the final approach utilizes automated text processing.

The first attempts in computer-based topic modeling started as early as in 1960s (Harway & Iker, 1964; Iker & Harway, 1965; Miles & Selvin, 1966, pp. 116–127 and were based on the earlier suggestions of analysis of word distribution from IBM (Luhn, 1957, 1958), inve stigation of "greater-than-chance" word co-occurrences (Osgood & Walker, 1959), and pioneering publications by Borko introducing a methodology of applying factor analysis to text modeling (Borko, 1962) and automated document classification (Borko, 1961; Borko & Bernick, 1963).

The original IBM research tackled a problem specific to the time when reading discipline-related journals of abstracts was a must for scientists as they tried to keep up to date with relevant research published in multiple outlets around the globe. Such journals frequently hired personnel to read and summarize scientific papers; the resulting abstract would then be published. However, "The abstracter's product is almost always influenced by his background, attitude, and disposition." (Luhn, 1958, p. 159). The IBM proposal was to create "auto-abstracts" of scientific papers in machine-readable form by selecting the most representative sentences of a paper. First, the list of the most significant words and word combinations would be created based on their frequency (but excluding the common words such as articles) and accounting for inflections. The presence and distance between these significant words would then be used to capture the "significant sentences" from the text, hence generating the abstract. While simplistic in terms of statistical analysis, the approach has resulted in a surprisingly coherent text summarization, as evident from sample abstracts provided in the publication.

The abovementioned simplicity of statistical measures was tackled by Osgood and Walker (1959), who suggested a large battery of statistical measures to capture person-specific text patterns. The immediate research goal was to reliably identify the writing of people in danger of committing suicide. The proposed mechanical approaches included text

* Corresponding author.
E-mail addresses: andrei.kirilenko@ufl.edu (A.P. Kirilenko), svetlana.step@ufl.edu (S. Stepchenkova).

https://doi.org/10.1016/j.tourman.2024.105007
Received 12 February 2024; Received in revised form 12 July 2024; Accepted 19 July 2024
Available online 31 July 2024

redundancy, "filling the blanks" (predictability of a blanked word based on its surroundings), counting of syllabi and words per sentence, and many others.

One particular problem with the abovementioned scholarship was the sheer number of unique words encountered in texts. Borko (1962) successfully dealt with the resulting high-dimensionality problems of linguistic analysis by demonstrating the effectiveness of exploratory factor analysis in mapping the word-space into concept-space. The derived factors were interpretable as the main topics of the analyzed text. These early ideas are still relevant for today's automated text and image processing and their traces appear in data pre-processing, statistical analysis, and even in word embedding and transformer architectures.

While these early approaches are still effective, especially when applied to texts using consistent terminology and eloquent language (Ma & Kirilenko, 2020), there have been developed multiple other methodologies targeting summarizing text meaning into a limited set of topics. A review of these approaches is found elsewhere (Churchill & Singh, 2022; Vayansky & Kumar, 2020).

In tourism literature, the most popular (according to Egger & Yu, 2022) text modeling approach is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This is not surprising given that LDA has been used in 80% of social media studies outside the computer science discipline (Laureate et al., 2023). The popularity of LDA, however, does not translate into the method's universality. The fundamental idea behind LDA is that each document is a mixture of hidden topics, and the topics are mixtures of words, both following a Dirichlet probability distribution. The model aims to find the hidden topics that best explain the observed word pattern in the documents. This framework dictates important LDA limitations that are often disregarded by researchers. To begin with, LDA relies on the accurate estimation of parameters related to the distribution of topics over documents and the distribution of words over the topics, requiring the existence of sufficiently lengthy documents capable of effectively encompassing a diverse range of topics. Additionally, the LDA algorithm necessitates a substantial amount of textual data and substantially long documents (Laureate et al., 2023) to ensure precise inference of the underlying topic distributions. The presence of discordant or extraneous documents[1] not well aligned with common topics present in data, a common occurrence in social media datasets, significantly undermines the quality of the inferred topics. Some of these limitations have been addressed in method extensions (hybrid-LDA, supervised LDA, hierarchical LDA, and many others), reviewed by Jelodar et al. (2019), but, fundamentally, the approach is still subject to criticism due to its intrinsic instability and sensitivity to data noise.

In the past few years, a new crop of transformer-based language models such as Google's Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and OpenAI's Generative Pre-trained Transformer (GPT)[2] (Radford et al., 2018, 2019) has become increasingly popular. In leading tourism journals (*Tourism Management* and *Journal of Travel Research*), however, there is only a handful of examples of BERT (Zhang et al., 2023) and none of GPT topic analysis (note that several authors used GPT and BERT for sentiment analysis), but it is only a matter of time when studies using these methods will appear in the tourism literature in substantial numbers. Yet, it is highly problematic to advance theoretical knowledge and provide practical recommendations with real life consequences for the economic and social wellbeing of people on potentially shaky ground of a methodology that has not been properly tested with respect to its suitability for various types of tourism-related data. Therefore, this paper makes an important step in the direction of cross comparing the three large classes of automated methods for text analyses, LDA, BERT, and GPT, for applicability in tourism research. We provide a comprehensive comparative analysis of contrasting tourism-related types of data along three dimensions: structured and coherent vs. poorly structured and noisy records, short vs. long records, and small vs. large document corpus sizes. Here, we use the term "comprehensive" to underscore the variety of textual data types examined, each data type presenting its own challenges for analysis as explained in detail in sections 2.1-2.4.

We developed five criteria to compare algorithm performance.

- *Effective topic extraction:* extracted topics should be coherent, meaningful, and interpretable in terms of real-world tourism-related concepts;
- *Thematic representation of document collection:* a majority of the documents in the corpus should be distinctly associated with the extracted topics, reinforcing the algorithm's efficacy in capturing the predominant themes present in the dataset;
- *Scalability:* the algorithm should be able to process a diverse range of data, from short to long documents and from small (hundreds of records) to large numbers of documents;
- *Robustness:* the algorithm should be able to process noisy and incomplete texts typical for tourism-related social media data, returning consistent results across multiple runs; and
- *Explainability:* a researcher should be able to understand and explain how the topics were obtained, reinforcing trust in results.

The first dimension of *effective topic extraction* is self-explanatory as it directly relates to the goal of topic modeling. A mathematically perfectly derived topic (for example, LDA returns a topic in the form of statistically related keywords) is useless unless a human analyst is able to make sense of it. The *thematic representation* dimension characterizes a different side of the goal of topic modeling: a human analyst does not only want to derive meaningful topics from a collection of documents, but they also want the derived set of topics to cover the entire dataset. While this goal is hard to fully achieve, e.g., due to the presence of discordant and extraneous documents, the percentage of documents mapped to well-interpretable topics should be as high as possible. Note however that while these two dimensions are complementary in their relation to the overall goal of topic modeling, they are competing in terms of algorithmic implementation (Blei, 2012). As the percentage of documents related to a specific topic increases, the topics become more and more fine-grained (related to fewer and fewer documents), becoming less useful for human interpretation (Chang et al., 2009). While there are statistical measures such as topic coherence[3] that allow evaluation of topic evaluation, we evaluated the interpretability of extracted topics manually using criteria suggested by Mimno et al. (2011, pp. 262–272). The reason for our decision was to ensure that the topics make sense not only mathematically but also to a human. The percentage of documents related to non-interpretable topics was then used as a measure for the second dimension.

Scalability and *robustness* are derivatives of the first two dimensions as they affect both topic extraction and thematic representation. Noisy data containing misspellings, slang, abbreviations, pause-fillings ("um",

[1] Discordant documents differ significantly in content from most documents within a dataset while extraneous documents are irrelevant. For example, a review focused on a guest's experience in a tour bought at a hotel would be discordant in a hotel review dataset, while a message "The weather was nice" would be extraneous. The difference is that discordant documents may still be relevant in a larger context.

[2] There are multiple versions of OpenAI's GPT models; well-known ChatGPT is a frontend to GPT-3 (free version) and GPT-4 (paid subscription). For clarity, we use the name GPT to refer to either ChatGPT, GPT-3, or GPT-4 unless the difference between the models is important. This study uses GPT-4 model.

[3] Topic coherence refers to the degree to which the topics make sense together and can be understood as distinct themes. Mathematically, topic coherence measures the likelihood of topic words to co-occur in the same context compared to random co-occurrence. There are several similar coherence measures, the most popular in the literature being the U-mass coherence.

"uh"), spam, etc. is common in social media (Agarwal et al., 2007); robust algorithms should be able to successfully tackle noisy data without substantial reduction in the quality of extracted topics (Agarwal et al., 2007). The problem of noise becomes especially acute when analyzed documents are short (Li et al., 2018) or when the number of documents is small (Tran et al., 2013); scalable algorithms are able to produce meaningful topics from short and long texts as well as corpora of small and large sizes.

Finally, the *explainability* of the process in which the topics were derived makes a topic model trustworthy (Linardatos et al., 2020). Unfortunately, machine learning models are "black box", which makes their mechanism poorly explainable. Still, some models such as LDA, are functioning on widely understood principles, producing a list of keywords which are then interpreted by a human. Comparatively, the way results are produced by ChatGPT are very hard to impossible to explain.

The rest of the paper is organized as follows. First, we introduce the data challenges and topic models in the Literature review. Then, we present the data and our research methodology. Next, we provide details on text analysis using three different text model classes: LDA, BERT, and GPT. Finally, research implications are reviewed in the Discussion section.

2. Literature review

2.1. Data challenges

Topic modeling algorithms excel in extracting latent structures from text data characterized by well-defined topics, clear topical coherence (that is, the semantic similarity between the topics), contextual continuity, and consistent terminology. In the tourism and travel context, such data are frequently found in shared stories detailing personal experiences at specific destinations. Conversely, topic modeling becomes problematic when dealing with noisy, heterogeneous (covering multiple loosely related themes), or disorganized textual corpora containing colloquial language, ambiguous semantics, and frequent topic shifts. On social media, fans' comments on posts by social media influencers introduce additional hurdles due to the prevalence of slang, abbreviations, and contextually dependent language, making it harder for algorithms to discern coherent topics amidst the noise. The ability to extract main discussion points from both the coherent and heterogeneous content types is equally important for automated topic modeling.

Another dimension of corpus complexity relevant to topic modeling is document length. Short documents such as tweets typically address only one point and can be considered to reflect a single topic. On the other hand, short documents tend to be sparse, with fewer words and less redundancy. This sparsity can result in less reliable statistical patterns, making it harder to discern meaningful topics. Nevertheless, some methods such as principal component analysis (PCA), Non-Negative Matrix Factorization (NMF), and LDA have demonstrated their ability to extract content from short documents successfully (Albalawi et al., 2020). Conversely, long documents, such as bloggers' online stories, present the challenge of "information overload" by combining numerous topics.

Finally, the size of the corpus, that is, the number of documents, plays a pivotal role in the effectiveness of topic modeling. A larger corpus provides a more comprehensive and representative sample of the underlying content, allowing algorithms to capture a broader range of themes and patterns. In a substantial corpus, the frequency and co-occurrence of words across diverse documents offer a more stable foundation for identifying latent topics, enhancing the robustness of the model. Additionally, a sizable corpus reduces the effect of noise and contributes to better generalization, enabling the model to discern overarching topics that are not merely artifacts of specific document subsets.

2.2. LDA

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a probabilistic model which considers each document a mixture of topics, distributed according to Dirichlet distribution. Similarly, each topic is considered a mixture of words with Dirichlet frequency distribution. LDA approach has proposed an efficient way of learning the hidden distributions; the word distribution over topics is then interpreted in terms of the main themes hidden in the collection of documents, while topic distribution over documents helps to understand the statistical properties of the entire collection.

LDA is widely used by tourism and hospitality academics and practitioners. It was successfully utilized by Guo et al. (2017) to extract dimensions of customer satisfaction from online hotel reviews. Kirilenko et al. (2021) used LDA to highlight challenges in the analysis of reviews posted by dissatisfied travelers. Williams et al. (2023) applied LDA to research online representations of war tourism in Ukraine and found a new form of "hybrid war tourism". Jia (2020) employed the method to compare the motivation and satisfaction of Chinese and U.S. tourists in restaurants. Overall, we counted 29 papers using LDA in the Tourism Management journal alone. A recent bibliometric analysis of tourism and hospitality research by Koseoglu et al. (2022) confirms this observation by stating that LDA "is a core and commonly used model within the topic modeling family" (p. 318).

An important aspect of LDA's popularity is its wide availability through a range of tools with user-friendly interfaces. For instance, LDA is the topic learning method behind Qualtrics survey analysis, frequently employed by social scientists. For more advanced users, LDA is available through a variety of software libraries for R and Python enthusiasts, as well as online through Amazon Web Services. LDA however poses numerous challenges, being computationally expensive and sensitive to parameter selection. Since LDA is a probabilistic method, it returns a different solution every time the method is used, making results non-reproducible (note however that robust solutions are typically very similar). Finally, from a real-world perspective, LDA is unable to capture the semantics of natural language, which requires careful data pre-processing joining semantically close terms, correcting typos, removing stop words, etc. Some of the LDA limitations were relaxed in multiple method modifications, yet the latter core challenge related to semantics was unaddressed. Importantly, LDA assumptions and limitations rarely if ever are accounted for in tourism literature and those academics who use LDA algorithms through other software may even be unaware of such.

2.3. BERT

The introduction of word embedding and transformers was the major development in topic modeling that addressed the inability of topic modeling methods to capture text semantics. Word embedding is a method of capturing the context of document words, making it possible to understand the semantic relationship between the words. Consider the following sentence in a review about Miami Beach destination: "Beach sand was crispy." LDA, similar to other popular approaches, starts with coding each word in a dataset: for example, the word "beach" may be coded as "1", "sand" as 3, and "crispy" as 5 (the word "was" would typically be discarded during pre-processing). Codes 2, 4, and others would be assigned to words from other reviews. Then, the sentence would be coded as "1010100 … 0" with the number "1" in positions 1, 3, and 5 indicating the presence of words "beach", "sand", and "crispy" in the sentence and 0s in other positions indicating the absence of other words. The entire dataset then would be represented by an $m \times n$ matrix where m is the number of reviews and n is the number of unique words in reviews.[4] One problem with this "bag of words" approach is

[4] Other ways of matrix representation are available.

that the connection between the words is not maintained. The second problem is that m tends to be very large, making the analysis challenging. LDA (and similar algorithms) deals with the latter problem by introducing a parameter limiting the value of n and discarding rare words, but the former issue is unaddressed. The embedding mechanism, in contrast, deals with both problems by representing each word with a vector of rational numbers instead of an integer index. The specific representation of a word is selected by a machine learning algorithm in such a way that semantically similar words have geometrically close vector representation. For example, the words "beach" and "sand" are likely to appear in the same context, making them semantically close. These words will be represented with similar numerical vectors. On the opposite, the words "beach" and "crispy" are unlikely to appear together in other documents; thus, they will be represented by dissimilar vectors. The vector length is selected to be much smaller than m, typically 50 to 300, solving the dimensionality problem. Popular models such as Word2Vec, GloVe, and FastText have been successfully used to preprocess texts, which were then analyzed using traditional topic modeling methods (Luo et al., 2021). In the tourism context, Egger (2024) has an excellent discussion on using embedding for tourist segmentation.

Transformers, first introduced by the Google Brain team, are neural network algorithms that build up on the word embedding ideas by introducing additional novel mechanisms. Instead of analyzing each word of a text separately like in LDA, the transformers analyze the entire sentence at once. The second improvement is the "self-attention" mechanism (Vaswani et al., 2017), which forces the model to analyze connections between the separate words, which in turn helps the model to learn the context in which the words are used. The self-attention apparatus is central to transformers' capabilities in generating new texts or extracting text meaning. Importantly, transformer model training can be parallelized, making it possible to train models on very large datasets (0.5 trillion tokens for GPT-3 (Brown et al., 2020)).

Both BERT and GPT use Transformers architecture, albeit in a different way. BERT, developed by Google, focuses on the bidirectional context in its training process. Here, "bidirectional" means that in the process of training, BERT is masking random words in sentences and trains the algorithm to predict those masked words based on the surrounding words. This approach allows BERT to "understand" the context in which a specific word is appearing in documents, which is especially useful when masked words have multiple meanings. BERTopic uses a pre-trained BERT model to convert the input documents into numerical vectors with values capturing the context in which the words appear. Then, the dimensionality of these embeddings is reduced. Finally, the embeddings are used to cluster the documents and each cluster is assigned a name.

Successful application of BERT depends on the match between the corpus that the model was trained on and the analyzed documents' domain. In this report, we used a multi-purpose model all-MiniLM-L6-v2 trained on over 1 billion sentences from a wide variety of media, which is considered a good trade-off between accuracy and performance (Grootendorst, 2022). Multiple other sentence embedding models have been developed to capture the context specific to domains of interest, including finances, medicine, and computational biology. In the tourism domain, TourBERT (Arefeva & Egger, 2022) was trained on tourism reviews and descriptions of tourist services, attractions, and sights and was reported to demonstrate superior performance compared to the base version. Unfortunately, experiences in using TourBERT outside this research group are very limited and inconclusive (Carrillo et al., 2023). We experimented with TourBERT analysis of our data but did not find significant differences in results compared with BERTopic. Very few examples of using TourBERT in tourism area include a comparative study of different algorithms applied to Twitter data (Egger & Yu, 2022), a contrasting research of Airbnb reviews in urban and rural locations (Sánchez-Franco & Rey-Moreno, 2022), and an investigation into concerns over the "cannabis tourism" (Lerksuthirat et al., 2023).

2.4. GPT

LDA, BERT, as well as many other conventional topic modeling approaches return analysis results in the form of a list of keywords describing the topics. This list further needs an expert interpretation to produce a meaningful topic. Frequently, the quality of interpretation is low, which is "the biggest hurdle for accepting statistical topic models" (Egger & Yu, 2021). As opposed to BERT, discussed in the previous section, GPT, developed by OpenAI, is trained to predict the next words in a sequence. This unidirectional training enables GPT to generate a coherent text, relevant to the context provided in a prompt. This is why GPT-type models, while not specifically designed for topic modeling, excel in the summarization of the documents in the form of a coherent text rather than a list of keywords.

Exploratory studies have confirmed the ability of GPT models to produce meaningful topics when applied to Amazon product reviews (Thompson & Mimno, 2020) and political speech (in combination with LDA) (Shrestha et al., 2023), as well as in topic detection, e.g., in document classification according to their genre (Kuzman et al., 2023), presence of hate speech (Huang et al., 2023), or political stance (Zhang et al., 2022). We, however, are unaware of any related publication dealing with tourism and hospitality data.

3. Data

The data were selected with two considerations in mind. First, datasets need to be tourism-related data, preferably in the form of user-generated content (UGC) as it is a widely used source of data for tourism and hospitality research (Lu & Stepchenkova, 2015). Second, the data should represent all three dimensions of data type – text coherence, document length, and corpus size. We chose stories of travel experiences and comments on social networks, and six datasets were obtained from TripAdvisor, YouTube, and Weibo platforms. The TripAdvisor data related to two of Costa Rica's destinations: the Arenal volcano and the Corcovado reserve. TripAdvisor's stories, which often include reviews of hotels, eco-lodges, parks, and guided tours, are typically coherent longer documents that use regular grammar and have minimal noise (Arenal text and Corcovado text datasets). The titles of those reviews are very short documents; however, they are still coherent, with proper grammar (Arenal titles and Corcovado titles datasets). Practical relevance for the inclusion of datasets with review titles is justified by prior studies: e.g., Yang et al. (2020) examined the effect of perceived consistency between review text and title on review helpfulness; another example is using the title and the text in fake review detection (Banerjee, 2022; He et al., 2022). The Arenal and Corcovado text and titles datasets are a match in terms of coherence and length of the documents; however, they differ in size (with the Arenal data being larger) because of the mismatched

Table 1
Three dimensions of datasets under analysis. Bold font indicates the dimensions presenting major analysis challenges.

Data set	Subset	Dimensions		
		Text coherence	Document length (words)	Corpus size (documents)
Li Ziqi	YouTube	**Noisy**	**Short** (mean 17.2; median 10)	**Small** (N = 444)
	Weibo	**Noisy**	**Short** (mean 9.7; median 6)	Large (N = 7171)
Corcovado	Titles	Coherent	**Short** (mean 4.9; median 4)	**Small** (N = 618)
	Text	Coherent	**Long** (mean 130.6; median 78)	**Small** (N = 618)
Arenal	Titles	Coherent	**Short** (mean 4.1; median 3)	Large (N = 13,582)
	Text	Coherent	Long (mean 79.1; median 60)	Large (N = 13,582)

Corcovado and Arenal data are obtained from TripAdvisor.

popularity of the two destinations (Table 1).

The noisy, poorly structured data type was represented by two datasets collected from YouTube and Weibo platforms. These datasets consisted of comments to videos published by a prominent Chinese blogger Li Ziqi. Both datasets exemplified noisy documents of short length left by Chinese audiences on both platforms. Understandably, the size of the Weibo dataset was much larger than the size of the YouTube dataset, which provided a desired contrast. The comments were translated from traditional and simplified Chinese to English; a random set of 100 translated comments was validated by a bilingual expert who confirmed the correct translation of 99 comments. The characteristics of the datasets are summarized in Table 1. Note that the table does not include long noisy documents, as we are unaware of such data in the tourism context.

To facilitate understanding of results, a short description of the context behind the data is necessary. Arenal Volcano National Park is one of the primary tourist destinations in Costa Rica with 1.5 million visitors annually. Aside from the active volcano, the notable points of interest include waterfalls, geothermal springs and spas, a large lake, animal sanctuaries, and premium accommodations. As opposed to Arenal Volcano's mass tourism destination, Corcovado National Park, situated on the Osa Peninsula in southwestern Costa Rica caps the daily number of visitors at 330. Corcovado is one of the global biodiversity hotspots, including rainforests, mangrove swamps, coastal habitats, and an astonishing array of iconic wildlife such as tapir, jaguar, and anteater. The park promotes adventure and ecotourism, providing a model for the sustainable coexistence of vibrant ecosystems and human enjoyment.

Li Ziqi (in tables, LZQ) holds the Guinness World Record for having the "most subscribers for a Chinese language channel on YouTube" (18 million) while being also extremely popular on Weibo (26 million subscribers). Effectively, Li Ziqi acts as a promoter of Chinese culture, especially among the Millennials (Matei, 2020). Ziqi is widely credited with de-facto stimulating rural tourism in China and opening traditional Chinese culture to potential foreign travelers; in addition, her content is a perfect example of virtual tourism, the phenomenon developed during COVID-19 (Jiao et al., 2022; Westcott & Wang, 2021). The Li Ziqi dataset represents fan's comments on her most popular videos featuring traditional Chinese crafts ("Using bamboo to make … furniture", 60 million views), food ("… Snacks for Spring Festival", 53 million views), calligraphy ("The Scholar's Four Jewels of China", 160 million views), and rural life ("The Life of Cotton", 81 million views). We collected all Chinese language comments on these videos from YouTube and Weibo. One can expect differences in discussion topics as the Weibo platform mainly represents Mainland China fans who use Simplified Chinese script while Chinese language users from other countries tend to react on YouTube and use Traditional Chinese script.

All data was scraped from TripAdvisor (tripadvisor.com), YouTube (youtube.com), and Weibo (Weibo.com) websites, respectively, using a custom Python code with Selenium library. The data covers the period 2018–2022 (Weibo and YouTube) and 2011–2021 (TripAdvisor).

4. The models

Custom Python code was developed for all three models. In essence, the code (1) read review texts; (2) transformed the texts into the format required by the respective method implementation (LDA, BERTopic, or GPT-4); and (3) extracted the topics using the software library for the respective model. Specifically, LDA was using the Latent Dirichlet Allocation method from the sklearn library. BERTopic used bertopic library, HDBSCAN clustering, sklearn CountVectorizer, and pretrained sentence transformer all-MiniLM-L6-v2, all available through GitHub online developer platform.[5] Finally, the GPT-4 model was accessed with a Python code using OpenAI's API interface.[6]

For LDA, the following parameters were used: document-topic density factor $\alpha = 0.1$; topic-word density factor $\beta = 0.001$; number of terms 400. The optimal number of topics was selected by first running the model for the number of topics ranging from 5 to 150 with a step of 5. After determining the interval with best-interpretable topics, the model was executed again for the number of topics within this interval with a step of 1 to determine the number of topic parameter returning the best-interpretable topics. The final number of topics was specific to the dataset. For BERTopic, we used the default settings for the majority of options to replicate the most likely scenario of its use. The n-gram range was set at 2, the vectorization model was CountVectorizer, the dimensionality reduction algorithm was UMAP, the clustering algorithm was HDBSCAN, the number of topics was found automatically, with the minimum topic size of 10 documents, and c-TF-IDF was used for topic representation.

While LDA and BERTopic topic extraction (the third step as described in the previous paragraph) was straightforward and did not require much effort beyond calling the respective function from a software library, GPT topic modeling was significantly more involved. Successful text analysis with GPT models depends on the researcher's ability to give clear, concrete, and highly specific instructions on the analysis goal, the steps to achieve this goal, and the desired output format, as well as data domain, examples, and other relevant information. An emerging discipline of prompt engineering (OpenAI, 2023) deals with structuring the input of a GPT model in a way that aids in obtaining desired results. An additional consideration was GPT's limitation on the length of input (at the moment of writing, approximately 80,000 words with the experimental gpt-4-1106-preview model and 20,000 with a "regular" gpt-4-32K model), which is smaller than our data size. Note that LDA and BERTopic do not have any specific input limit other than reasonable processing time.

To meet GPT limitations, we broke GPT data processing into three parts as follows.

Part 1. Topic extraction. The text is separated into blocks confirming GPT model limitations; the topics are extracted from each of the segments.

goal = "Find the most prominent topics in the following documents"
steps = "

1. *Break the list of documents onto separate documents using the '\n' symbol as a separator;*
2. *When a document contains a spelling error, correct the error;*
3. *When a document contains an emoticon, replace the emoticon with a corresponding word;*
4. *Find no more than 20 most prominent topics common for all documents"*

actAs = "a person trained in summarizing a text"
format = "a table with the topic index in the first column and the topic text in the second column"
prompt = "Forget all prior prompts.

Your goal is to {goal}, acting as {actAs}.
To achieve this, take a systematic approach by: {steps}.
Present your response in markdown format, following the structure: {format}.
The list of documents is as follows: {text}"[7]

Part 2. Topic merging. The extracted topics are merged.

goal = "Find the most prominent topics in the following list of topics"

[5] http://github.com.

[6] https://platform.openai.com/docs/introduction.

[7] {text} is the analyzed document. The prompt is identical in stages 1, 2, and 3.

steps = "

1. *Break the list of topics into separate topics using the '\n' symbol as a separator;*
2. *Find two most similar topics. Two topics are the most similar if the semantic difference between these two topics is the smallest one;*
3. *Join two most similar topics into one topic;*
4. *Repeat steps 2, 3, and 4 until there are no more than 20 common topics"*

actAs = "a clustering algorithm"
format = "a table with the topic index in the first column and the topic text in the second column"

Part 3. Mapping documents to the extracted topics.
goal = "match the documents to the list of topics. The output should contain all documents, their matching topics, and matching scores"
steps = "

1. *Break the list of documents onto separate documents using the '\n' symbol as a separator;*
2. *Break the list of topics into separate topics using the '\n' symbol as a separator;*
3. *Each document starts with the document index followed by the document text;*
4. *For each document, starting with document 1, do the following steps:*
5. *Find three best matching topics from the list of topics;*
6. *For each of the three matching topics compute the matching score;*
7. *When there are no well matching topics, assume the topic is 'Other' and the matching score 0."*

The overall GPT topic modeling framework is shown in Fig. 1. We included the Part 1 code in Appendix 1.

5. Results

5.1. Topic development

The results of LDA and BERTopic models is a list of keywords that define documents' topics, which require further interpretation based on the documents with high loading on these topics. Topic modeling with GPT allows us to skip this step. Table 2 shows an example of keywords representing the first topic for each of the corpora and models, together with a document loading on this topic. For the entire set of keywords and topics please refer to Appendix 1.

Occasionally, we found it impossible to intelligently interpret the keywords in terms of an overarching topic, even after consulting with documents loading on these keywords; in this case, we marked the topic as a "Mix". Further, the outliers, i.e. isolated or small clusters of documents in the embedding space were labeled as "Other". An example of such a document (Corcovado text corpus, BERTopic model) is a park review concentrated on human waste: "… Like the broken window theory of crime, all this plastic waste encourages littering amongst the tourists and I saw quite a few bottles that were clearly not washed in from the ocean". We did not find another review focusing on human garbage in the entire dataset. These categories are arbitrary and depend on model parameters, e.g. the number of outliers can be reduced (potentially, to none) by relaxing limitations on the minimum cluster size. Tables 3–8 show the final topics as well as the percentage of documents loading on clearly defined topics other than outliers.

5.2. Topic validation

To validate the derived topics, for each of the method/dataset combinations we randomly selected 100 documents together with the topics they were assigned (each document could have up to 3 topics assigned). For a fair comparison, we excluded all document-topic assignments for which the respective confidence indicator was below 50%. For LDA, that means we used only the most prominent topic of a document. BERTopic model provided one topic per document while GPT provided up to three topics per document. Together, 2298 topic assignments were manually validated (Table 9). The raters were provided with a table containing the following fields: (1) Title of the text (where relevant); (2) Text of a review or online comment; and (3) The assigned topic(s). The raters were then asked to fill in an additional field as "1" (topic does not reflect the title/text/comment) or "2" (topic does reflect the title/text/comment). This procedure formed the base for calculating the percentage of correctly identified topics.

Note that a significant part of the documents was not assigned any specific topic (BERTopic), or all topics assigned to the documents had assignment probabilities (LDA) below 0.5 or matching scores (GPT) below 50%. This part was quantified as the percentage of documents for which a topic was not identified. In one case of an extreme mismatch between the tool and the dataset (BERTopic applied to Corcovado reviews), 91% of the documents were excluded from the topic assignment, even though for any specific dataset the most appropriate tool was able to assign a topic to 62% (YouTube) to 99% (Corcovado and Arenal review texts) of documents. Because of that, the final overall tool performance was computed as the percentage of documents assigned a correct topic, which we calculated as a product of the percentage of documents with the assigned topic and the percentage of correct document assignments (Table 9). For the best methods fitting a specific dataset, this overall performance varied from circa 50% for noisy Yutube and Weibo data to over 80% for coherent TripAdvisor reviews.

To exclude a confirmation bias in topic validation, 300 topic assignments were validated by an additional rater who was not a part of this research. The Interrater agreement was 82%, ranging from 88% for the correctness of the assignment of the most prominent topic to 73% for the least prominent topic of a document. One example of a disagreement between the raters was about the correctness of an assignment of a message about a wildlife refuge visit to the topic "Wildlife watching". While one of the raters judged this assignment correct, the other argued that there was a more fitting topic "Ecotourism and conservation".

5.3. Topic comparison

In the Introduction, we defined five dimensions important for understanding topics contained in a collection of documents: effective topic extraction, thematic representation, scalability, robustness, and transparency. This section reviews the skill of topic modeling algorithms across these five dimensions.

5.3.1. Effective topic extraction

The performance of LDA and BERT varied across the datasets. Frequently, LDA and BERT have identified many similar topics. For example, for YouTube comments (Table 3), both engines successfully identified video content (Bamboo crafts, paper making), cultural aspects of the content (Popularization of Chinese culture), welcoming to Malaysia (to study bird's nest production), and various expressions of commenters' admiration for Li Ziqi. In the Weibo dataset, both LDA and BERT have similarly recognized many topics such as Chinese New Year best wishes, "cute" grandmother, asking Li Ziqi to wear gloves for garden work, etc. (Table 4). More rare topics were identified only by one model: e.g., the topic "Li Ziqi steals wool [to make a calligraphy brush]" on YouTube was recognized only by BERT. This should not be treated as an advantage of one model over the other as model sensitivity can be managed by changing its parameters. One important generalization, however, is that contrary to BERT, LDA tends to generate multiple semantically similar topics (Li Ziqi is beautiful, like a fairy,[8] omnipotent, powerful, and amazing).

Similar observations can be made on Corcovado review titles

[8] The actual reference is to a Celestial Maiden of Chinese mythology.

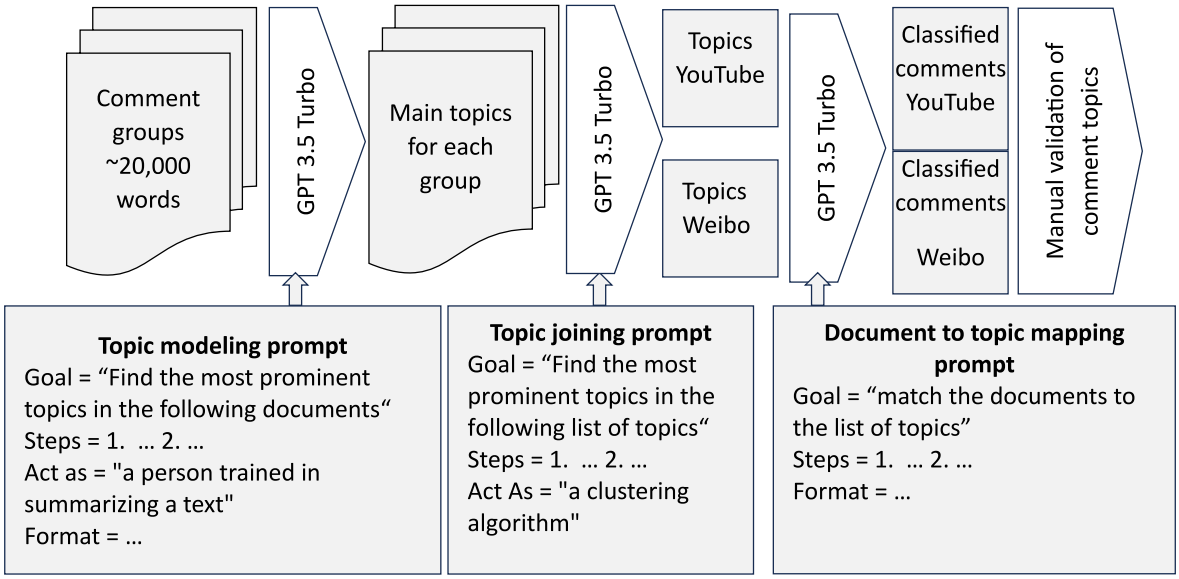


Fig. 1. GPT topic modeling framework.

(coherent but short documents, small corpus – Table 5). While at first glance LDA has produced more topics, many of those were nearly identical (e.g., amazing place, great experience, great hike, great nature), while BERTopic's results were more generic (c.f., "amazing place"). For a larger Arenal corpus (Table 6), however, BERTopic has clearly outperformed LDA by identifying multiple unique topics missing in LDA output: price, a need for tour guide, zipline experience, park challenges ("not for everyone"), etc. Interestingly, the topics extracted from review titles are similar to the ones generated from the full review text, supporting the practice of using only the titles to understand e.g. Instagram photography content.

Compared to BERTopic and LDA, GPT was able to extract meaningful and highly interpretable topics from all six datasets. Furthermore, the extracted topics were already interpreted in real-world terms and required little editing from the researcher's side. Overall, GPT has clearly overperformed both LDA and BERTopic in the efficiency of the topic extraction dimension.

5.3.2. *Thematic representation of document collection*

In addition to the identification of the topics, the extracted topics should correctly represent the document collection. That is, the majority of the documents should be clearly related to one or more of the topics. There was a clear difference between the performance of the algorithms here. The most striking is that only 9% of Corcovado reviews (and 20% of Corcovado titles) were related to one of the topics extracted by BERT (Tables 6 and 8). Similarly, LDA was clearly underperforming on noisy data from YouTube comments (Table 3), with only 30% of documents related to one of the identified topics. Overall, LDA and BERTopic performance varied across the datasets while GPT was able to identify topics of the majority (i.e., at least 60%) of the documents from all six datasets.

5.3.3. *Scalability*

Processing of the small-size dataset such as Corcovado text and title as well as YouTube (Table 7) was clearly failed by BERTopic, which identified only a few topics, while the rest of the topics were impossible to interpret. The underperformance of BERT on some data, also discussed in previous sections, directly relates to scalability issues. The major issue is the BERTopic assumption of only one topic per document,[9] which makes it less suitable for long reviews (Grootendorst, 2022), especially typical for the Corcovado text dataset. This problem is further complicated by the small number of reviews in the Corcovado data (Abuzayed & Al-Khalifa, 2021). BERTopic algorithm relies on cluster analysis, which is the most effective when applied to large datasets of short documents, as present in review titles and social media comments.

As opposed to BERTopic, the effective application of LDA requires multi-topic documents. For instance, Amazon Comprehend recommends at least 3-sentence long documents for its LDA-based topic extraction algorithm[10]. Even though the size of the Corcovado text dataset falls short of LDA requirements, it still was able to extract multiple meaningful topics related to park lodges, attractions, and visiting recommendations (Table 7). Similarly, LDA has outperformed BERTopic on Arenal review texts by extracting more specific reviews, even though BERTopic was able to partially compensate for the long document size by taking advantage of a large corpus (Table 8). Meanwhile, GPT was seemingly able to return meaningful and relevant topics for all scales.

5.3.4. *Robustness*

LDA was clearly outperformed by both BERTopic and GPT in processing small noisy YouTube dataset (Table 3): LDA identified few meaningful topics and only 30% of comments were mapped back to these topics. While processing a larger noisy Weibo dataset was seemingly successful, producing 18 meaningful topics, many of these topics had high similarities. Compared to that, BERTopic, while returning fewer topics, has also found important details missing in LDA processing, e.g. cute animals, chestnut picking, and discussion of fan ranking on Weibo. Both GPT and BERTopic were able to extract multiple meaningful and semantically different topics, despite the prevalence of slang in content.

[9] There are ways to circumvent this limitation, e.g., by breaking each document into smaller parts, deriving one topic for each part, and then assigning all derived topics to the original document.

[10] https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html.

Table 2
Example of topic development for all corpora and instruments. Only the largest topic and a random document loading on this topic are presented. "Mix": a topic hard to identify from the keywords. Similar shadings are used to emphasize related datasets.

		Corpus	Topic	Keywords (LDA, BERT) / representative document (LDA, BERT, GPT)
LDA	LZQ	YouTube	Mix	time, long, long time, know, delicious, film, life, shoot, special, make / "**Like a special program, Qishi is really filming her daily life, but now someone is helping her shoot the camera** … "
		Weibo	Mix	little, young, envious, fruit, tree, lady, young lady, fairy, little fairy, craftsman / "**What is the big yellow fruit**?"
	title	Corcovado	Great experience	great experience, experience, fantastic, amazing, amazing experience, wildlife experience, lifetime, great, Pedrillo, great walk, incredible wildlife / "**Simply fantastic**"
		Arenal	La Fortuna waterfall	waterfall, beautiful waterfall, expensive, beautiful place, place, swim, swim waterfall, serene, hike waterfall, waterfall hike / "**Serene**"
	text	Corcovado	Mix	sleep, light, Puerto, thing, highlight, heat, Jimenez, stop, baby, Puerto Jimenez / "**See my review … regarding taking a private plane from PJ to Estacion Biolgica Sirena. Our highlight was seeing a sleeping Tapir**"
		Arenal	Horse ride	Natural, horse ride, natural, awesome, ride, horse, attraction, horseback, beauty, love, feel, Fortuna waterfall / "**Natural Feel I got there.. and hope come again in next year**"
BERTopic	LZQ	YouTube	LZQ admiration	ziqi, li, good, bless, heart, hair, wanwan, amazing, world, super / "**I love you Li Ziqi**"
		Weibo	Good morning/night	good, fairy, morning, night, amazing, really, awesome, video, know, god / "**The video is so well done!**"
	title	Corcovado	Mix	park, national, wildlife, place, Rica, costa, hike, experience, Corcovado, jungle / "**Amazing experience**"
		Arenal	Waterfall hike	waterfall, hike, worth, beautiful, lake, Arenal, great, steps, nice, walk / "**Awesome waterfall**"
	text	Corcovado	Mix	park, guide, day, Sirena, station, Corcovado, saw, hike, tour, monkeys / ***the document is a 3,177 words trip report***.
		Arenal	Waterfall hike	waterfall, steps, water, worth, hike, swim, falls, beautiful, stairs, walk / "**Great experience, great photos involved walking down several stairs to get to the basin on the falls, able to dip in the water** …"
GPT	LZQ	YouTube	Appreciation of videos	**Ziqi's video moved me very much… Thank you for your video, it is really a beautiful culture, let us pass it on together.**
		Weibo	Admiration for skills	**… I think she can do anything except make atomic bombs…**
	title	Corcovado	Wildlife sightings	**Exquisite Wildlife**
		Arenal	Waterfalls	**Awesome waterfall**
	text	Corcovado	Hiking Trails	**… It's hard work walking during all day…**
		Arenal	Scenic views, nat. beauty	**The rain forest was amazing. The wild life, hot springs and landscape were like being in the Garden of Eden…**

5.3.5. Explainability

All three models are "black box", yet LDA's and, at a letter degree, BERTopic results are easier to understand. LDA's output is a set of topics, and for each document, it provides the distribution of topics and, for each topic, the distribution of words. These distributions are easy to interpret and can be presented as a list of words associated with each topic along with their probabilities. Similarly, the topics identified by BERTopic are based on the most significant words and phrases that differentiate one keyword cluster from another. This makes it relatively easier to explain the results in terms of the prevalent themes or topics present in the documents. These patterns of words that create topics are readily available for the researcher's analysis. GPT however provides only the final topics. In the words of ChatGPT, "GPT generates text based on the patterns it learned during training. While it can produce coherent and contextually relevant text, the challenge lies in understanding exactly how the model arrives at a particular output. The decision-making process is distributed across the many layers and parameters of the model, making it a "black box" in terms of interpretability." This explanation is hardly helpful in forming trust in model results. The transparency issue is discussed in more detail in the next section.

6. Discussion

Evidently, GPT is a winner on the first four dimensions important for topic identification in data relevant to tourism research (Fig. 2). It was able to return comprehensive, highly interpretable, and relevant to the real-world topics for all datasets, including the noisy ones, and at all scales. Further, most dataset documents (from 60% for the short and noisy dataset to nearly 100% for the cohesive Arenal review dataset) were related to one of these topics. This ensures that the algorithm not only identifies topics but also accurately represents the thematic content of the entire document collection. GPT, however, is an extreme case of a "black box" algorithm.

The black box problem (Castelvecchi, 2016; Rudin, 2019) refers to a

Table 3
Topic comparison: Noisy data, short documents, small corpus (Li Ziqi YouTube comments)[a].

LDA (30% identified)	BERTopic (62% identified)	GPT (61% identified)
Bamboo crafts	Bamboo crafts	Admiration for Li Ziqi's talents
Chinese culture	Come to Malaysia	Appreciation of Li Ziqi's videos
Come to Malaysia	Like to watch video	Aspiration to live a similar lifestyle
LZQ admiration	LZQ admiration	Beauty and tranquility of nature
LZQ is beautiful and talanted	LZQ happy life after suffering	Chinese New Year celebrations
Paper and ink making	LZQ steals wool	Connection to childhood memories
Treasure culture	LZQ works hard	Desire to learn and replicate skills
	Paper and ink making	Educational value of content
	Popcorn making	Emotional impact of videos
	Popularization of China culture	Global reach and influence of content
		Hard work behind scenes
		International viewership and admiration
		Love for Chinese food and snacks
		Pastoral and rural life
		Portrayal of traditional Chinese values
		Promotion of Chinese culture
		Sense of peace and relaxation from video
		Significance of traditional Chinese artifacts
		Simplicity and authenticity of content
		Traditional Chinese crafts and skills

[a] The (% identified) indicates a percentage of documents with successfully identified main topics as follows. LDA: other than "Unknown" and over 10% loading; BERTopic: not an outlier or a mix; GPT: not "Other".

Table 4
Topic comparison: Noisy data, short documents, large corpus (Li Ziqi Weibo comments)[a].

LDA (39% identified)	BERTopic (80% identified)	GPT (60% identified)
Bamboo crafts	Background music	Admiration for LZQ relationship with grandmother
Chinese New Year	Bamboo crafts	Admiration for skills and talent
Envy of the [rural] life	Chestnut picking	Affection for animals
Food	Cute animals	Beauty and craftsmanship in traditional Chinese arts
Good morning to LZQ	Food	Beauty of the environment
Good night to LZQ	Forwarding the video	Comments on video content and production
Happy new year	Good morning/ night	Cultural significance of Traditional Chinese culture
Like to watch video	Happy new year	Desire to learn or request for information
LZQ admiration	I am a fan	Desire to live a life similar to LZQ
LZQ grandma is cute	Iron Fan ranking	Discussion of Chinese cultural items
LZQ is beautiful	LZQ admiration	Discussions about traditional versus modern life
LZQ is like a fairy	LZQ grandma is cute	Emotional connection viewers feel with LZQ
LZQ is omnipotent	LZQ is omnipotent	Expressions of hunger or desire to eat
LZQ is powerful woman	Wear gloves for work	Expressions of support and encouragement
LZQ wear gloves for work	Paper and ink making	Interest in Chinese New Year traditions
Paper and ink making	Waiting for a new video	LZQ content is teaching about Chinese culture
Traditional Chinese culture		Making of traditional Chinese snacks and food
		Personal reflections and nostalgia
		Positive emotional responses (amazement, envy …)
		Preserving and inheriting traditional craft
		Richness of agriculture in LZQ environment
		Simplicity and tranquility of rural life

[a] The (% identified) indicates a percentage of documents with successfully identified main topics as follows. LDA: other than "Unknown" and over 10% loading; BERTopic: not an outlier or a mix; GPT: not "Other".

lack of trust in artificial intelligence (AI) algorithms. Brożek et al. (2023) deconstructed the black box problem into four issues: opacity (a limited understanding of the process by which the conclusions were derived), justification (providing a reason for a specific outcome), unpredictability (humans' aversion to "surprises") and the strangeness (humans aversion to "soulless" algorithm). The authors argue that contrary to common beliefs, the opacity problem is easy to deal with: even though the underlying concepts of an algorithm may be extremely complex, simplified explanations are readily available. In fact, we have provided these explanations in Section 2. Similarly, a new field of "Explainable AI" (Linardatos et al., 2020; Tjoa & Guan, 2020) is tackling the justification problem by developing special applications targeting explaining how a black-box solution was achieved. Several methods have been theorized to help in understanding GPT results, including GPT models trained on explanation data (Hassija et al., 2023). The justification problem is related to the first two: once a person can "understand" the way AI functions and how it has arrived at a particular solution, it becomes easy to justify accepting this solution.

We argue that the core of the GPT topic modeling black-box problem relates to the strangeness aspect. Indeed, LDA has already become a common instrument for tourism academics. LDA's output is a set of topics, and for each document, it provides the distribution of topics and, for each topic, the distribution of words. These distributions should further be interpreted by a researcher and can be presented as a list of words associated with each topic along with their probabilities. Similarly, BERTopic output represents clusters of words that require further interpretation. A human investigator is a key element in this process, using an AI not unlike more familiar tools (e.g., SPSS). GPT, however, seemingly excludes the human from the decision-making process, returning a final list of topics without turning to the researcher's expertise beyond the initial prompt. That makes GPT extremely "alien". Further challenging the notion of AI strangeness, GPT is not unknown of "hallucinating", that is, presenting incorrect information as if it were a fact. This is why careful validation of GPT-generated topics is required. This validation not only creates trust in GPT results but also eliminates the strangeness problem by returning the human researcher as a final decisionmaker.

Both LDA and BERTopic were successful in their own domains. LDA was extremely successful in extracting highly detailed topics from large datasets of long cohesive documents as found in users' reviews. Meanwhile, BERTopic was the most successful in processing short data as found in review titles. In addition, it was very robust against noisy data which did not require any specific treatment except filtering a standard list of stop words. BERTopic, however, had the worst scalability among the algorithms, being unable to deal with data outside its domain, especially with longer documents. To overcome the one topic per document restriction of BERTopic, one recommendation is to break large documents into smaller parts or even separate sentences. This can also help to solve the problem with a small dataset size. This is why we generally ranked BERTopic capabilities above LDA's.

One limitation of both BERTopic and, especially, LDA is the high specificity of generated topics. For example, LDA processing of Weibo

Table 5
Topic comparison: Coherent data, short documents, small corpus (Corcovado titles)[a].

LDA (86% identified)	BERTopic (20% identified)	GPT (96% identified)
Amazing adventure	Amazing place	Accessibility and Transportation
Amazing place	Corcovado national park	Accommodations and Facilities
Beautiful hike	Must see	Adventure and Exploration
Beautiful place	Nature love	Biodiversity and Conservation
Corcovado guide	Plan ahead	Comparisons to Other Parks and Destinations
Corcovado national park	Ranger stations	Conservation Efforts and Environmentalism
Costa Rica	Guide recommendation	Cultural and Educational Value
Great experience	Stay overnight	Difficulty and Challenge of Trails
Great hike	Wildlife	Disappointments and Unmet Expectations
Great nature	Worth money and effort	Experiences with Specific Guides or Tours
Jungle walk		Guide Quality and Importance
La Sirena ranger station		Hiking and Trekking Experiences
Nature love		Natural Beauty and Scenery
Rain forest		Park Management and Regulations
Rainforest hike		Rainforest Ecosystem and Flora
Tour guide		Recommendations and Tips for Visitors
Wildlife		Remote and Isolated Location
Wonderful park		Unique and Unforgettable Experiences
		Weather Conditions and Preparedness
		Wildlife and Animal Sightings

[a] The (% identified) indicates a percentage of documents with successfully identified main topics as follows. LDA: other than "Unknown" and over 10% loading; BERTopic: not an outlier or a mix; GPT: not "Other".

comments has resulted in multiple topics related to Li Ziqi's character: "Li Ziqi admiration", " Li Ziqi is beautiful", "Li Ziqi is like a fairy", " Li Ziqi is omnipotent", and "Li Ziqi is a powerful woman". Semantically, the similarity between these topics is high. In comparison, while GPT has returned more topics, these topics are, on one hand, more general (e.g., all abovementioned topics are combined into two: "Admiration for skills and talent" and "Positive emotional responses". Similarly, separate topics related to crafts (bamboo furniture, paper making, etc.), prominent in LDA and BERTopic outcomes, were combined by GPT into one general topic of "Preserving and inheriting traditional craft".

Overall, GPT is a universal instrument for topic modeling, summarizing document data using easily comprehensible formulations, robust against data noise (slang and typos), and requiring little pre-processing. It is also highly scalable, being effective in all domains of document and corpus sizes. However, it is too early to discard other methods successfully used in tourism literature. We suggest that the main issue with GPT-type generative language models is trust. While our research has shown excellent correspondence between the topics and related documents, the strangeness of GPT models is extremely challenging. Further, the guidelines for ethical ways of using large language models (LLMs) are largely unexplored. For instance, the amazing performance of ChatGPT relies on model training on one petabyte of data mostly derived from web crawling, that is, the data inadvertently "donated" by the public. How the responsibility for LLM outcomes should be assigned should the results prove incorrect or damaging? Should the LLM use be reported to the public and does it somehow undermine research outcomes? In our opinion, in all these respects LLMs are not different from any other analytical tool, but this area is severely underinvestigated.

Table 10 presents an overall conclusion on three contemporary topic modeling approaches. BERTopic-type models are useful for topic

Table 6
Topic comparison: Coherent data, short documents, large corpus (Arenal titles)[a].

LDA (64% identified)	BERTopic (82% identified)	GPT (81% identified)
500 steps waterfall hike	Asis Proyecto tour	Accessibility and transport
Arenal Volcano	Baldi Hot Springs Resort	Accommodation and camping
Baldi Hot Springs Resort	Beautiful, amazing	Adventure sports and adrenaline
Gorgeous place	Exceeding/ disappointing	Conservation and sustainability
Great experience	Experience, education	Crowds and peak times
Great trip	Gorgeous place	Customer service and staff interactions
Great view	Hanging bridge	Family-friendly activities
La Fortuna waterfall	Hidden gem	Food and dining experiences
Lake Arenal	Hot springs	Guided tours and educational value
Nice hike	La Fortuna waterfall	Hiking trails and difficulty
Tour	Must see	Hot springs relaxation
Waterfall climb, swim	Need guide	Local culture and history
	Not for everyone	Park facilities and cleanliness
	Overpriced/worth it	Park fees and value for money
	Relaxation	Safety and park regulations
	Specific tour guide	Scenic views and photography
	Waterfall hike	Swimming and water activities
	Wildlife	Waterfalls and natural beauty
	Zipline	Weather conditions and preparation
		Wildlife sightings

[a] The (% identified) indicates a percentage of documents with successfully identified main topics as follows. LDA: other than "Unknown" and over 10% loading; BERTopic: not an outlier or a mix; GPT: not "Other".

Table 7
Topic comparison: Coherent data, long documents, small corpus (Corcovado reviews)[a].

LDA (80% identified)	BERTopic (9% identified)	GPT (99% identified)
Clothing to wear	Nature love	Adventure and exploration
Crossing river	Rainforest	Beaches and coastal areas
Flora and fauna	Ranger stations	Educational aspects
Jimenez lodge	Wildlife	Flora and plant life
La Leona trail	Worth the trip	Photography opportunities
La Sirena ranger station		Accessibility and travel options
Lodge		Accommodations and camping options
Manuel Antonio park		Best month/season to visit, weather and climate
Nice path		Biodiversity, conservation, eco-friendly practices
Pedrillo ranger station		Comparison with other national parks
Puerto Jiménez [airport]		Costs and fees
Puma Valley trek		Difficulty and physical challenge
Specific guide recommendation		Guides and tours
Waterfall		Hiking trails
Wildlife		Park facilities and safety
		Park regulations and restrictions
		Planning and reservations
		Remote and untouched wilderness
		Scenic views and natural beauty
		Visitor experience and expectations
		Visitor tips and recommendations
		Wildlife watching

[a] The (% identified) indicates a percentage of documents with successfully identified main topics as follows. LDA: other than "Unknown" and over 10% loading; BERTopic: not an outlier or a mix; GPT: not "Other".

Table 8
Topic comparison: Coherent data, long documents, large corpus (Arenal reviews)[a].

LDA (97%/100% identified)	BERTopic (93% identified)	GPT (99% identified)
Arenal volcano trail	Asis Proyecto tour	Accessibility, travel options, and parking
Arrive early, crowd	Baldi Hot Springs Resort	Accommodations and camping options
Baldi Hot Springs Resort	Boat ride	Adventure, zip-lining, horse riding, rafting
Clothing to wear	Cerro Chato hike	Crowds and best times to visit
Costa Rica	Hanging bridge	Ecotourism and conservation
Food	Hot springs	Entrance fees and value
Gorgeous place	La Fortuna waterfall	Family-friendliness and activities for children
Hanging bridge	Lake Arenal kayaking	Food and dining options
Highly recommend	Main POI	Guided tours and educational value
Horse ride	Traveling from Liberia	Hiking trails
Hotel stay	Traveling from Monteverde	Hot springs
Traveling from Monteverde	Waterfall hike	Local culture, communities, and town visits
Make sure to have shoes	Wildlife tour	Overall satisfaction and recommendations
Proyecto Asis tour		Park facilities
Rain forest		Safety and security
Rest stop		Scenic views, natural beauty, and photography
River swimming		Souvenirs, shops and local crafts
Time to spend		Waterfalls and swimming
Topic name		Weather and climate
Travel to park		Wildlife watching
Waterfall climb, swim		
Wildlife		

[a] The (% identified) indicates a percentage of documents with successfully identified main topics as follows. LDA: other than "Unknown" and over 10% loading; BERTopic: not an outlier or a mix; GPT: not "Other".

extraction from short documents, e.g. to evaluate the main interests of social media influencers' followers (Kirilenko et al., 2023). It can also be used for long documents such as found in travel blogs, but only after breaking the documents into shorter parts, ideally into sentences. Using LDA is more advisable in this context. Finally, when comprehensive, highly interpretable results of topic modeling are desirable, GPT is an instrument of choice. Compared to earlier methods such as LDA, GPT results can be further improved by feeding validation results back into the model as examples of valid or invalid document classifications (a technique known as one-shot or few-shot prompting). This procedure, however, would require an extra validation step increasing human effort requirements.

We suggest that these conclusions are extendable to a highly related task of natural language processing, sentiment analysis. LDA has successfully been used for sentiment analysis, including the tourism domain (Putri & Kusumaningrum, 2017). Similarly, BERT-based models were successfully applied to tourism data (Viñán-Ludeña & de Campos, 2022). While we are not aware of GPT applications in tourism or hospitality literature, in other domains it has proven to deliver a superior performance (Kheiri & Karimi, 2023).

Another area for further improvement of analytic methods available for academic researchers is the specialization of instruments. One example of such specialization is a BERT-type model trained on tourism domain data by Arefieva and Egger (2022). We tested their model on our data and found its performance similar to BERTopic, yet the results of a specialized model would be preferable in terms of trust. Similarly, there are multiple versions of LDA focusing on alleviating its weak points such as a fixed number of topics. For GPT-type models, one can expect fast progress including fine-tuning the models on tourism data, automated generation of instructions, post-processing, and many others. For example, the results of GPT v. 4 have been vastly improved compared with GPT v.3.5. Meanwhile, an ongoing online discussion of the latest (at the time of the second revision of this paper) GPT v. 4o has alleged some reduction in results quality as compared to the original GPT v.4. This observation is supported by reports on extremely high costs of running the model ($0.7 million/day), driving the developers to seek lower-cost solutions, either by developing new computer chips or by streamlining the code (Mok, 2023). Hence, we envision two competing tendencies: (1) Improving model performance and (2) reducing costs of

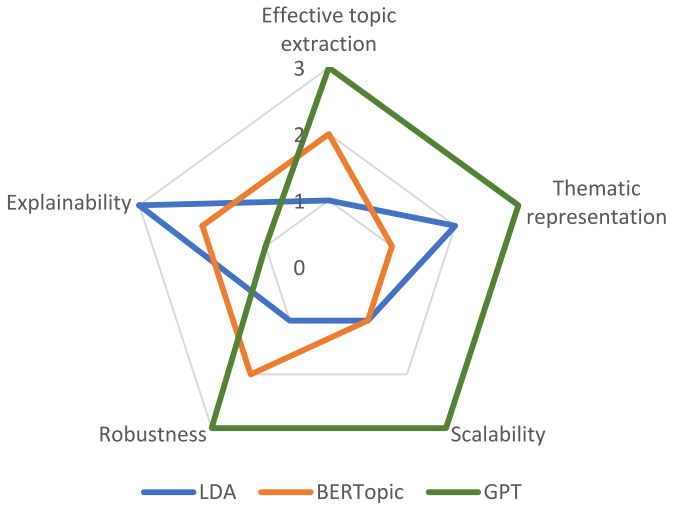


Fig. 2. Ranking of three analyzed topic modeling approaches from 1 (the lowest) to 3 (the best).

Table 9
Topic validation results. Effective topic extraction (% documents for which a topic was identified), identification precision (% correctly identified topics), and the overall % of documents with correctly identified topics for six datasets/three methods combinations.

Dataset	% topic identified			% correctly identified			Overall quality		
	LDA	BERT	GPT	LDA	BERT	GPT	LDA	BERT	GPT
YouTube	30%	62%	61%	59%	80%	86%	0.18	0.50	0.52
Weibo	39%	80%	60%	56%	60%	82%	0.22	0.48	0.49
Corc. title	86%	20%	96%	67%	75%	89%	0.58	0.15	0.85
Corc. review	80%	9%	99%	54%	90%	84%	0.43	0.08	0.83
Aren. title	64%	82%	81%	67%	54%	96%	0.43	0.44	0.78
Aren. review	97%	93%	99%	63%	89%	87%	0.61	0.83	0.86

Table 10
Instrument applicability domains.

Data category	Dataset size	LDA	BERTopic	GPT[a]
Online comments, tweets	Small (≪1000)	–	+	+
Online comments, tweets	Large (≫1000)	+/–	+	+
Short reviews, titles (1–2 sentences)	Small (≪1000)	+/–	–	+
Short reviews, titles (1–2 sentences)	Large (≫1000)	+/–	+	+
Typical reviews (3+ sentences)	Small (≪1000)	+	–	+
Typical reviews (3+ sentences)	Large (≫1000)	+	+/–	+

[a] GPT classification requires validation.

running the models.

A recent review of social media short text analysis by Laureate et al. (2023) has revealed that automated topic modeling is well adopted in academic research, with the highest percentage of studies (42%) published in social sciences. Meanwhile, the authors' sample of 189 publications comprised of a systematic Ebsco® and Web of Science® search has resulted in a single publication in leading tourism journals (by Kirilenko et al. (2021) in Tourism Management)). Even though the analyzed sample comprised only a small part of 1284 initially collected publications, reduced in a rigorous screening process, e.g., applying quality criteria, study outcomes suggest that tourism academic research is behind many other areas of social sciences in the analysis of large volumes of social media. Further, we suggest that the majority of published studies concentrate on a very small sector of social media, such as textual data published in TripAdvisor. Contrasting, Laureate et al. (2023)counted 14 social media platforms used in academic research. Notably, there are many platforms missing in tourism research that represent specific countries and cultures. These platforms are as diverse as CoachSurfing (inexpensive travel oriented at free stays with locals), FlyerTalk (a community of frequent flyers), Blued (a Chinese gay platform), and Vkontakte (Russian users). We envision one future development in tourism research in the diversification of platforms and types of research media to include images, videos, and sounds. This expansion, however, is restricted by methodological limitations.

Despite the explosive development of new numerical methods of analysis and innovative research frameworks, the percolation of these methods in the tourism literature is extremely slow, especially in comparison with the adoption rate of these methods in the tourism and hospitality industry. In our opinion, there is a lack of academic literature systematically and comprehensively testing these methods on typical tourism and hospitality data used in research. Importantly, our paper presents only a limited first attempt at such testing. Even though the approaches that we utilized represent (in our opinion) the most important and perspective developments in text analysis, each approach includes a variety of implementations, and each implementation can be fine-tuned using multiple parameters. The approaches and parameters that we have used are not necessarily the "optimal" ones; moreover, we strongly suspect that the choice and tuning of the analysis tools depend on data characteristics, which are still to be discovered. Once this bottleneck is cleared, the new, easier to use tools based on ChatGPT-like AI assistance will provide an easy-to-use suite of contemporary methods allowing analysis of diverse multimedia data.

7. Conclusion

New GPT-type large language models deliver huge progress in the analysis of tourism-related data, but it is not time to discard the proven methods of topic analysis. The main challenge is the issue of trust, mainly related to GPT-type black-box models' "strangeness". In our opinion, this strangeness comes from the seeming elimination of a researcher from the decision-making process, unlike with other topic modeling approaches. Because of that, GPT analysis, while superior in terms of the effectiveness of topic extraction, thematic representation of document collection, scalability, and robustness against data noise, should be used only in conjunction with a thorough validation of its outcomes. We provide recommendations on application domains for three types of models: a generative probabilistic model (LDA), a model combining embedding to capture the semantic relationship between document's words with clustering (BERTopic), and a Generative Pre-trained Transformer model (GPT), which is focused on the generation of intelligible texts. These recommendations can be useful for both academics, exploring social media data and large surveys to advance theoretical knowledge, and for practitioners, seeking to provide trustworthy data analysis for their stakeholders fast and efficiently.

Impact statement

With ChatGPT disruption, Artificial Intelligence tools such as ChatGPT have already been used in various fields, from genetics to liberal arts. In tourism research, however, the progress has been slow; we are unaware of any application truly digging into the potential of large language models (LLM) as a research tool. Our manuscript explores one frequently used research task, which is the extraction of topics from tourism reviews. The manuscript provides a comprehensive cross-examination of the two most significant LLMs, BERTopic and GPT (the model behind ChatGPT). We introduce a typology of contrasting tourism review data and compare LLM performance with currently the most frequently used instrument for automated topic analysis, Latent Dirichlet Allocation (LDA).

In terms of practical significance, we provide recommendations on application domains for three types of models: a generative probabilistic model (LDA), a model combining embedding to capture the semantic relationship between the document's words with clustering (BERTopic), and a Generative Pre-trained Transformer model (GPT), which is focused on the generation of intelligible texts. We also discuss using prompt engineering to optimize GPT performance. These recommendations can be useful for both academics, exploring social media data and large surveys to advance theoretical knowledge, and for practitioners, seeking to provide trustworthy, fast, and efficient data analysis for their stakeholders.

Regarding the theoretical contribution, we discuss the issue of the slow percolation of new AI tools in tourism literature. In our view, the main challenge is the issue of trust in the black-box models' "strangeness". This strangeness comes from the seeming elimination of a researcher from the decision-making process, unlike with other topic modeling approaches. GPT analysis, while superior in terms of the effectiveness of topic extraction, thematic representation of document collection, scalability, and robustness against data noise, seemingly eliminates a researcher from the data analysis process and as such is an extreme case of "strangeness". We discuss ways of reducing this misconception.

CRediT authorship contribution statement

Andrei P. Kirilenko: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Svetlana Stepchenkova:** Writing – review & editing, Writing – original draft, Validation, Conceptualization.

Declaration of competing.interest

None.

Acknowledgment

The authors are very thankful to the University of Florida student Hannah Oh for her participation in the data validation procedure. We are also thankful to Jing Yang, M.S. for the validation of the translation from Chinese.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tourman.2024.105007.

References

Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: An experimental study on BERTopic technique. *Proc. Comput. Sci., 189*, 191–194.
Agarwal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How much noise is too much: A study in automatic text classification. In *Presented at the seventh IEEE International Conference on data mining (ICDM 2007)* (pp. 3–12). IEEE.
Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Front . Artif. Intell., 3*, 42.
Arefeva, V., & Egger, R. (2022). When BERT started traveling: TourBERT—a natural language processing model for the travel industry. *Digital, 2*, 546–559.
Arefieva, V., & Egger, R. (2022). TourBERT: A pretrained language model for the tourism industry. *arXiv preprint arXiv:2201.07449*.
Banerjee, S. (2022). Exaggeration in fake vs. authentic online reviews for luxury and budget hotels. *International Journal of Information Management, 62*, Article 102416.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77–84.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
Borko, H. (1961). Automatic document classifications using a mathematically derived classification system. *System Development Corp. FN-6164*. Santa Monica, CA.
Borko, H. (1962). The construction of an empirically based mathematically derived classification system. *Presented at the Proceedings of the May 1-3, 1962, spring joint computer conference* (pp. 279–289).
Borko, H., & Bernick, M. (1963). Automatic document classification. *Journal of the ACM, 10*, 151–162.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
Brożek, B., Furman, M., Jakubiec, M., & Kucharzyk, B. (2023). The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artif Intell Law*. https://doi.org/10.1007/s10506-023-09356-9
Carrillo, J. C., Beltran, V., Sebastia, L., & Onaindia, E. (2023). *SmartTur+ ECO: A conversational recommender system for tourism*.
Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News, 538*, 20.
Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing Systems* (pp. 288–296).
Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys, 54*, 1–35.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*.
DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics, 41*, 570–606.
Egger, R. (2024). Vectorize me! A proposed machine learning approach for segmenting the multi-optional tourist. *Journal of Travel Research, 63*(5), 1043–1069.
Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in instagram data: A topic modelling comparison. *Tourism Review, 77*, 1234–1246.
Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology, 7*, Article 886498.
Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management, 59*, 467–483.
Harway, N. I., & Iker, H. P. (1964). Computer analysis of content in psychotherapy. *Psychological Reports, 14*, 720–722.
Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Comput*, 1–30.
He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science, 41*, 896–921.
Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.
Huang, F., Kwak, H., & An, J. (2023). Is chatgpt better than human annotators? Potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv: 2302.07736*.
Iker, H. P., & Harway, N. I. (1965). A computer approach towards the analysis of content. *Behavioral Science, 10*, 173–182.
Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*, 15169–15211.
Jia, S. S. (2020). Motivation and satisfaction of Chinese and us tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management, 78*, Article 104071.
Jiao, Y., Meng, M. Z., & Zhang, Y. (2022). Constructing a virtual destination: Li ziqi's Chinese rural idyll on YouTube. *Journal of Teaching in Travel & Tourism, 22*, 279–294. https://doi.org/10.1080/15313220.2022.2096178
Kheiri, K., & Karimi, H. (2023). *Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning*. arXiv preprint arXiv:2307.10234.
Kirilenko, A., Emin, K., & Tavares, K. C. (2023). Instagram travel influencers coping with COVID-19 travel disruption. *Information Technology & Tourism*, 1–28.
Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021). Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management, 83*, Article 104241. https://doi.org/10.1016/j.tourman.2020.104241
Koseoglu, M. A., Yick, M. Y. Y., King, B., & Arici, H. E. (2022). Relational bibliometrics for hospitality and tourism research: A best practice guide. *Journal of Hospitality and Tourism Management, 52*, 316–330.
Kuzman, T., Mozetic, I., & Ljubešic, N. (2023). *Chatgpt: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification*. ArXiv, abs/2303.03953.
Laureate, C. D. P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artif Intell Rev, 56*, 14223–14255. https://doi.org/10.1007/s10462-023-10471-x
Lerksuthirat, T., Srisuma, S., Ongphiphadhanakul, B., & Kueanjinda, P. (2023). Sentiment and topic modeling analysis on twitter reveals concerns over cannabis-containing food after cannabis legalization in Thailand. *Healthcare Info. Res, 29*, 269–279.
Li, X., Wang, Y., Zhang, A., Li, C., Chi, J., & Ouyang, J. (2018). Filtering out the noise in short text topic modeling. *Information Sciences, 456*, 83–96.
Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy, 23*, 18.
Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management, 24*, 119–154.
Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development, 1*, 309–317.
Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*, 159–165.
Luo, Y., He, J., Mou, Y., Wang, J., & Liu, T. (2021). Exploring China's 5A global geoparks through online tourism reviews: A mining model based on machine learning approach. *Tourism Management Perspectives, 37*, Article 100769.
Ma, S., & Kirilenko, A. P. (2020). Climate change and tourism in English-language newspaper publications. *Journal of Travel Research, 59*, 352–366.
Matei, A. (2020). Country life: The young female farmer who is now a top influencer in China. The Guardian. Jan. 28 2020. URL: https://www.theguardian.com/lifeandstyle/2020/jan/28/li-ziqi-china-influencer-rural-life 7.25.24.
Miles, J., & Selvin, H. C. (1966). A factor analysis of the vocabulary of poetry in the seventeenth century. *The computer and literary style*.
Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Presented at the Proceedings of the 2011 conference on empirical methods in natural language processing*.
Mok, A. (2023). ChatGPT could cost over $700,000 per day to operate. *Microsoft is reportedly trying to make it cheaper* [WWW Document]. Business Insider. URL https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4, 7.11.24.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, Article aac4716.
OpenAI. (2023). *Prompt engineering*.
Osgood, C. E., & Walker, E. G. (1959). Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal and Social Psychology, 59*, 58.
Putri, I. R., & Kusumaningrum, R. (2017). Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. *Presented at the journal of physics: Conference series*. IOP Publishing, Article 012073.
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*, 9.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.
Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychology and Marketing, 39*, 441–459.
Shrestha, K. M., Wood, K., Goodman, D., & Mistica, M. (2023). Do we need subject matter experts? A case study of measuring up GPT-4 against scholars in topic evaluation. *Presented at the proceedings of the seventh workshop on natural language for artificial intelligence (NL4AI 2023) co-located with 22th international conference of the Italian association for artificial intelligence (AI* IA 2023)*.
Thompson, L., & Mimno, D. (2020). *Topic modeling with contextualized word representation clusters*. arXiv preprint arXiv:2010.12626.
Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems, 32*, 4793–4813.

Tran, N. K., Zerr, S., Bischoff, K., Niederée, C., & Krestel, R. (2013). Topic cropping: Leveraging latent topics for the analysis of small corpora. In *Research and advanced echnology for digital libraries: International Conference on theory and practice of digital libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3* (pp. 297–308). Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*, Article 101582.

Viñán-Ludeña, M. S., & de Campos, L. M. (2022). Discovering a tourism destination with social media data: BERT-based sentiment analysis. *J. Hospit. Tourism. Technol., 13*, 907–921.

Westcott, B., & Wang, S. (2021). China's rural tourism boom [WWW Document]. CNN. URL https://www.cnn.com/travel/article/china-rural-tourism-pandemic-cmb-intl-hnk/index.html, 7.12.24.

Williams, N. L., Wassler, P., & Fedeli, G. (2023). Social representations of war tourism: A case of Ukraine. *Journal of Travel Research, 62*, 926–932.

Yang, S., Yao, J., & Qazi, A. (2020). Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. *Information Processing & Management, 57*, Article 102179.

Zhang, B., Ding, D., & Jing, L. (2022). *How would stance detection techniques evolve after the launch of chatgpt? arXiv preprint arXiv:2212.14548*.

Zhang, H., Liu, R., & Egger, R. (2023). Unlocking uniqueness: Analyzing online reviews of Airbnb experiences using BERT-based models. *J. Trav. Res.00472875231197381*.

Andrei Kirilenko, Ph.D. is an Associate Professor at the Department of Tourism, Hospitality & Event Management at the University of Florida. His research is focusing on interaction between humans and environment with concentration on the impacts of climate change and sustainability issues. He is especially interested in the research of social and mass media and big data analysis.

Svetlana Stepchenkova, Ph.D. is a Professor at the Department of Tourism, Hospitality and Event Management at the University of Florida. Her research interests are in marketing communications, branding, and positive image building. She studies tourism behavior and the effectiveness of destination promotion in situations of strained bilateral relations between nations. She is also interested in usability of user-generated content for managerial decision making in destination management.