https://doi.org/10.31410/tmt.2023-2024.291

A Comprehensive Analysis of Online Reviews in the Srem Region through Topic Modeling

Olivera Grljević[1]
Mirjana Marić[2]

Received: October 16, 2023
Revised: June 27, 2024
Accepted: July 10, 2024

Keywords:
Topic modeling;
Online reviews;
Tourist preferences

Abstract: *This chapter employs topic modeling to reveal the public stance and perception of tourist destinations in the Srem region, providing insights into their diverse appeal and variety of tourist profiles. Social media is the touch point with visitors of tourist attractions and consumers of tourist services. The collection of online reviews of tourist attractions in the Srem region is populated and used for modeling the hidden thematic structures. The authors identified an optimal model with 14 distinct topics through extensive experimentation, centered around nature, relaxation, shrines, and museum history. The topics indicate various tourist profiles active, gastronomic, leisure-seeking, history-loving, and family-oriented tourists. Knowing about the audience is valuable for targeted marketing strategies. The authors extracted and analyzed subsets of reviews related to Monasteries, Museums, Nature, and Nature Reserves indicating specific preferences of tourists within these categories, such as historical relevance of museums, use of modern technologies in exhibitions, or children-friendly content in nature reserves, and improvement areas, such as condition of roads, control of forest cutting, or garbage disposal management. This research offers clearly defined methodological steps and valuable insights for marketing and tourism development in the Srem region.*

1. INTRODUCTION

Sifting through and finding relevant content became challenging due to the information overload caused by the abundance of online data. Efficient methods for extracting valuable information are developed. This chapter focuses on textual data, as it makes up approximately 80% of the Internet's content (Anandarajan et al., 2019; Dixon, 2023), and two such methods: keyword extraction and topic modeling. Keyword extraction is relevant because keywords depict the most significant information within text documents. Topic modeling is a technique that utilizes keywords to uncover hidden thematic structures, known as topics, within a collection of texts (referred to as a corpus) (Maier et al., 2018). Beyond merely discovering topics within documents, topic modeling aims to reveal how these topics are interconnected and how they evolve. It also proves valuable for tracking trends, emotions, rumors, and the driving forces that influence people's consumption of services or products, and as such it *enables rational economic decisions, learning about clients, and turning data into knowledge* (Ubiparipovic et al., 2020). Topic modeling seamlessly integrates both qualitative and quantitative analytical aspects (Maier et al., 2018) and, as it reveals knowledge about the tourist experience and provides insights often missed or neglected by marketers (Li et al., 2018; Shafqat & Byun, 2020), it is well-suited for exploratory and descriptive analyses (Banks et al., 2018).

Consumers of travel-related services often voice their opinions and attitudes through social media, such as online reviews, comments, discussion boards, or blog posts, making the application

[1] Univeristy of Novi Sad, Faculty of Economics in Subotica, Segedinski put 9-11, 24000 Subotica, Serbia
[2] Univeristy of Novi Sad, Faculty of Economics in Subotica, Segedinski put 9-11, 24000 Subotica, Serbia

of topic modeling particularly relevant for the tourism sector. By analyzing and categorizing these texts into topics tourism managers can understand traveler preferences, emerging trends, and the various aspects of tourist experiences. In the age of data-driven decision-making, topic modeling has become a necessity for the tourism sector that offers a competitive edge.

Authors of the research presented in this chapter use topic modeling to gain insight into public opinion about various tourist attractions, uncover tourists' preferences, and appeal to different localities in the Srem region. By Srem region, authors refer to the Srem in Serbia and the Srijem region in Croatia. This chapter presents the methodological framework for topic modeling and the main results of conducted research. The methodological framework includes the following: 1) data collection, 2) data pre-processing, 3) selection and application of topic modeling algorithm, 4) evaluation of resulting topics, and 5) reporting on the results. For the data collection, the authors used social media as the touch point with visitors of tourist attractions and consumers of tourist services of the Srem region. Therefore, the corpus comprises Google online reviews of tourist attractions in the Srem region. The data pre-processing refers to data cleansing, standardization, and representation in a vectorized form suitable for topic modeling algorithms. The proposed methodological framework suggests the application of topic modeling over well-pre-processed data, which will allow for extracting hidden thematic structures in the corpus, identification of the most popular topics visitors mention, and exploration of content and vocabulary used within particular topics. The research results point to 14 distinct topics depicting general public opinion on the Srem region's tourism potential. Salient keywords are used for detailed exploration of these topics. The results uncovered insights on tourist profiles useful for targeted marketing campaigns. The analysis was furthered with topic modeling per categories of tourism destinations: Monasteries, Museums, Nature, and Nature reserves, which has uncovered specific tourist preferences elaborated in more detail in the last section of the chapter. The business benefits of topic modeling are twofold. Insights gained from topic modeling can improve marketing campaigns and the tourism offer of the region itself, while the results of topic modeling can be used to further expand data analysis, in terms of data annotation and sentiment analysis.

The chapter is structured as follows. Section *2. Related work* provides an overview of the literature on topic modeling in the tourism sector. Section *3. Methodology* gives an overview of the research methodology. The following subsections are detailed elaborations of each methodological step, which are illustrated through the case study of the Srem region: *3.1. Data Collection and Cleansing* provides details about the collection of texts, the formation of the corpus, and the reduction of noise in the text; *3.2. Data Pre-processing* is devoted to the issues of data quality increase, normalization, and transformation of data into a format suitable for selected analytical methods; *3.3. Modeling the Hidden Thematic Structures: Method Selection* is an overview of available methods for topic modeling and the decision-making process for method selection, while section *3.3.1. The Latent Dirichlet Allocation* provides details on the selected analytical method, i.e., the approach to topic modeling. Evaluation of topic models is an essential step to finding the optimal parameters for a selected method. Details on the evaluation process and available metrics are presented in section *3.4. Evaluation and Selection of Topic Model*, while experiments, resulting models, evaluation of models, and discussion on results are presented in section *4. Results and Discussion*. Concluding remarks on the conducted research are presented in section *5. Conclusion*.

2. RELATED WORK

Topic modeling is an analytical approach for discovering hidden topics in document collections. It enables automated summarization of information and a better understanding of large amounts of data. Topic modeling algorithms rely on the assumption that documents can be described and expressed by a limited number of fundamental concepts, the so-called topics, which are differently expressed in each document in the corpus (Egger, 2022). Learning about fundamental concepts is performed without supervision, based on the mutual occurrence of keywords in similar texts, does not require pre-labeled data, and can be categorized as a technique of unsupervised learning about texts (Grljević, 2023). The result of topic modeling is *k* number of topics. Each topic comprises a collection of highly correlated keywords.

The literature on topic modeling in the tourism sector is limited. As a potential reason, the authors Papilloud and Hinneburg (2018) cite the gap in knowledge and skills needed to apply topic modeling in tourism-related research, such as programming, statistics, and the basics of mathematical modeling. Research on topic modeling applications in tourism over the past five years can be categorized into four main areas:

1) Research focusing on understanding tourists' perceptions (Kim et al., 2019; Wen et al., 2020; Yu & Egger, 2021; Zou, 2020).
2) Research centered on analyzing tourists' preferences (Hu et al., 2019; Kim et al., 2019; H. Lee & Kang, 2021; Shafqat & Byun, 2020; Vu et al., 2019).
3) Research investigating tourists' emotions and sentiments (Ali et al., 2022; Calheirosa et al., 2017; Shafqat & Byun, 2020; Yu & Egger, 2021).
4) Research focused on extracting insights to enhance management practices in the tourism industry.

Although these research areas are general and reflect the primary research focus, the research itself cannot be classified into only one of these areas. Contributions or practical implications have a direct connection with managerial activities in respective fields, such as research presented in papers (Yu & Egger, 2021) and (Kim et al., 2019) contribute to destination management and marketing, the research presented in the paper (Wen et al., 2020) contribute to public land management and policy, while the author of the paper (Zou, 2020) provides actionable insights for the restaurant industry to better accommodate niche customers. Aside from a main focus, the research often addresses aspects from other research areas, such as analysis of preferences or perceptions is often coupled with sentiment analysis, such as research presented in the papers (Ali et al., 2022; Calheirosa et al., 2017; Shafqat & Byun, 2020; Yu & Egger, 2021).

Authors researching ***tourists' perceptions*** highlight the importance of analyzing user-generated content to understand evolving tourist perceptions and experiences. The focus is mainly on the way tourists regard, understand, or interpret various aspects of tourism or travel-related services and experiences. Topic modeling in this context allows automated discovery of insights about tourists' attitudes from publicly available texts, which would otherwise remain as an unused potential and source of business knowledge. Such insights can refer to:

a) *Shift of interest* (increase or decline) *in tourist offers* through time. Authors Kim et al. (2019) explored shifting tourist perceptions at Jeju Island UNESCO heritage sites, offering practical insights for destination marketing and management in Jeju Island. Kim et al. (2019) noticed that the popularity of "Sunrise Peak" in Jeju Island declined. Over time,

visitors became less satisfied with simply visiting heritage sites and began seeking more adventurous experiences in these areas.

b) Identification of particular *sentiments associated with tourists' experience*. Yu and Egger (2021) investigated over-tourism through tourist experiences at crowded attractions, providing practical implications for destination management organizations. Authors highlighted that tourists addressing the over-tourism feel the most negative towards safety and security (Yu & Egger, 2021).

c) Identification of key *factors influencing customer satisfaction*. Authors Wen et al. (2020) explored a niche market of tourists with food allergies and utilized topic modeling to analyze their dining experiences. Their practical contributions are recommendations to the restaurant industry to accommodate these niche tourists. Wen et al. (2020) emphasized good communication between personnel and allergy customers, as well as clearly stated allergen information that mostly contributes to customer satisfaction.

d) Identification of *factors influencing public acceptance or rejection of changes in business practice or policies*. Zou (2020) used topic modeling to understand public perception and acceptance of national park fee increases, offering theoretical and practical contributions to public land management. The author analyzed public acceptance of fee increases at the National parks in the USA and identified distributive justice principles and place attachment as the main influential aspects (Zou, 2020).

Research in topic modeling dealing with ***tourist preferences*** focuses on identifying commonalities and differences among tourist groups in terms of preferences (Vu et al., 2019) and behavior or activities (H. Lee & Kang, 2021). For these purposes, topic modeling is often complemented by sentiment analysis (Shafqat & Byun, 2020). Preferences can also be observed in terms of customer dissatisfaction. Authors of papers (Hu et al., 2019; Kirilenko et al., 2021) suggest directing topic modeling on reviews with negative semantics, contrary to using the overall corpus when some cues of dissatisfaction go unobserved. Topic modeling in service of tourist preferences analysis can:

a) Contribute to *travel product development and tour recommendations*. Vu et al. (2019) analyzed travel itineraries and discovered implicit activity preferences, identified common itinerary types, and revealed popular tourist activities.

b) Enhance *tourism management* by monitoring tourists' preferences, behaviors, and activities and monitoring trend changes. H. Lee and Kang (2021) classified Flicker users into residents and tourists, categorized tourism activities using topic modeling, and identified appealing factors for different regions of attraction, intending to provide timely insights to tourism management organizations.

c) Aid to *customer satisfaction improvement* by offering insights valuable for effective customer dissatisfaction management in tourism and travel-related businesses. Hu et al. (2019) aimed to uncover the root causes of customer dissatisfaction in the hotel industry by identifying topics within hotel reviews, classifying them, and exploring how dissatisfaction varies across hotel categories (defined by star ratings). They identified ten negative topics, predominantly related to service, facilities, and pricing, which provide valuable insights into customer dissatisfaction. Authors Kirilenko et al. (2021) highlighted the challenges in analyzing diverse reviews and argue that practitioners in the tourism and hospitality industry seeking to understand and effectively address customer satisfaction and dissatisfaction should more carefully approach the interpretation of topic models developed over negative reviews. Their suggestion is to deploy manual verifications and dictionary-based approaches to improve interpretability.

Analysis of tourists' preferences and perceptions is often complemented by sentiment analysis (Shafqat & Byun, 2020; Yu & Egger, 2021), as already mentioned. Sentiment analysis is dealing with ***tourists' emotions and sentiments***. According to authors Pang and Lee (2008), sentiment suggests a stable opinion that reflects one's feelings, while emotions represent positive (pleasant, relaxing) or negative (nervous, angry) human feelings (Grljević, 2016). Dealing with emotions in the context of tourism and travel-related services is important as people form emotions before any information processing and human behavior is strongly influenced by them (Hudson et al., 2015). In addition, revealing general topics within a corpus is often not informative enough for business decision-making. Sentiment analysis helps identify leading sentiment across topics. This approach was used in various types of analysis:

a) Authors Yu and Egger (2021) explored over-tourism and identified that tourists' feelings are generally positive, with the highest sentiment scores associated with aspects unrelated to crowding, such as art and culture. Safety, security, service quality, and queues received the lowest sentiment scores. However, topics related to social interactions and the overall atmosphere were rated more positively.
b) Authors Shafqat and Byun (2020) researched possibilities of enhancing recommendations of under-emphasized tourist spots, which should be endorsed more effectively and interestingly to attract more tourists.
c) As a means of improving customer experience, authors Calheirosa et al. (2017) used topic modeling to identify topics characteristic of eco (green) hotels and applied sentiment analysis, which unveiled that improvements should be directed towards food services and hotel amenities, while authors Ali et al. (2022) used topic modeling to identify topics related to tourism attractions and sentiment analysis helped to model in detail the satisfaction and dissatisfaction of tourists according to 4 identified general topics (atmosphere on public areas, shopping experience, citizen's behavior, and overall touristic experience).

Extracting knowledge to ***enhance managerial activities***. Regardless of the research focus, the results of topic modeling directly contribute to the improvement of managerial practices. They offer actionable recommendations that can be applied to promotional efforts (Kim et al., 2019; Vu et al., 2019), the management of tourism destinations (H. Lee & Kang, 2021; Shafqat & Byun, 2020; Yu & Egger, 2021), the development of training or educational modules for personnel (Wen et al., 2020), or the enhancement of services and tourist offers (Ali et al., 2022; Calheirosa et al., 2017; Hu et al., 2019).

The literature review reveals several shortcomings in the research methodologies of the papers examined. Most papers lacked detailed explanations of their methodological steps, including issues such as inadequate data preparation (Calheirosa et al., 2017; Kim et al., 2019) and the absence of evaluations of resulting topic models (Calheirosa et al., 2017; Hu et al., 2019; Kim et al., 2019; Shafqat & Byun, 2020; Yu & Egger, 2021). Furthermore, essential information such as the data source and language of analyzed texts was frequently missing (Calheirosa et al., 2017; Kim et al., 2019). Data preparation was inconsistent across studies, with all the papers implementing some form of word reduction, but only a few conducting specific data preparations to enhance data quality (Ali et al., 2022; Kirilenko et al., 2021; Wen et al., 2020). The predominant method for topic modeling in tourism-related research is probabilistic LDA modeling, often using online reviews as primary textual data. However, no analysis of corpus characteristics or document length was provided. Corpus sizes varied significantly, ranging from a small of 400 reviews to a large of 147,000, and all research, which provided language information, used texts written in English. This chapter aims to address these methodological gaps by emphasizing data

cleansing, pre-processing, and model evaluation to enhance data quality and reproducibility. Additionally, it distinguishes itself by applying topic modeling to Serbian language texts, which due to limited language resources and tools is an under-researched language.

3. METHODOLOGY

Although the Srem region has a wealth of tourist attractions, they are not equally popular or visited by tourists. In an attempt to understand the varying appeal of destinations and to enhance the promotion of under-emphasized tourist spots, the goals of this research paper are to uncover hidden thematic structures of tourist destinations and attractions in the Srem region, to identify the most popular topics visitors mention, and to define various tourist profiles useful for targeted marketing activities by exploring the content and vocabulary used within particular topics. Social media is used as a touch-point with visitors, particularly online reviews in which visitors freely share opinions and write about the overall experience. The research methodology comprises five steps: data collection and cleansing, data pre-processing, topic modeling, evaluation of topic models and optimal model selection, and interpretation of topics and discussion of the resulting topic model.

Data collection and cleansing implies activities related to the development of a dataset, i.e., corpus. A more detailed description of data collection and corpus development methodology is presented in the paper (Grljević et al., 2019), while this chapter provides a general overview. The corpus is a collection of online reviews of tourist destinations and attractions in the Srem region. Online reviews were acquired from Google review sites. As user-generated content, online reviews are often written in informal and colloquial writing style without respect to grammar and require initial data cleansing to remove or correct such irregularities from the content (Kovačević et al., 2020). Data cleansing also implies irrelevant content removal, such as empty pages and reviews containing only interrogative or factual sentences that do not reflect subjective viewpoints.

Data pre-processing, in combination with data cleansing, has the task of raising the quality of the data set, as well as transforming the data into a format suitable for selected machine learning algorithms. Pre-processing steps performed on the collected data include: 1.) Tokenization – splitting streams of texts into meaningful elements, such as words, which are referred to as tokens. 2.) Removal of stop words. Stop words refer to frequent but uninformative words. 3.) Stemming – reduction of words to their word stem by removing suffixes from the end of the words (Grljević, 2023). Vectorization is the next step that involves the transformation of pre-processed online reviews into a numerical format that topic modeling algorithms understand, i.e., it refers to a vector representation of data.

Modeling topic structures implies the selection, application and fine-tuning of a topic modeling algorithm. In the specific research, the authors applied LDA topic modeling. In order to identify the optimal number of topics, it is necessary to develop a larger number of topic models for different sizes of k and to *evaluate the resulting topic models*. In the specific research, the authors applied a coherence measure and determined that the optimal size of k is 14 on the overall corpus. The following sections present each methodological step in more detail.

3.1. Data Collection And Cleansing

Online reviews reflect the opinions, attitudes, preferences, and tourists' dissatisfactions. Hence, Google reviewing sites are the data source for the research presented in this chapter. The authors crawled Google review site of each of the following predefined destinations or attractions

in the Srem region: Fruska Gora, Obedska Bara, Borkovac Lake, Zasavica, each monastery at Fruska Gora, Orthodox church Indjija, the church of St. Demetrius, Archaeological site Sirmium, Imperial palace Sirmium, Sava Sumanovic memorial house, Sava Sumanovic art gallery, Art museum Sid, Museum of beekeeping, Museum of Srem, Eltz, Ilok Castel, Vinkovci museum, and Vucedol. These destinations are grouped into four major categories: Monasteries, Museums, Nature, and Nature reserves, as illustrated in Table 1. All available online reviews were collected up to 2019 when data collection took place. To improve the quality of the data, the collected data was cleaned as follows:

1. People write online reviews without respect for grammar or spelling. Each collected online review was checked for spelling and grammar errors and was corrected accordingly using Aspell libraries and an Aspell plug-in called DspellCheck.
2. Interrogative sentences and sentences containing historical facts do not express the subjective viewpoints of individuals. Generally, data analysts do not consider them in the analysis. However, in rare cases, people use these types of sentences to express certain emotions or use rhetorical questions to express displeasure. For this reason, a linguist[3] examined each interrogative and factual sentence to determine whether they contained subjectivity and made a decision about their retainment in the dataset.
3. Data cleansing implied the removal of excessive whitespaces and special characters, which are not part of the written language. For this purpose, the authors used Python.
4. Tourist from all around the world are welcome to write online reviews in their native language. The authors used a custom-written Python translator to translate reviews written in languages other than Serbian to Serbian and convert Cyrillic to Latin alphabet.
5. Diacritical marks (*č*, *ć*, *š*, *ž*, and *đ*) pose a limitation to data pre-processing or analytical tasks. In all collected reviews, diacritical marks were replaced with equivalents (*cx*, *cy*, *sx*, *zx*, and *dx*). For this purpose, the authors used Python and regular expressions.

The final data collection, corpus, comprises 1908 documents – online reviews. Table 1 provides information on the number of reviews per category. We can observe that the corpus is imbalanced in favor of nature-related online reviews, which might pose a problem during the modeling phase. This issue is addressed in the section *4. Results and Discussion*.

Table 1. Overview of destinations per category and the number of collected reviews

Destination category	Subcategory	Number of reviews
Monastery	Besenovo, Divsa, Grgeteg, Jazak, Krusedol, Novo Hopovo, Orthodox church Indjija, Petkovica, Privina glava, Sisatovac, St. Demetrius, Staro Hopovo, Velika Remeta, and Vrdnik.	326
Museum	Beekeeping Museum, Sava Sumanovic Gallery, Sid Museum, Sirmium, Srem Museum, Vinkovci Museum, Vucedol, Eltz, and Ilok Castel.	309
Nature	Fruska Gora	942
Nature reserve	Borkovac lake, Zasavica, and Obedska bara.	331

Source: Own research

3.2. Data Pre-processing

The goal of data pre-processing is to improve data quality and transform data in a format suitable for analytical and machine learning models (Grljević, 2023). In any analytical project, data pre-processing is a crucial task that consumes the most time and effort of the analyst. Handling

[3] Gordana Svilengaćin, Master in Serbian philology–linguistics

unstructured data like online reviews presents a unique challenge, mainly due to the extensive vocabulary used. Even short documents can contain numerous words, phrases, and symbols with context-dependent meanings. Each unique word in the corpus becomes a dimension or attribute used to describe documents, leading to high dimensionality. Therefore, effective data preparation is key to reducing dimensionality and identifying a simplified set of attributes (Feldman & Sanger, 2007).

A corpus of online reviews, collected and cleansed as presented in the previous section, is the input for the pre-processing. All documents (online reviews) undergo identical pre-processing procedures. As the research goal is to identify prevailing topics based on the prominent keywords, the focus has been on the most granular level: individual words. Therefore, each document is tokenized (split) into its constituent words (tokens). To reduce the dimensionality of the data, the authors implemented the following steps of data pre-processing:

1. Lowercasing. The capitalized, uppercased, or lowercase versions of the word are treated as different unique tokens (e.g., *Great*, *GREAT*, or *great*) and, as such, contribute to the increase in the dimensionality of the data. To reduce variations in the vocabulary and lower the corpus dimensionality, it is necessary to standardize the case. The authors opted to lowercase the corpus content.
2. Removal of punctuation. The punctuation does not add new knowledge or information relating to topics and it was removed from the content before further analysis.
3. Alphanumeric words. Words containing both numbers and letters, i.e., alphanumeric content were removed from online reviews.
4. Filtering of stopwords. Stopwords refer to frequent words in spoken and written language that do not add new knowledge or attribute to the semantics of the text, such as conjunctions (Grljević et al., 2022). Every word in the corpus documents is examined to ascertain whether it corresponds to a stopword, and if so, these stopwords are subsequently eliminated from the documents.
5. Word stemming. Stemming is the process of reducing words to their base. It is achieved by removing the suffix from the word (e.g., *begins* or *beginning* is stemmed to its' root *begin*). Given the corpus comprises texts written in the Serbian language, the authors of this chapter used a custom-built stemmer for the Serbian language (Milosevic, 2012).
6. Short words. Words containing two or less than two characters were removed from the analysis after stemming. These words can be considered as stopwords.

Table 2. illustrates examples of online reviews after pre-processing. The second column contains the original online review and the third column contains pre-processed and stemmed content that was used as input for topic modeling.

Table 2. Illustration of preprocessed online reviews

No.	Comment	Pre-processed
1	It will look nice. (srb. *Lepo će izgledati*)	lep izgled
2	Very quiet place. Very nice interior and very old building. (srb. *Vrlo mirno mesto. Vrlo lep enterijer i vrlo stara zgrada*)	vrl mir mest vrl lep enerijer vrl star zgrad
3	Soul paradise. (srb. *Raj za dušu*)	raj dusx

Source: Own research

To apply topic modeling algorithms pre-processed texts should be vectorized (Grljević et al., 2022). Vectorization relies on vector space model representation, which implies the representation of each document *d* as a vector. In the research presented in this chapter, *d* refers to one

online review pre-processed as aforementioned. For this purpose, the authors chose a well-known approach to text representation: the bag-of-words (BoW), which has proved to be a well-suited approach to text representation in similar tasks (Liu, 2012). The BoW model uses the so-called document-term matrix (DTM). The matrix rows represent pre-processed documents (*m*). The matrix columns represent unique words, tokens (*n*), derived from the corpus. The matrix values can refer to frequencies of occurrence in documents or weighted values (Grljević et al., 2022). The authors of this chapter used weighted values. In particular, the Term Frequency – Inverse Document Frequency (TF-IDF) score is used to weight vector components according to their contribution to the document content (Salton & McGill, 1986). More information on TF-IDF measures is available in (Grljević, 2023). The resulting matrix is the input for the topic modeling in the consequent step.

3.3. Modeling the Hidden Thematic Structures: Method Selection

When selecting a suitable topic modeling method, one should consider the characteristics of the corpus data and the possibilities and limitations of available methods. In terms of the corpus characteristics, it is essential to observe the volume of available data, the data types the corpus comprises, and the length of each document. Generally, the need for a large amount of data and significant adjustments to achieve acceptable results limits topic modeling (Liu, 2012). In the research presented in this chapter, the authors use the corpus comprised of 1908 online reviews. Due to the sparsity of data, one can anticipate that topic modeling will: a) easily recognize general and frequent topics in the corpus, b) have difficulty recognizing locally frequent but globally rare aspects, which are often the most useful in, e.g., applications of sentiment analysis since they are most relevant to the specific entity that the user is interested in (Liu, 2012). As the goal of the research presented in the paper is to identify the general topics that tourists talk about, topic modeling will successfully meet the research needs and contribute to solving the research goal.

Topic modeling methods do not show uniform results in working with all types of texts (newspaper articles, scientific papers, blogs, tweets, online reviews, queries, comments, or e-books), (Tang et al., 2014). The main problems are reflected in insufficient data, too short texts that lack context, or, in the case of e-books, the representation of many topics. Custom data preparation can solve these problems (Tang et al., 2014), such as dividing the e-book into pages, where each page becomes a document, or aggregating tweets with the same attribute into longer documents, where the aggregation criterion can be the author or other meta-feature. In the research presented in this chapter, the corpus comprises online reviews. In general, online reviews tend to be short. The authors analyzed the corpus after pre-processing to assess the suitability of the most common approaches to the topic modeling. Figure 1. illustrates the distribution of online reviews according to the number of words. The majority of online reviews have less than 10 words. Analysis showed that, on average, reviews have 6.48 words. Based on these statistics, the authors conclude that online reviews are rather short and they should strive towards a topic modeling approach suitable for dealing with short texts.

The authors Abdelrazek et al. (2023) grouped available approaches to topic modeling into four categories: algebraic, fuzzy, probabilistic, and neural. Each group exhibits general limitations and advantages, while within the same category, various algorithms use different baseline mathematical models to shape topics. An assessment of the performance of topic modeling approaches on short texts is relevant given the corpus used in the research presented in this chapter comprises short online reviews.

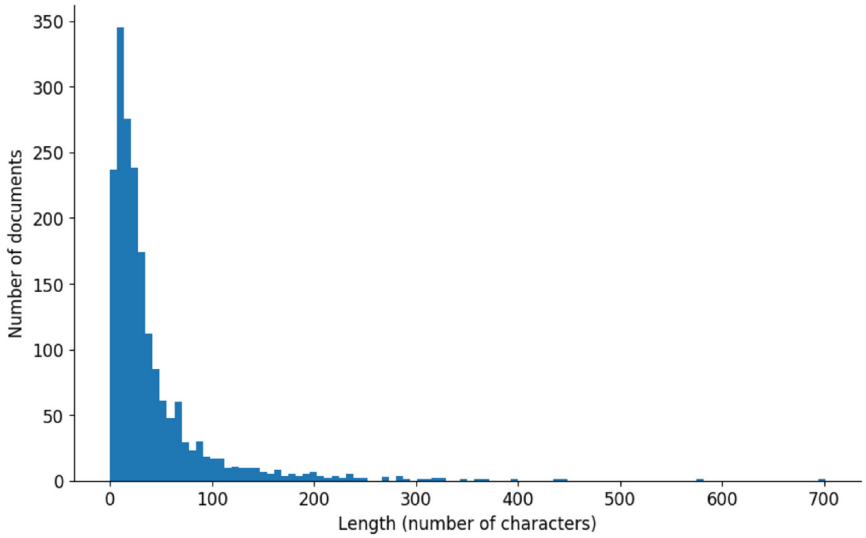


Figure 1. Distribution of reviews according to the total number of words
Source: Own research

Algebraic topic modeling methods, such as Latent Semantic Analysis (LSA) or Non-negative matrix factorization (NMF), are based on linear algebra. These methods are simple, intuitive, and relatively efficient (Abdelrazek et al., 2023). They perform DTM decomposition using various mathematical methods and thus represent dimensionality reduction techniques. The LSA method decomposes the DTM matrix by applying singular value decomposition (SVD), which determines unique terms that represent the underlying concepts manifested within the data (Hutchison et al., 2018). SVD reduces sparse matrices and enables the preservation of similar structures among matrix columns by representing each separate document (Hutchison et al., 2018). LSA is typically used as a dimension-reduction or noise-reducing technique. As such, it is not suitable for the research presented in this chapter. The NMF method is specialized for dealing with non-negative values, such as the values of DTM matrices that represent the frequencies of words in the document (Paatero & Tapper, 1994), (D. D. Lee & Seung, 1999). NMF decomposition implies decomposition of the original *n* tokens by *k* topics *(W)* and *k* topics by the original *m* documents *(H)* (Egger, 2022), which results in two non-negative matrices *W* and *H* that are consequently multiplied (dot product), (Papilloud & Hinneburg, 2018). The application of the NMF method for topic modeling has gained popularity due to its capability to automatically collect rare and essential features from a set of non-negative data vectors (Aghdam, 2022). However, the authors of the research presented in this chapter did not choose this approach due to its data greediness and sensitivity to selected parameters, which requires time-consuming and computationally expensive hyperparameter tuning.

Fuzzy topic modeling methods are based on clustering techniques. Forming clusters begins with extracting topics from the documents, following the assignment of words to them (Abdelrazek et al., 2023). Authors Abdelrazek et al. (2023) have emphasized that the advantage of this approach lies in their capability to overcome the sparsity problem, which makes them effective in modeling topics within short text, such as tweets or online reviews. Medical research extensively uses fuzzy topic modeling, while there are no resources nor benchmark models for the tourism sector. Therefore, the authors of this chapter did not consider a fuzzy approach to topic modeling.

In general, *probabilistic topic modeling methods*, such as Latent Dirichlet Allocation (LDA), are simple, intuitive, extensible, and interpretable, while inference becomes complicated with

increased model complexity (Abdelrazek et al., 2023). LDA topic modeling determines recurring patterns based on the co-occurrence of terms in documents (Grljević, 2023). The LDA models have a particular specificity. All documents from the corpus share the same set of topics, while they express topics in different degrees - the probability distribution of topics per document differs. Each topic includes all words from the corpus vocabulary, while the probability distribution of keywords per topic differs. Keywords with the highest probabilities best reflect the observed topic. LDA is the prevailing topic modeling approach in research dealing with short texts (Laureate et al., 2023) and in research relating to topic modeling in the tourism sector, as the literature review of this chapter pointed. Authors Laureate et al. (2023) identified that 79.79% of research studies dealing with short text, encompassed in their systematic literature review, used LDA topic modeling. The LDA has limitations. Topics change over time. If the corpus spans several years, the order of the documents in the corpus becomes relevant as it depicts changes in topics through time. The LDA approach has proven to be unstable in those situations. Any change in the order of the fed documents during training can influence the resulting topics (Abdelrazek et al., 2023). In the research presented in this chapter, authors do not observe time dimension, and this characteristic of the LDA modeling approach does not pose a limitation for the research. Due to the prevalence of the LDA approach in tourism-related research that provides the possibility of benchmarking the results, the authors of this chapter decided to model hidden thematic structures in the corpus using the LDA topic modeling approach.

Neural topic modeling methods are based on deep learning techniques with variations in data representation, such as lda2vec (Moody, 2016) or text embeddings obtained using Bidirectional Encoder Representations from Transformers (BERT) pre-trained language model (Devlin et al., 2019). These methods are often referred to as a black box, as they lack interpretability of model parameters, and it often remains unknown why the model is performing or not performing well (Abdelrazek et al., 2023). Deep learning is generally data greedy and should not be a choice of approach when working with small-scale corpora, such as the corpus used in the research presented in this chapter.

3.3.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation, introduced by Blei et al. (2003), is a widely used method for uncovering underlying topics within a collection of documents. LDA identifies recurring patterns in documents by analyzing word co-occurrence (Grljević, 2023). The goal is to extract sets of keywords that often appear together in the corpus. The sets of keywords are referred to as topics. LDA operates on the premise that words appearing together are related to the same topic. Each document in the corpus is represented as a distribution of topics, and each topic is characterized by a distribution of keywords. LDA makes the following key assumptions, (Grljević, 2023):

1. All documents in the corpus share the same set of topics, but individual documents express topics to varying degrees, resulting in different topic probability distributions.
2. Each topic encompasses all words in the vocabulary, but the distribution of words across topics varies. High-probability words within a topic represent its essence.
3. Word order in documents is not considered when generating LDA topic models.
4. In the resulting LDA model, the extracted topics are independent.

LDA is an iterative algorithm that produces results through a series of steps in model training and fine-tuning phases. Figure 2 illustrates this iterative process. The input of the LDA model is a corpus comprised of M documents, denoted as $\{d_1,d_2,...,d_M\}$. Each document is a sequence of N words

$(w_1,w_2,...,w_N)$, where w_N is the nth word in the sequence. The unique words, *W*, from across the corpus collectively form the vocabulary. LDA uses words in vectorized form. Given the input, the primary objectives of LDA include 1) Determining *K*, the number of topics in the corpus, which is a predefined parameter set before executing LDA; 2) Estimating *Φ*, the topic-word distribution, which specifies the likelihood of each word appearing within a given topic; and 3) Estimating *θ*, the document-topic distribution, which specifies the probability of each topic's presence in a particular document. In Figure 2, *W* denotes the words in the documents, while *Z* signifies the topics associated with the document and words within the document. In the modeling phase, LDA aims to infer the hidden topic structure by estimating the values of *Φ*, *θ*, and *Z* given the observed words *W* in the documents. To accomplish this, LDA relies on two hyperparameters, alpha (α) and beta (β). These hyperparameters play pivotal roles in shaping the number of topics characterizing each document and the number of words describing these topics.

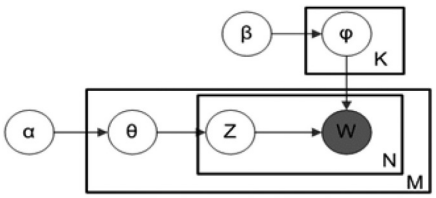


Figure 2. The iterative process of LDA modeling of hidden thematic structures
Source: Blei et al., 2003

As Figure 2 illustrates, LDA topic modeling begins with the initialization of parameters α and β. The parameter α is related to the distribution of topics across documents, the document-topic distribution. Documents are represented with a grater mixture of topics, the higher the α value (Grljević, 2023). The parameter β is related to the distribution of words by topics, the topic-word distribution. Topics will contain more vocabulary words, the higher the β value (Grljević, 2023). As α and β values affect the granularity and diversity of topics identified by LDA (Egger, 2022), selecting their appropriate values is crucial for the quality and meaningfulness of topic modeling results. The optimal values of these parameters are unknown in advance. Therefore, selecting their optimums involves experimentation and fine-tuning to achieve the desired level of topic granularity and diversity, aligning with the specific objectives of the analysis. Parameter values are varied during the experimentation (Grljević, 2023), and the quality of the resulting models is assessed. For this purpose, techniques for automated parameter adjustments are used, such as grid-search.

3.4. Evaluation and Selection of Topic Model

The purpose of model evaluation is to assess the generalizability of the resulting topic model and its independence from the data or specific case study (Wallach et al., 2009). It is advisable to combine statistical evaluation parameters and evaluation of model interpretability by a human expert. Human experts' role is to analyze leading keywords within the topics and to provide insights on the quality and meaningfulness of the topics and the possibility of their interpretation. Repeating keywords across topics is a signal that too many topics are selected.

Different statistical parameters can evaluate the quality of topic models, such as model stability, topic coverage, or coherence. *Model stability* refers to its sensitivity to small changes in the input data. The model is stable if it consistently concludes the presence of similar topics regardless of changes in the data (Greene et al., 2014). *Coverage* evaluates how well the resulting topic

model covers a predefined set of concepts, i.e., reference topics that the model is expected to reveal (Korenčić et al., 2021). The topics' *coherence* represents one of the key and most frequently used measures in evaluating the quality of topic models. The authors of the research presented in this chapter used the coherence in evaluation of the resulting topic models. A coherent model can be considered a model that includes all or most of the facts (Egger, 2022). In other words, topic coherence measures the frequency of co-occurrence of leading words in a topic. For a given list of words $T = \{w_1, \ldots, w_n\}$, coherence is defined as (Egger, 2022; Rosner et al., 2013):

$$C(T) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{p(w_m, w_l) + \frac{1}{D}}{p(w_l)} \quad (1)$$

The probability $p(w_m, w_l)$ is the quotient of the number of documents containing both words w_m и w_l and the total number of documents in the corpus D. A smoothing factor of $1/D$ is introduced to avoid zero-based calculations. The probability $p(w_l)$ is the quotient of the number of documents containing word w_l and the total number of documents in the corpus D. To generate topics that can be easily interpreted, in the modeling process several models are built with different parameters settings (Maier et al., 2018) and the model with the highest coherence score is selected. Most often, the number of topics that ends the accelerated growth of topic coherence provides meaningful and interpretable topics, while a slightly larger number can be chosen for more detailed subtopics.

4. RESULTS AND DISCUSSION

The search for the optimal value of K, which provides the best fit for the corpus and meaningful interpretation of topics, requires extensive experimentation. In the research presented in this chapter, authors experimented with different values of $\alpha = [0.01, 0.31, 0.61, 0.91]$ and $\beta = [0.01, 0.31, 0.61, 0.91]$ hyperparameters and considered symmetric and asymmetric distributions as there is no prior knowledge about topic distributions. The symmetric distribution indicates an even distribution of topics throughout the document, while an asymmetric distribution favors certain topics. For each combination of parameter values, authors built topic models for K ranging from 2 to 15 and evaluated the resulting topic models using coherence as the goodness-of-fit measure. Different percentages of the corpus are used to validate persistence in the results: 25%, 50%, 75%, and total corpus. Using Python and grid-search for hyperparameter tuning, authors built 1560 models in total with varying parameters and validation sets. The model with the optimal K has the highest coherence score. The results indicate that the highest coherence is 0.717 for the model with 14 different topics, $\alpha = 0.61$, and $\beta = 0.91$.

According to the results, 14 topics reflect tourists' overall perception of the Srem region. The most salient keywords in all 14 topics, presented in Figure 3, indicate that people mostly talk about nature, rest, and walks (srb. *priroda*, *odmor*, *šetnja*) in national park Fruška Gora, following shrines in monasteries (srb. *svetinje*), and the history in museums (srb. *istorija*). In Figure 3. these keywords are marked by red rectangles. Other prevailing words are adjectives and adverbs indicating pleasant feeling and satisfaction related to the visits, such as beautiful (srb. *lepo, prelepo*), wonderful (srb. *divno, predivno*), excellent (srb. *odlično*), super (srb. *super*), interesting (srb. *zanimljivo*), the most beautiful (srb. *najlepše*), very (srb. *jako*), phenomenal (srb. *fenomenalno*), good (srb. *dobro*), or enjoyment (srb. *uživanje*, which is noun). From the prominent keywords, the authors conclude that visitors to the Srem region had a positive experience, and they recommend these destinations as a must-see.

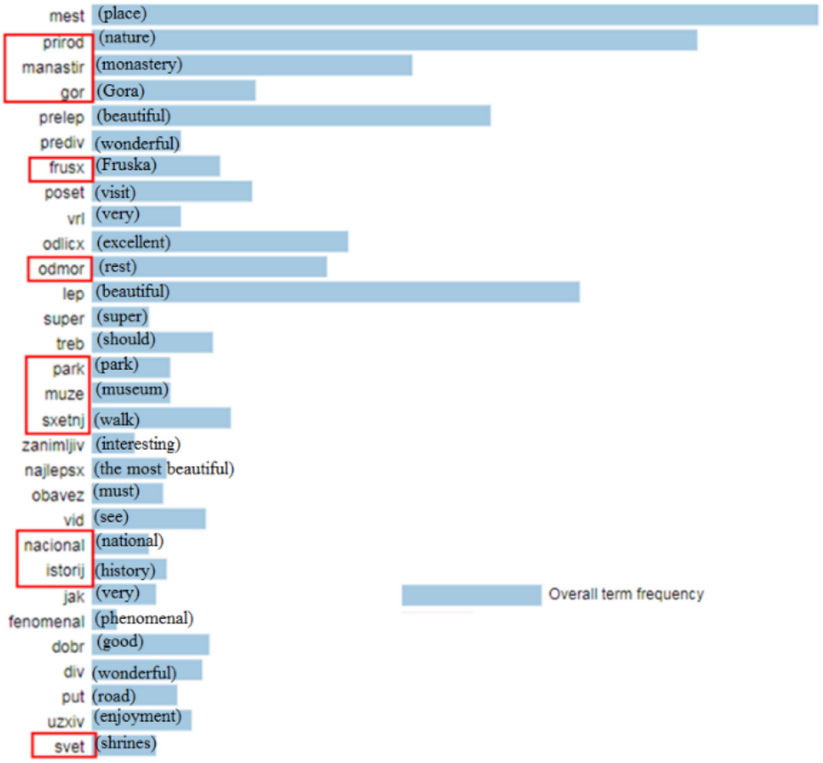


Figure 3. The most salient keywords in the corpus*

* Words are stemmed and thus illustrated in stem form. The illustration is generated in pyLDAvis.

Source: Own research

Each online review from the corpus is labeled with the topic with the highest probability. To assign meaningful names to topics authors used these labels and topic keywords. Table 3 illustrates 14 distinct topics, their indicative names, keywords that characterize a particular topic, and the percentage of corpus tokens a topic covers. Given the imbalanced data collection, it was expected that topics relating to nature and Fruška Gora would be dominant (Topic 7: *Body and soul rest* - 25.4% coverage, Topic 11: *General enthusiasm for Fruška Gora* - 11.9% coverage, and topics with less coverage Topic 2: *Parties and trash*, Topics 12: *Trail markings and organization*, Topic 13: *Phenomenal beauty in every season*, and Topic 14: *Spiritual oasis*), followed by topics related to monasteries (Topic 4: *Shrines and monasteries*), and museums (Topic 1: *Museum exhibits*, Topic 10: *Sirmium*, Topic 8: *Extraordinary museums' history*). Topic 3: *Family time*, Topic 5: *Maintenance and arrangement of picnic areas*, and Topic 6: *General delight* are either general topics or refer to similar categories of online reviews (Museums & Monasteries or Nature & Nature reserve).

The resulting topics reveal the variety of profiles of visitors to the Srem attractions and destinations: active tourists (Topic 5 and Topic 12), gastronomic tourists (Topic 9), tourists searching for relaxation and peace (Topic 7 and Topic 13), history-lovers (Topic 1, Topic 8, and Topic 10), spiritually-oriented (Topic 4 and Topic 14), or family-oriented tourists (Topic 3). Marketers can use this knowledge to design campaigns aiming at different tourist profiles. For these purposes, marketers can use identified keywords to uncover particularities related to various tourist profiles or extend the analysis to extract n-grams. N-grams represent two or more related words frequently appearing together in the corpus. Also, within the identified topics, only one topic - Topic 2, is strictly focused on the criticisms of tourists. Topic 2 indicates a problem related to the garbage left after parties, which should understood as a suggestion for improvement for resort authorities.

Table 3. Top keywords per topic and topics' corpus coverage*

Topic no. / name	Keywords	% of token coverage
Topic 1: ***Museum exhibits***	museum, culture, setting, modern, Vučedol, pleasant, superb, arrangement, end, hours, exhibits, past, interesting, simple, site, Vučedol, level, collection	6.1%
Topic 2: ***Parties and trash***	garbage, extra, fountain, sometime, ruin, gather, banister, drunk, parties, chaos, gathering, carousel, arises, swamp, broken, made, youth, cans, staircase, thrown, broken, glasses	3.8%
Topic 3: ***Family time***	beautiful, interesting, hiking, setting, relaxed, greenery, place, trash, simple, everywhere, curator, preserved, map, colony, Flavian, excellent, place, instructive, works, everywhere, extraordinary, family, experience, leadership, Sirmium, extra, to flourish	4.4%
Topic 4: ***Shrines and monasteries***	monastery, beautiful, visit, nice, see, must, history, valuable, holy place, visit, church, museum, old, Orthodoxy, Serbian, bad, little, exception, thing, peace	15.6%
Topic 5: ***Maintenance and arrangement of picnic areas***	super, fine, recommendation, water, neat, maintained, mosquitoes, tame, extra, export, trails, atmosphere, trash, spirituality, entertainment, dog, wonderful, viewpoint, church gate, protected, pleasant, friendly, horror, arranged	4.3%
Topic 6: ***General delight***	extra, nice, great, beautiful, divine, interesting, fantastic, simple, wonderful, peace, paradise, pearl, fantasy, experience, incredible, gorgeous, potential, untapped, destination, oasis, spirituality, recommendations, exceptional	3.6%
Topic 7: ***Body and soul rest***	place, nature, beautiful, excellent, rest, peace, walk, good, wonderful, enjoyment, trails, relaxation, city, right, air, soul, full, clean, temple, shrine, trip, drive, pleasant	25.4%
Topic 8: ***Extraordinary museums' history***	historic, extraordinary, castle, museum, time machine, old, objects, delight, recommendations, interesting, building, iconostasis, god, extra, ruin, breath, best	4.2%
Topic 9: ***Food***	extra, healthy, fish, mangulica, meat, donkey, milk, specialty, catch, dried, cheese, tree, marathon, cut, hiking, visited, shame, seedling, trees, management	3.8%
Topic 10: ***Sirmium***	very, beautiful, best, view, imperial, near, palace, Roman, peace, history, city, museum, Mitrovica, don't miss, scenery, exhibition, excavations, civilization, beginning, excellent	5.9%
Topic 11: ***General enthusiasm for Fruška Gora***	Fruška, Gora, beautiful, national, park, the most beautiful, mountain, Serbia, Vojvodina, beauty, forest, paradise, love, good	11.9%
Topic 12: ***Trail markings and organization***	earth, always, perfect, return, paradise, smile, extra, constantly, signposts, organization, emergence, possibilities, ready, unmarked, stray, occasionally, corner, thoroughfare, specificity, crowd, populate, endemic, maintained	3.8%
Topic 13: ***Phenomenal beauty in every season***	phenomenal, extra, beautiful, poor, gift, possibilities, pale, recreational, tower, great, divine, fantastic, interesting, top, winter, simple, wonderful, glad, paradise, pearl, peace, summer, fantasy, experience, incredible, potential	3.6%
Topic 14: ***Spiritual oasis***	extra, beautiful, great, divine, interesting, fantastic, simple, wonderful, paradise, pearl, fantasy, experience, peace, incredible, gorgeous, potential, untapped, unique, destination, oasis, spirituality, recommendations, extraordinary, delighted, weak	3.6%

* The keywords are ordered from the highest probability for the given topic to the lowest among the top 30 keywords. Some topics have illustrations for less than 30 words because certain Serbian words have the same translation in English, or the keywords could be treated as domain stopwords.

Source: Own research

Authors of papers (Hu et al., 2019; Kirilenko et al., 2021) suggest that using the overall corpus for topic modeling analysis, especially when focusing on negative aspects, leads to the omission of some dissatisfaction cues. Although LDA has proven to be a good choice for unraveling general and frequent topics in the corpus, which depict the overall preferences of tourists of the Srem region, to dive deeper into the opinions and topics related to different categories of tourist attraction, the authors of this chapter have repeated experiments on four subsets of the corpus, i.e., on collected online reviews about Monasteries, Museums, Nature, and Nature reserves. Authors built models for various values of α and β hyperparameters (0.01, 0.31, 0.61, 0.91), different values of K (2-15), as well as validation sets (25%, 50%, 75%, 100%). Table 4 presents the results of optimal parameters for each corpus subset. Results indicate that topics related to Nature are the most diverse (14 topics), while people use the narrowest set of topics to express their stance towards Nature reserves (7 topics). For each corpus subset, the authors have analyzed topic models built using the optimal parameters illustrated in Table 4. The resulting topics for each category of online reviews are an extension of the general 14 topics discovered over the entire corpus.

Table 4. Optimal parameters for topic models per destination category subsets

	Highest coherence	K	Alpha	Beta
Monasteries	0.743	10	Asymmetric	0.91
Museums	0.663	9	Asymmetric	0.91
Nature	0.691	14	0.61	0.91
Nature reserve	0.731	7	Asymmetric	0.91

Source: Own research

Within 10 topics uncovered for the Monasteries category, the authors observe that visitors appreciate the peacefulness and spirituality surrounding the monasteries and churches. Special praise is given to the iconostasis, frescoes, and interior. People also emphasize the historical importance of monasteries. In particular, the facts that Fruska Gora monasteries were the home of the relics of the holy prince (srb. *knez*) Lazar, the place where Dositej Obradović became a monk, or the Brankovich endowment.

Within 9 topics uncovered for the Museums category, the authors observe visitors appreciate the possibility of getting closer to this region's history through museum exhibitions and the use of modern technologies, such as projections. What was not noticeable within 14 general topics is the particular interest in the beekeeping museum, placed within a winery, and the tastings they offer.

Aside from admiration for the nature of Fruska Gora, the authors observe that visitors offer criticism and suggestions for improvement. Uncovering knowledge about dissatisfaction is particularly important as it provides directions for improvements. Among 14 topics detected for the Nature category, one topic is dedicated to the bad condition of the roads, the reckless cutting of forests, and the garbage found everywhere on Fruska Gora. One topic refers to family visits, where visitors emphasize picnics, barbecues, walks, biking, and hiking as activities.

Within 7 topics uncovered for the Nature reserve category, the authors observe that visitors appreciate the possibility of family visits, children-friendly content, delicious food, and availability of donkey milk, known for its notable health effects. One topic is dedicated particularly to boat rides.

Table 5 provides an overview of the most salient keywords for each model (Monasteries, Museums, Nature, and Nature reserves). Some of the keywords presented in Table 5, such as beautiful (srb. *lepo*), nature (srb. *priroda*), place (srb. *mesto*), and vacation (srb. *odmor*), are salient for

all four categories of online reviews. However, certain words are indicative of a particular review category. In Table 5 these words are colored differently. The red words point to topics related to Monasteries, the blue words to Museums, the green to Nature, and the yellow to Nature reserves. The authors used the Voyant tool[4] to analyze the context surrounding these salient keywords. Visitors of the Srem monasteries and churches advise others to visit these localities (e.g. phrases: "*must visit" or "make sure to stop by"*), they emphasize the spirituality surrounding the monasteries (e.g. phrases: *spiritual experience*, *spiritual peace*, *spiritual relaxation*), and indicate monasteries as a heaven for the soul, a perfect place for spiritual and physical rest. Blue salient words related to museums indicate that people love wine and honey degustations, and visitors highlight the cultural and historical content at museum's exhibitions. Green salient words are related to Nature and, in particular, to Fruška Gora. They indicate visitors value marked trails for hiking and biking and clear air. Tourists emphasized the roads are in bad condition and poorly maintained. Yellow salient words are words particularly indicative of Nature reserves in Srem. Based on the context analysis, the authors could reveal that people value boat rides, flora and fauna diversity, and the possibility of purchasing donkey milk.

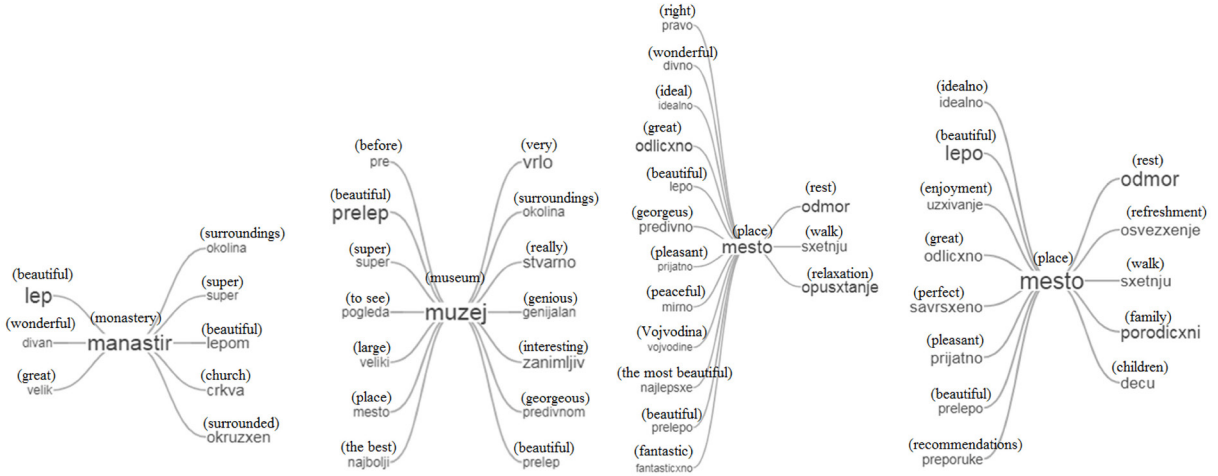
Table 5. Overview of leading keywords per review categories

Monasteries	**Museums**	**Nature**	**Nature reserve**
Beautiful (*lepo, prelepo, predivno*)	Super (*super*)	Nature (*priroda*)	Place (*mesto*)
Place (*mesto*)	Wine (*vino*)	Place (*mesto*)	Nature (*priroda*)
Monastery (*manastir*)	Beautiful (*lepo, prelepo*)	Holiday (*odmor*)	Holiday (*odmor*)
Peace (*mir*)	Great (*odlično*)	Gora	Beautiful (*predivno*, *lepo, prelepo*)
One (*jedan*)	Culture (*kultura*)	Great (*odlično*)	Ride (*vožnja*)
Nature (*priroda*)	Honey (*med*)	Fruška	
Should (*treba*)	Place (*mesto*)	Beautiful (*lepo, prelepo, predivno*)	Relaxed (*opušteno*)
Spirituality (*duhovnost*)	Museum (*muzej*)	Road (*put*)	Peace (*mir*)
Holiday (*odmor*)	A bit (*malo*)	Trail (*staza*)	A bit (*malo*)
Most beautiful (*najlepše*)	Old (*staro*)	Walk (*šetnja*)	Does not have (*nema*)
Soul (*duša*)	Good (*dobro*)	Relaxed (*opušteno*)	Boat (*brodić, brod*)
Visit (*obići*)	Visit (*posetiti*)	Most beautiful (*najlepše*)	See (*videti*)
Must (*obavezno*)	Really (*stvarno*)	Peace (*mir*)	Where (*gde*)
Holy (*sveto*)	Vučedol	Should (*treba*)	River (*reka*)
Very (*vrlo*)	Very (*vrlo*)	Enjoyment (*uživanje*)	More (*više*)
Paradise (*raj*)	Something (*nešto*)	Bad (*loš*)	Good (*dobro*)
History (*istorija*)	Before (*pre*)	Monastery (*manastir*)	Some (*neke*)
See (*videti*)	Winery (*vinarija*)	Mountain (*planina*)	
Road (*put*)	A lot (*puno*)	Park	Walk (*šetnja*)
Krušedol	Exhibition (*izložba*)	A bit (*malo*)	Should (*treba*)
A bit (*malo*)	Beekeeping (*pčelarstvo*)	Air (*vazduh*)	Located (*nalazi*)
Very (*jako*)	Located (*nalazi*)	Serbia (*Srbija*)	Come (*doći*)
Super (*super*)	Year (*godina*)	Perfect (*savršeno*)	Species (*vrste*)
Exceptionally (*izuzetno*)	History (*istorija*)	Vojvodina	Park
Worth (*vredi*)		Town (*grad*)	Milk (*milk*)
Fruškogorski		Picnic area (*izletište*)	Reserve (*rezervat*)
Gora		National (*nacionalni*)	Really (*zaista*)
		Clear (*čist*)	Can (*može*)

Source: Own research

[4] Voyant tool https://voyant-tools.org/.

The word tree is a tool within the Voyant suite. It allows exploration of the use of keywords in different phrases in the corpus. Figure 4 illustrates word trees for four sub-corpora. By default, the most common term is used as the first-word tree root. In our case, the first words are indicative of the category they belong to: Figure 4a: monastery, Figure 4b: museum, Figure 4c and Figure 4d: place. Word trees provide additional context. Visitors emphasize the surroundings and the ambient in which the monasteries are located – Figure 4a; the fun and interesting of museums – Figure 4b; the nature and Fruška Gora are perfect for rest, walks, and relaxation – Figure 4c; nature reserves are a great choice for the rest, refreshment, walks, family time, and children – Figure 4d.



a) Monasteries b) Museums c) Nature d) Nature reserve

Figure 4. Word trees for four sub-corpora

Source: Own research

5. CONCLUSION

By applying topic modeling over large amounts of unstructured data, such as online reviews or social media posts, hospitality or tourism managers can enhance their understanding of tourists' preferences, sentiments, and interests hidden in prevailing topics, as well as identify distinct visitor profiles and trends. Discovered knowledge and insights can be used for enhancement of resource allocation, marketing, or services, shifting the overall tourism offer to a more personalized and tailored one.

The research presented in this chapter presents the results of an analysis of online reviews of tourism destinations and attractions in the Srem Region. Authors applied Latent Dirichlet Allocation topic modeling, experimented with various combinations of alpha (α) and beta (β) hyperparameters, which control the document-topic and topic-word distributions, respectively, to build various models, and used coherence score as a measure of a model quality to determine the model with the optimal number of topics (K). The result of the experimentation is 1560 LDA models. The optimal model, with a coherence score of 0.717, revealed that 14 topics best represented the tourists' perceptions of the Srem region. These topics encompassed various aspects, including nature, relaxation, historical sites (monasteries and museums), and overall satisfaction, as indicated by the prevalence of positive adjectives in the reviews. The authors identified diverse visitor profiles, the active tourists, and those oriented toward gastronomy, leisure, history, and family. Their unique perspectives on tourism in the Srem region provide valuable insights for destination marketing and the design of marketing campaigns.

Furthermore, the authors extended the analysis to subsets of reviews related to specific categories of attractions, such as Monasteries, Museums, Nature, and Nature reserves. The analysis revealed distinct preferences and concerns of visitors within each category. For instance, visitors to monasteries emphasized spirituality and historical significance, while those interested in museums valued wine and honey tastings and exhibitions. The authors also identified specific sets of keywords typical for each category, which can be used for designing marketing strategies tailored to each group of visitors.

The research has limitations which will be addressed through future research. The corpus could be extended with additional data sources, such as travel blogs, forums, and news posts. This will provide a broader perspective on the preferences of the Srem region visitors. The results are informative and we can conclude that LDA has done a good job of modeling hidden thematic structures within a corpus which has revealed the general preferences of tourists of the Srem region. However, analysis of available methods points to potential in other topic modeling approaches. The future research directions will be focused on extending the experiments onto various topic modeling algorithms and benchmarking their performance over Serbian texts.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems, 112*, 102131. https://doi.org/10.1016/j.is.2022.102131

Aghdam, M. H. (2022). A novel constrained non-negative matrix factorization method based on users and items pairwise relationship for recommender systems. *Expert Systems with Applications, 195*, 116593. https://doi.org/10.1016/j.eswa.2022.116593

Ali, T., Omar, B., & Soulaimane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. *MethodsX, 9*, 101894. https://doi.org/10.1016/j.mex.2022.101894

Anandarajan, M., Hill, C., & Nolan, T. (2019). *Practical text analytics: Maximizing the Value of Text Data.* Springer Cham. https://doi.org/10.1007/978-3-319-95663-3

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App). *Journal of Business and Psychology, 33*, 445–459. https://doi.org/10.1007/s10869-017-9528-3

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(2003), 993-1022. Retrieved from https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Calheirosa, A., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling. *JOURNAL OF HOSPITALITY MARKETING & MANAGEMENT, 26*(7), 675–693. https://doi.org/10.1080/19368623.2017.1310075

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* (pp. 4171–4186). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423.pdf

Dixon, S. (2023). *Media usage in an internet minute as of April 2022.* Statista.

Egger, R. (2022). *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications.* Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-88389-8

Feldman, R., & Sanger, J. (2007). *The text mining handbook - Advanced approaches in analyzing unstructured data.* Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511546914

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How Many Topics? Stability Analysis for Topic Models. *Machine Learning and Knowledge Discovery in Databases*, 498-513. https://doi.org/10.1007/978-3-662-44848-9_32

Grljević, O. (2016). *Sentiment u sadržajima sa društvenih mreža kao instrument unapređenja poslovanja.* Subotica, Srbija: Autorski reprint.

Grljević, O. (2023). *Analiza sadržaja društvenih medija: Napredni pristupi analizi nestrukturisanih podataka.* Subotica: Ekonomski fakultet u Subotici.

Grljević, O., Bošnjak, S., Pavlićević, V., & Pavlović, N. (2019). Analysis of public stance on tourism destinations in Srem/Srijem region. In V. Bevanda, & S. Štetić (Eds.), *4th International Thematic Monograph: Modern Management Tools and Economy of Tourism Sector in Present Era* (pp. 267-290). Beograd: Association of Economists and Managers of the Balkans in cooperation with the Faculty of Tourism and Hospitality, Ohrid, North Macedonia. https://doi.org/10.31410/tmt.2019.267

Grljević, O., Bošnjak, Z., & Kovačević, A. (2022). Opinion mining in higher education: a corpus-based approach. *Enterprise Information Systems, 16*(5), 1773542. https://doi.org/10.1080/17517575.2020.1773542

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management, 72*, 417-426. https://doi.org/10.1016/j.tourman.2019.01.002

Hudson, S., Roth, M., Madden, T., & Hudson, R. (2015). The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees. *Tourism Management, 47*, 68-76. https://doi.org/10.1016/j.tourman.2014.09.001

Hutchison, P. D., Daigle, R. J., & George, B. (2018). Application of latent semantic analysis in AIS academic research. *International Journal of Accounting Information Systems, 31*, 83-96. https://doi.org/10.1016/j.accinf.2018.09.003

Kim, K., Park, O., Barr, J., & Yun, H. (2019). Tourists' shifting perceptions of UNESCO heritage sites: lessons from Jeju Island-South Korea. *Tourism Review, 74*(1), 20-29. https://doi.org/10.1108/TR-09-2017-0140

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

Korenčić, D., Ristov, S., Repar, J., & Šnajder, J. (2021). A Topic Coverage Approach to Evaluation of Topic Models. *IEEE Access, 9*, 123280-123312. https://doi.org/10.1109/access.2021.3109425

Kovačević, A., Grljević, O., Bošnjak, Z., & Svilengaćin, G. (2020). The Linguistic Construction of Sentiment Expressions in Student Opinionated Content: A Corpus-based study. *Poznań Studies in Contemporary Linguistics, 56*(2), 207–249. https://doi.org/10.1515/psicl-2020-0006

Laureate, C. D., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-023-10471-x

Lee, D. D., & Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791. https://doi.org/10.1038/44565

Lee, H., & Kang, Y. (2021). Mining tourists' destinations and preferences through LSTM-based text classification and spatial clustering using Flickr data. *Spatial Information Research, 29*, 825–839. https://doi.org/10.1007/s41324-021-00397-3

Li, W., Guo, K., Shi, Y., Zhu, L., & Zheng, Y. (2018). DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. *Knowledge-Based Systems, 146*, 203-214. https://doi.org/10.1016/j.knosys.2018.02.004

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Haussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

Milosevic, N. (2012). Stemmer for Serbian language. *arXiv*, arXiv:1209.4471v1. https://doi.org/10.48550/arXiv.1209.4471

Moody, C. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *arXiv*, arXiv:1605.02019v1.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*(2), 111-126. https://doi.org/10.1002/env.3170050203

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135. http://dx.doi.org/10.1561/1500000011

Papilloud, C., & Hinneburg, A. (2018). *Qualitative Textanalyse mit Topic-Modellen: Eine Einführung für Sozialwissenschaftler.* Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-21980-2

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2013). Evaluating topic coherence measures. *Neural Information Processing Systems Foundation (NIPS 2013).* https://doi.org/10.1145/2684822.2685324

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval.* New York, NY, USA: McGrawHill, Inc.

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

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. *Proceedings of the 31st International Conference on Machine Learning. 32(1)*, pp. 190-198. Beijing, China: JMLR: W&CP. Retrieved from http://proceedings.mlr.press/v32/tang14.pdf

Ubiparipovic, B., Matkovic, P., Marić, M., & Tumbas, P. (2020). Critical factors of digital transformation success: A literature review. *Ekonomika preduzeća, 5-6*(septembar-oktobar 2020), 400-415. https://doi.org/10.5937/EKOPRE2006400U

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

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation Methods for Topic Models. *Proceedings of the 26 th International Conference on Machine Learning* (pp. 1105–1112). Montreal, Canada: ACM. https://doi.org/10.1145/1553374.1553515

Wen, H., Park, E., Tao, C.-W., Chae, B., Li, X., & Kwon, J. (2020). Exploring user-generated content related to dining experiences of consumers with food allergies. *International Journal of Hospitality Management, 85*, 102357. https://doi.org/10.1016/j.ijhm.2019.102357

Yu, J., & Egger, R. (2021). Tourist Experiences at Overcrowded Attractions: A Text Analytics Approach. In W. Wörndl, C. Koo, & J. Stienmetz, *Information and Communication Technologies in Tourism 2021.* Cham: Springer. https://link.springer.com/chapter/10.1007/978-3-030-65785-7_21

Zou, S. (2020). National park entrance fee increase: a conceptual framework. *Journal of Sustainable Tourism, 28*(12), 2099-2117. https://doi.org/10.1080/09669582.2020.1791142