Identifying hidden semantic structures in Instagram data: a topic modelling comparison

Roman Egger and Joanne Yu

Roman Egger is based at the Department of Innovation and Management in Tourism, Salzburg University of Applied Sciences, Salzburg, Austria.
Joanne Yu is based at the Department of Tourism and Service Management, Modul University Vienna, Vienna, Austria.

Abstract

Purpose – *Intrigued by the methodological challenges emerging from text complexity, the purpose of this study is to evaluate the effectiveness of different topic modelling algorithms based on Instagram textual data.*

Design/methodology/approach – *By taking Instagram posts captioned with #darktourism as the study context, this research applies latent Dirichlet allocation (LDA), correlation explanation (CorEx), and non-negative matrix factorisation (NMF) to uncover tourist experiences.*

Findings – *CorEx outperforms LDA and NMF by classifying emerging dark sites and activities into 17 distinct topics. The results of LDA appear homogeneous and overlapping, whereas the extracted topics of NMF are not specific enough to gain deep insights.*

Originality/value – *This study assesses different topic modelling algorithms for knowledge extraction in the highly heterogeneous tourism industry. The findings unfold the complexity of analysing short-text social media data and strengthen the use of CorEx in analysing Instagram content.*

Keywords *Instagram, Machine learning, LDA, Topic modelling, CorEx, NMF*

Paper type *Research paper*

语意大解密：主题模型的比较

研究目的：基于对文本复杂性的兴趣，本研究以*Instagram*文本数据为基准，旨在比较不同主题建模的算法的有效性。

研究方法：本研究以标有*#darktourism*的*Instagram*帖子作为背景，评估直观理解（*LDA*），相关解释（*CorEx*）和非负矩阵分解（*NMF*）在分析与黑暗观光相关的帖子的实用性。

研究结果：*CorEx*分析出*17*个新兴的黑暗景点和活动，亦胜过*LDA*和*NMF*。虽然*LDA*能探讨出较多的主题数，但它们的内容几乎重复。同样的，尽管*NMF*适用于短文本数据，但它提取出主题相当笼统且不够具体。

原创性：透过将营销和数据科学学科相结合，本研究为分析非结构化的文本奠定了基础，并证实了*CorEx*在分析短文本社交媒体数据（如*Instagram*数据）中的效益。

关键字：*Instagram; LDA; NMF; CorEx;* 主题建模; 机器学习

纸张类型：研究论文

Received 20 May 2021
Revised 22 August 2021
Accepted 9 September 2021

Author contribution:
1. Conception or design of the work: Roman Egger
2. Data collection: Roman Egger
3. Data analysis: Roman Egger
4. Data interpretation: Roman Egger and Joanne Yu
5. Drafting the article: Joanne Yu
6. Critical revision of the article: Joanne Yu
7. Final approval of the version to be published: Roman Egger and Joanne Yu

Identificación de estructuras semánticas en datos de Instagram: una comparación de modelos de temas

Resumen

Propósito : *Intrigado por los desafíos metodológicos que surgen de la complejidad del texto, este estudio evalúa la efectividad de diferentes algoritmos de modelado de temas basados en datos textuales de Instagram.*

Metodología : *Al tomar publicaciones de Instagram con #darktourism como contexto de estudio, esta investigación aplica la asignación de Dirichlet latente (LDA), la explicación de correlación (CorEx) y la factorización matricial no negativa (NMF) para descubrir experiencias turísticas.*

Resultados : *CorEx supera a LDA y NMF al clasificar los sitios y actividades oscuros emergentes en 17 temas distintos. Los resultados de LDA son homogéneos y se superponen, mientras que los temas extraídos de NMF no son lo suficientemente específicos como para obtener conocimientos profundos.*

Originalidad : *Este estudio evalúa diferentes algoritmos de modelado de temas para la extracción de conocimiento en la industria del turismo. Los hallazgos revelan la complejidad de analizar datos de redes sociales de texto corto y fortalecen el uso de CorEx para analizar el contenido de Instagram.*
Palabras llave : *Instagram; LDA; NMF; CorEx; modelado de temas; aprendizaje automático*
Tipo de papel : *Trabajo de investigación*

Introduction

Marketers have embraced social media to enable new marketing strategies by connecting with the audience, increasing brand awareness and influencing purchase intention (Leung *et al.*, 2019). Platforms such as Facebook, Twitter and emergingly, Instagram, empower users to share feelings and experiences with other individuals in real-time (Shawky *et al.*, 2019). Commonly known as user-generated content (UGC), each individual is able to create content in the form of text, videos, or images online (Dedeoğlu *et al.*, 2021). As an experience-based industry (Sthapit *et al.*, 2020), tourism is exactly where UGC and social media come into play. The democratisation of knowledge, driven by technologies and easy access to information (Martini and Buda, 2020), allows tourists to share their experiences during a multiphasic travel journey (Leung *et al.*, 2019). To date, studies already witness a paradigm shift in developing marketing strategies and conducting scientific research by analysing UGC (Mariani, 2019; Köseoglu *et al.*, 2020). Particularly, Instagram has been recognised by marketers due to its popularity among tourists for sharing travel-related experiences (Arefieva *et al.*, 2021).

Nonetheless, understanding tourist experiences through Instagram is yet to be mainstream in marketing (Filieri *et al.*, 2021). Because UGC is unstructured by nature (Bharadwaj *et al.*, 2020), the complexity of texts which often involves hashtags, emojis and slangs, may hold researchers back in analysing Instagram content. While Instagram is known as a visual platform, text captions play an important role in conveying meanings that may not be evident visually (Munoz and Towner, 2017). Seeing that textual content of Instagram posts, which mainly consist of very short texts, are largely under-researched in tourism (Cip *et al.*, 2019), researchers urgently need to cope with the methodological challenges emerging from text complexity based on short-text data.

This study aims to evaluate different topic models based on Instagram textual data to facilitate tourism practitioners and researchers in capturing hidden semantic structures and patterns of travel experiences shared on social media. Previous studies have primarily used Latent Dirichlet Allocation (LDA) as a topic modelling algorithm, with both topic selection and evaluation often left underexposed. The often poor quality of identified topics is the biggest hurdle for accepting statistical topic models outside the machine learning community (Cai *et al.*, 2018). Thus, this study presents alternative topic modelling approaches and compares their results so as to support the decision-making when selecting appropriate approaches. By combining the disciplines of marketing and data science with tourism, it helps marketers and researchers to evaluate the effectiveness and suitability of different techniques.

Literature review

Data analytics in tourism: the importance of text

At the intersection of computer science, statistics and machine learning, data science emerges as an essential tool for business intelligence (Mariani *et al.*, 2018b) and fosters analytic techniques (Mariani and Wamba, 2020). With a particular focus on tourism and hospitality, applications can be seen from analysing online reviews (e.g. on TripAdvisor), UGC (e.g. on Facebook and Instagram) and consumer behaviour (e.g. on airline platforms), among others (Mariani *et al.*, 2018a). By transforming unstructured information into

interpretable patterns (Bharadwaj *et al.*, 2020), data analytics challenges our previous philosophical and epistemological assumptions on how knowledge in social sciences is developed (Kitchin, 2014). Different from traditional approaches where theories need to be tested and verified (Soares *et al.*, 2021), data-intensive scientific discovery breaks the boundaries of state-of-the-art knowledge (Kitchin, 2014) by allowing tourism researchers to derive novel insights exploratorily based on pertinent research objectives.

Social media has gained momentum in tourism marketing to gain insights into consumer experience (Gaffar *et al.*, 2021; Peco-Torres *et al.*, 2021). While analysing UGC on social media has become more commonplace in recent years, a strong focus is placed on Facebook, followed by YouTube and Twitter (Shawky *et al.*, 2019). Comparably, Instagram, as a key channel for marketing and one of the fastest-growing applications in tourism (Filieri *et al.*, 2021), still receives little attention (Arefieva *et al.*, 2021). As individuals (un)consciously convey their thinking styles, feelings and experiences through text (Bharadwaj *et al.*, 2020) or pictures (Munoz and Towner, 2017), scholars have initiated a call to analyse tourist experiences shared on Instagram. For instance, Arefieva *et al.* (2021) applied machine learning and concluded key concepts related to tourist destination image on Instagram. Filieri *et al.* (2021) adopted text analytics based on captions and hashtags of Instagram posts to investigate tourists' feelings at popular destinations. Likewise, Irawan *et al.* (2020) performed topic modelling to uncover tourists' perceptions towards attractions in India on Instagram.

Although Instagram is a visual platform, the value of textual components should not be overlooked because text illuminates the meaning embedded in pictures (Munoz and Towner, 2017). From linguistic viewpoints, since the context and disambiguation of meanings are often hidden in a piece of text, extracting key aspects from the corpus is necessary for text comprehension (Dash, 2008). In tourism, despite the prominence of travel photos, tourist experiences and destination attributes cannot be fully captured without unveiling the latent meaning from textual captions (Munoz and Towner, 2017). However, since UGC typically lacks metadata and is not organised in a predefined manner (Bharadwaj *et al.*, 2020), new methods such as topic modelling and machine learning are required for tourism practitioners to extract insights (Arefieva *et al.*, 2021).

Coping with unstructured text data: topic modelling solutions in tourism

Due to the concise and text-heavy nature of UGC, the foremost step is to demystify hidden semantic structures to gain a deep understanding of tourist experiences (Maier *et al.*, 2018). Among various kinds of text analysis methods, topic modelling serves as a novel approach in innovation analytics (Kakatkar *et al.*, 2020), business marketing (Reisenbichler and Reutterer, 2019) and emergingly tourism social media research (Kim *et al.*, 2019; Arefieva *et al.*, 2021). Rooted in machine learning and natural language processing (NLP), topic modelling identifies latent topics based on co-occurrences of terms from a corpus of text data and has an explorative character (Albalawi *et al.*, 2020). In general, topics can be identified either with supervised, unsupervised, or semi-supervised machine learning models (Mariani *et al.*, 2018a). In supervised methods, a machine is trained based on labelled data sets, which corresponds to a classification approach. Typical topic modelling is mainly performed by unsupervised algorithms that attempt to group unlabelled data into topics. More recent topic modelling models allow to guide the algorithm by seeding keywords, which corresponds to semi-supervised learning. Although topic modelling benefits exploratory research (Maier *et al.*, 2018), it is worth mentioning that interpretation relies heavily on researchers' domain knowledge. In most cases, the identified topics neither match human judgement, nor are they relevant and specific to the research context (Cai *et al.*, 2018). Additionally, algorithm evaluation is particularly critical because short-text posts often contain noisy data (Albalawi *et al.*, 2020).

Among several types of topic models, LDA is agreed as the most popular algorithm (Gallagher *et al.*, 2017). The intuition is that each document consists of several heterogeneous topics, while

each topic is composed of similar words. Its core strength is that the algorithm can infer the latent topic structure where no predefined classes are required (Maier *et al.*, 2018). Through an iterative process, words are re-assigned to a topic until convergence (Hannigan *et al.*, 2019). An example can be seen from the study of Kim *et al.* (2019), which concludes ten unique topics discussed by international tourists on UGC using LDA-based topic modelling. Nevertheless, as an unsupervised learning method, one major flaw lies in its high dependence on words and their frequency in a corpus, which may result in overlapping clusters (Hannigan *et al.*, 2019).

In addition to LDA, a newly semi-supervised approach used in recent social media literature that merits attention is correlation explanation (CorEx) (Rizvi *et al.*, 2019). CorEx is based on an information-theoretic framework, suggesting a process of gaining information when the outcome is unknown before the analysis (Gallagher *et al.*, 2017). Hence, unlike LDA focussing on generated topics, CorEx searches for informative representations that can best explain certain phenomena (Rizvi *et al.*, 2019). Meanwhile, CorEx provides the model with anchor words, allowing researchers to incrementally refine topics for higher interpretability of results (Gallagher *et al.*, 2017). Such flexibility empowers users to develop creative strategies and enforces topic separability. Recently, Arefieva *et al.* (2021) classified the destination image of Austria into 15 features based on Instagram pictures shared by tourists. That is, CorEx has proved itself a suitable method for classifying complex social media posts related to travel and tourism by considering multiple attributes within the data (Arefieva *et al.*, 2021).

Turning to another mathematical framework, non-negative matrix factorisation (NMF) is a linear-algebraic model capable of improving topic coherence (Blair *et al.*, 2020). Its applications range from artificial intelligence and computer vision to signal processing and bioinformatics. Lately, scholars suggest that NMF can transform sparse data into a dense matrix without losing latent information (Kujiraoka *et al.*, 2017). Moreover, when learning in noisy domains (e.g. social media) (Blair *et al.*, 2020), its strength is to detect meaningful topics without prior knowledge of the original data, making it particularly suitable for analysing real-time data, text and image (Albalawi *et al.*, 2020). Nonetheless, with regards to semantic interpretation, NMF may not always provide accurate results (Albalawi *et al.*, 2020) and produce less coherent topics than LDA due to weaknesses of non-unique factorisation (Wang and Zhang, 2021).

Methodology

While acknowledging that tourism involves both hedonic and negative/unfavourable experiences, the latter incident is rarely evaluated (Sthapit *et al.*, 2020). Visiting dark tourism attractions, for example, is representative of such context. Intending to explore the aspects of tourist experiences that might be neglected, this research applies dark tourism experiences as the study context based on narratives shared on Instagram. A detailed description of the case and the methodological procedures follow.

Study context: dark tourism

Dark tourism suggests the act of visiting places associated with death, suffering, disaster, or the seemingly macabre, such as concentration camps (Commane and Potton, 2019), battlefields and cemeteries (Light, 2017). Increasingly, the commodification of death and people's awareness of dark sites (Martini and Buda, 2020) has opened a new scientific paradigm for dark tourism. Particularly, Instagram functions as a new lens for tourists to exchange dark experiences (Carter-White, 2018) and adds value to dark tourism management (Commane and Potton, 2019). While Instagram encourages consumers to engage with memorial sites and historical events (Commane and Potton, 2019), earlier research mainly features a top-down method by first identifying a specific attraction as the study context or assessing dark experiences based on predefined measurements (Light, 2017).

Seeing that Instagram has caused the democratisation of heritage, the image and understanding of dark tourism can be re-contextualised by analysing Instagram posts (Carter-White, 2018) based on travellers' experiences towards tragic events and/or sites. Methodologically speaking, dark tourism remains as a niche market, allowing researchers to analyse almost the whole population and look at the phenomenon from a holistic perspective. Furthermore, even being a niche segment, its variety (perception-wise) is large enough to extract meaningful/multiple topics (Light, 2017). This is particularly important for topic modelling studies because a stable (time-wise) and diverse context is necessary.

Data collection and pre-processing

This research compares different topic modelling algorithms to evaluate their effectiveness in processing short-text data based on the case of dark tourism experiences. Out of the 33,881 Instagram posts tagged with #darktourism at the time of data collection in March 2020, a total of 26,581 public posts from 6,650 accounts were crawled using *Phantombuster*. Extracted data included captions, date of the post, check-in location, post URL and users' account type (i.e. personal/business). To exclude commercial content, 1,939 business accounts and their related posts were removed. Non-English posts were deleted, resulting in 12,835 posts published by 4,711 personal accounts.

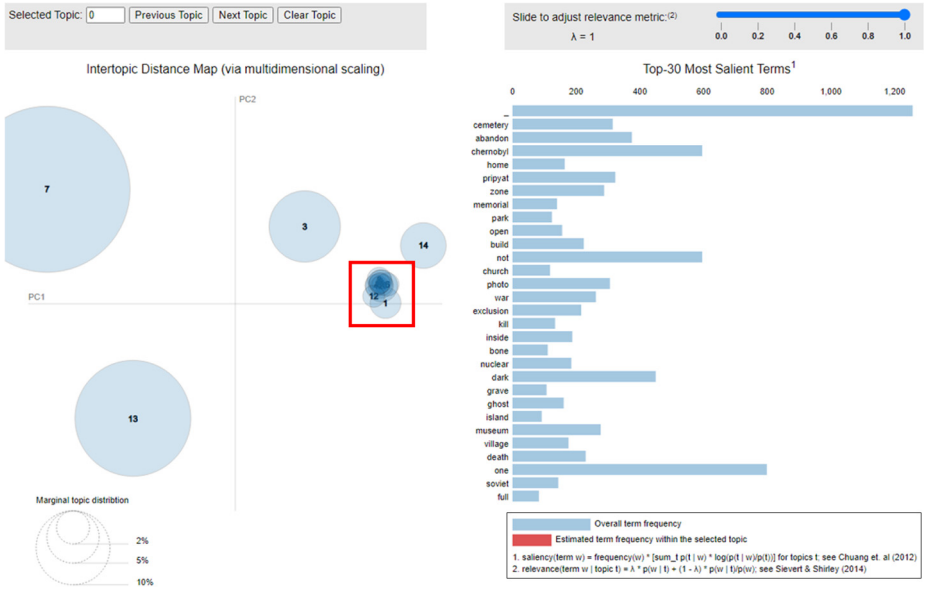
Next, data pre-processing was conducted using NLP modules in Python. The corpus was tokenised, stop words were removed, non-informative texts (e.g. numbers, emojis, abbreviations, unknown characters and usernames) were excluded. Slang words were reformed and diacritics were converted to a basic format. Stemming was performed using Porter Stemmer to remove word suffixes (e.g. changing to chang). Words were lemmatised to their actual base forms using WordNet Lemmatizer (e.g. changing to change). Finally, the text was transformed into vectors based on term frequency-inverse document frequency (*tf-idf*). Specifically, *tf-idf* suggests the importance of a term, being the logarithmically scaled quotient of the total number of documents in a corpus and the number of documents containing the word.

Procedures and results of topic modelling techniques

Three topic modelling techniques – LDA, CorEx and NMF – as discussed earlier, were applied to compare their performance and features. In LDA, a grid search, the process of finding optimal hyperparameters, on three essential hyperparameters (number of topics (K), alpha, beta) was performed. These hyperparameters should be defined beforehand with the aim of controlling the learning process. To identify the optimal hyperparameters, one always varies and the other hyperparameters remain constant. Since K is generated based on quantitative calculations, in the case that the results are ambiguous (e.g. non-sense results) or terms are incoherent, an evaluation of interpretability by experts is required (e.g. using intercoder reliability tests). After running several experiments, the best coherence score, suggesting the quality of the extracted topics, with 15 topics, was reached. In the next step, the grid search yielded a symmetric alpha value and a beta value of 0.91. A higher alpha indicates a greater mix of topics; a higher beta leads to more words pertaining to topics. To reinforce the decision, an intertropical distance map was generated with pyLDAvis (Islam, 2019). However, through a visual inspection, several overlapping topics were identified (Figure 1), which should be avoided (Passos *et al.*, 2011). Additionally, by means of grid search for an optimal hyperparameter setting, LDA cannot determine a suitable topic solution. The following section thus will not discuss the results of LDA in more detail.

With regards to NMF, NMF decomposes the term-document matrix into two factors: *W* and *H*. The former presents the connection between terms and topics; the latter implies the relationship between topics and documents (Chen *et al.*, 2019). *W* and *H* are always non-negative, minimising the risk of interpreting topics with negative entries for certain words

Figure 1 Visual inspection of LDA



Note: Link to the interactive visualisation https://tinyurl.com/darkt-LDA

(Wang and Zhang, 2021). The intuition is that researchers approximate the factors, although there is no guarantee whether the input matrix can be recovered. Typically, the number of subjects needs to be determined in advance, which is best done by calculating the coherence score. In this study, Gensim (Islam, 2019) was used to estimate the optimal number of topics. Seven topics could be identified, which did not overlap and may result in higher interpretability due to their small size. Human judgement, however, concluded that the topics were not specific enough for meaningful implications. Interpretation of such a decision is presented in the next section.

As for CorEx, since the number of topics depends on the distribution of total correlation (Gallagher *et al.*, 2017), researchers may decide whether to add more latent topics by observing the changes of the overall total correlation. After comparing different numbers of extracted topics to determine the optimal number, the inspection in this study returned a 17-topic solution with the highest correlation. To effectively extract latent factors in UGC, incorporating domain knowledge through anchor words could be used with CorEx to further optimise the results. For example, words such as "nuclear", "disaster" and "accident" could be used to extract a topic around these terms. By assessing the strength of anchor words, it facilitates researchers to decide to what extent the words are related to a topic (Gallagher *et al.*, 2017). However, this step was not proceeded in the research because the evaluation of anchor words corresponds to a more advanced theory- or domain-knowledge-driven approach, which is not present in LDA and NMF and would bias the comparison of the different methods.

Since CorEx outperformed NMF in terms of topic coherence and semantic interpretation, the 17 topics based on #darktourism posts generated by CorEx were used for further analysis. Explanations concerning the choice of algorithms are provided in the discussion section.

Results of the study context

Notably, since topic modelling still serves as a means for qualitative research, the results cannot be compared with a "value", but human judgement and interpretation are of high

relevance (Hannigan *et al.*, 2019). Table 1 provides an overview of the 17 identified topics of CorEx and seven topics of NMF. The identified topics of NMF are included for comparison. The topic names were given based on keywords with higher *tf-idf* weights, as explained earlier.

Regarding the site-specific topics, tourists discussed the Imperial Crypt in Vienna, the Berlin Wall and the abandoned textile factory in Germany, as well as the Old Concert Hall, the abandoned Olympic stadium and the Patarei fortress in Estonia. In Philadelphia, tourists were drawn to an annual professional wrestling tournament (i.e. Survival of the Fittest). In the Czech Republic, tourists visited Sedlec Ossuary. Meanwhile, some tourists described themselves as taphophiles interested in cemeteries/gravestones (e.g. Novodevichy Cemetery and Rookwood Cemetery); others highlighted sea forts as popular war-related sites.

Concerning topics related to past incidents, commonly mentioned events are the Chernobyl disaster and the Jeju uprising. Other topics include the Holocaust, concentration camps and Soviet nuclear weapons, which are more related to the Jewish community. Interestingly, tourists seem to enjoy urban exploration (urbex) and view abandoned environments as a

Table 1 Topics related to dark tourism

CorEx Topic	*Topic name*	*Keywords*	*NMF Topic*	*Keywords*
1	Chernobyl disaster	chernobyl, pripyat, ukraine, exclusion, chernobylexclusionzone	1	exclusion, chernobyl, pripyat chernobyl, zone ukraine
			2	pripyat, abandon, military, building, wheel
2	Urban exploration	urbex, stalker, urbandecay, urbexworld, utopia		NA
3	Cemetery impressions	cemetery, moscow, taphophile, russia, rookwood	3	cemetery, moscow, russia, vagankovskoe cemetery, novodevichie
4	Dark activities in Philadelphia	history, creepy, survivalofthefittest, philadelphia, gritty		NA
5	Nuclear power of the Soviet Union	nuclear, nuclearaccident, soviet communist, cccp, radioactive		NA
6	Graveyard impressions	graveyard, necropolis, freaks, darkness, fantasy		NA
7	War tourism	war, gunfight, abandoned, seaforts, seacoast		NA
8	Ghost hunting	creepy, wv, paranormalinvestigation, springs sanitarium, sweetspringsresort	4	home, sanitarium, sweet, springs, memorial home
9	Imperial Crypt in Vienna	crypt, vienna, imperial, habsburg, kaisergruft		NA
10	Jeju uprising	dark pictures, emotional, text, april, jejuisland		NA
11	History of Berlin	memory, loving, hearts, berlinwall, textile factory		NA
12	Sedlec Ossuary	kutnahora, bonechurch, sedlecossuary, czech republic		NA
13	Dark activities in Estonia	Patarei, concert hall, old, dame, stadium tallinn		NA
14	Historical Ainu culture	hiroshima, massachusetts, ainuhistory, ainupeople, ethnicclothes		NA
15	The Holocaust	deathcamp, auschwitz, konzentrationslager, holocausthistory		NA
16	Random travel posts at the abandoned sites	abandoned, keydestination, picoftheday, photo photooftheday	5	darktourism, tourist, tour, island, explore dark
			6	greeting america, london, field, cambodia, fiction
17	Travel photography	photographytravel, photography, instagram, gramtravel, travelworld	7	place, visit, time, year, make

Note: NA=not applicable

sort of utopia. The findings also disclosed a fascination for paranormal activities and ghost hunting in the Sweet Springs Sanitarium in West Virginia. Finally, the preservation of Ainu ethnic clothes is another emerging topic shared on Instagram.

Conversely, NMF only presents seven topics that are not diverse enough. For instance, topics 1 and 2 of NMF are both relevant to the Chernobyl disaster, leading to duplicated results. Site-specific topics include cemetery impressions in Russia and the Sweet Springs Sanitarium. Yet, unlike the keywords returned by CorEx for topic 8 (e.g. "paranormal investigation"), the keywords of NMF do not inform what kind of activities tourists were doing at the Sweet Springs Sanitarium. Furthermore, topic 5 of NMF contains several generic terms (e.g. "dark", "dark tourism"), while topic 6 seems to merge multiple contexts (e.g. "America", "London", "Cambodia"). Finally, several keywords extracted for NMF topic 7 (e.g. "time" "make" "year") appear to be either unrelated to dark tourism or not making sense.

Discussion

Since this study aims to inspect three different topic modelling techniques based on short-text data, extensive discussion on dark tourism is beyond the scope of this research. Additionally, it is important to keep in mind that there is no "best" topic modelling technique because the performance may vary, depending on the nature of the data sets. Table 2 summarises the advantages and disadvantages of the individual approach. Discussion of each algorithm, using the study's findings as examples, is as follows.

Table 2 Comparison of topic models

	Advantage	*Disadvantage*
LDA	• Good in finding coherent topics. • Able to deal with sparse input. • LDA-learned features are easy to be interpreted • No prior domain-knowledge required • Mixed membership (one document can contain several topics) • Provides full generative models with multinominal distribution over topics • Shows adjectives and nouns in topics	• Requires detailed assumptions and careful specification of hyperparameters • Topics are soft clusters which can result in overlapping topics • No objective evaluation metric available • Hard to find optimum number of topics (Number of topics needs to be specified by users) • Results are not deterministic and reliability and validity cannot be taken for granted • Assumes that topics are independent of each other (using word co-appearance frequency only; Word correlations are disregarded, relations between topics cannot be modelled)
NMF	• Mixed membership (one document can contain several topics) • Term-document matrix with *tf-idf* weighting can be used instead of raw word frequencies like in LDA • Computationally efficient and highly scalable • Easy to be implemented • No prior domain knowledge required	• Tends to provide incoherent topics • Hard to find optimum number of topics (Number of topics needs to be specified by users) • Implicit specification of probabilistic generative models
CorEx	• Does not require detailed assumptions and an underlying generative model • Has a good way to find optimum number of topics by increasing them until the correlation does not raise any more • Domain knowledge can be integrated through anchor words • Computational advantage for large datasets • Mixed membership (one document can contain several topics)	• Not suitable to process longer texts; shorter subdocuments are recommended • Uninformative seed-words need to be added to obtain more intuitive topics • Each word is assigned to only one topic

Generally, this study supports the effectiveness of CorEx in uncovering hidden insights when processing Instagram data. This can be seen from a number of site-specific topics barely mentioned in the existing literature (e.g. Sedlec Ossuary, the Imperial Crypt and Rookwood Cemetery). Moreover, the flexible nature of CorEx broadens the topic aspects (Gallagher *et al.*, 2017) by revealing ghost hunting at the Sweet Springs Sanitarium and urbex to be potentially interesting areas of investigation. Specifically, urbex can be visitations to the Chernobyl disaster site, sea forts, cemeteries, abandoned hospitals and sanatoriums.

Unlike LDA that often results in more generic models (Rizvi *et al.*, 2019), the findings are consistent with earlier social media literature (based on Twitter data), where CorEx outperforms LDA by returning the most coherent topics (Zhou *et al.*, 2019). Likewise, other scholars concluded that CorEx is more efficient because LDA often returns non-related terms (Alnusyan *et al.*, 2020). Despite that LDA remains a dominant topic modelling approach (Gallagher *et al.*, 2017), overlapping topics have always been an issue for learned LDA models due to the polysemy of words (Passos *et al.*, 2011). Similar to the results of LDA in this study, Cai *et al.* (2018) also affirmed duplicated topics as one of the fundamental problems because the divisions within a corpus would not be presented. These "bad" topics may arise when feature extraction lacks information for statistical learning (Cai *et al.*, 2018), especially in short texts that are noisy and sparse (Chen *et al.*, 2019). A potential solution for LDA could be to divide a large-scale textual corpus into segments to improve the comprehensiveness of the extracted topics. Ultimately, what is equally essential is hyperparameter tuning, referring to the process of selecting a set of hyperparameters to optimise modelling results. In this study, the calculation of topic coherence and hyperparameters such as the alpha and beta are considered essential to truly obtain an optimised model. By changing the hyperparameters, topic coherence provides a good hint regarding the frequency of the top co-occurred words in a topic (Maier *et al.*, 2018).

Although NMF appears to perform better than LDA especially when analysing social media posts (Chen *et al.*, 2019), scholars underlined that NMF has a lower capability in capturing ideas embedded in the text (Blair *et al.*, 2020). Similar to a study analysing short-text data on Facebook, NMF could not extract meaningful and logical topics (Albalawi *et al.*, 2020). Because NMF relies on the Frobenius norm (Chen *et al.*, 2019), word weights are not probability distributed. This potentially explains the study's findings that the NMF model only contains a small number of topics. Covertly, the findings may suffer from a poor-quality model where multiple topics are merged into one (Cai *et al.*, 2018). The issue of mixed topics also emerged in another research using NMF to extract insights from customer reviews (Kujiraoka *et al.*, 2017). To optimise topic modelling based on short texts, Chen *et al.* (2019) recently proposed a knowledge-guided NMF method by integrating the word–word semantic regularisation based on external knowledge from databases such as Wikipedia.

Conclusion

Theoretical and methodological contributions

Seeing the complexity of analysing short social textual data, this study broadens the scope of state-of-the-art data science methods for knowledge extraction to the highly heterogeneous tourism industry. Considering the broad application of topic models, this research initiates a call to detect trends and patterns based on UGC, closing the gap in data analytics for tourism marketing (Mariani *et al.*, 2018b). As topic modelling in tourism research is limited to LDA due to its probabilistic model with interpretable topics (Gallagher *et al.*, 2017), this study sheds light on other algorithms that could be more suitable for short-text data but are less known in tourism. While acknowledging the significance of using digital footprints to understand tourist behaviour (Möhring *et al.*, 2021), the linkages

between data science and tourism marketing are yet to be established (Mariani *et al.*, 2018a). This study thus lays the groundwork for applying topic modelling as an innovative method in exploratory studies to better comprehend short-text data from a bottom-up approach.

This study responds to a need for topic model evaluation (Reisenbichler and Reutterer, 2019) and provides insights for tourism scholars in choosing appropriate algorithms. When it comes to social media posts, the findings showcase the potential of CorEx, which remains a rarity in tourism (Arefieva *et al.*, 2021) and offer alternative guided modelling solutions for scholars interested in applying data analytics for marketing purposes. It is worth mentioning that the poor quality of topic models often results from an inadequate understanding of their application. In the case of UGC, it can be assumed that LDA is not an ideal method because the window size for LDA needs to be very small. Meanwhile, in many cases, insufficient information exists in tourism concerning the necessary steps of data pre-processing and how hyperparameters are defined. For instance, in some recent tourism studies, the default hyperparameters were used, and the number of topics was defined without estimating these parameters (Celuch, 2021; Yu and Egger, 2021). Consequently, non-ideal results might have been reported as a successful topic extraction. This research offers a critical view of the methodological evaluation, hoping to improve the acceptance of various topic modelling algorithms in tourism.

Practical implications

Drawing upon the intersection of marketing and data science, this study informs practices for various stakeholders. In general, since hashtags are an essential part of social media strategies, social media managers and advertisers, especially those on Instagram, are suggested to treat keywords (i.e. terms with higher *tf-idf* weights) as hashtags to create more informative posts. Effective use of hashtags improves engagement and exposure rate as they allow users to see all public posts related to a specific context. Meanwhile, the evaluation of three topic models provides social media managers a better understanding of methodological challenges (e.g. duplicated, generic and multi-context topics), which they may encounter during analysis. However, because posting practices vary across social media channels, it is important to keep in mind that this study is best applied to Instagram. As different social networks have distinct user experiences, cultures and demographics, marketers are advised to customise ads based on placements.

After all, the potential of analysing Instagram posts is unprecedented when it comes to optimising marketing strategies. Specific to tourism, to ensure congruity between the projected and perceived image (Arefieva *et al.*, 2021), this study presents CorEx as an ideal tool for destination managers to catch up with travel trends through identifying attractions that may not have been recognised. Take dark tourism as an example; Australian marketers could include Rookwood cemetery to personalise travel experiences for taphophiles. Likewise, marketers could include necropolises in a travel package, which appears to be more popular than cemeteries and graveyards. Other less-known activities, such as urbex and ghost hunting, that appears to be overlooked by marketers of dark tourism sites also demand attention. Although a country might not necessarily have places associated with tragic events, marketers are encouraged to explore abandoned buildings and construction sites to allow tourists to "touch" the history.

Limitations and recommendations

Since this research only evaluates three types of algorithms, scholars are encouraged to assess other modelling approaches such as Top2Vec or BERTopic that are also applicable to short-text data. Moreover, in addition to topic coherence used in this study, future research is recommended to conduct reliability checks to ensure the robustness of the topic solutions. One critical reason is that there are multiple topic modelling approaches,

each of which entails different human evaluations. Alternatively, to measure the coherence of the extracted topics, researchers can perform word- and topic-intrusion tasks (Chang *et al.*, 2009). Through manual groupings, the techniques can enlist a second pair of eyes for topic model evaluation because human understanding facilitates researchers' decision-making to change the hyperparameters or accept the results. Turning to Instagram posts, as experiences might be different across cultures, researchers may consider retrieving the metadata of posts such as Instagram users' country of origin if presented on user profiles. To continue extending tourism literature applying topic models, anchor words could be defined for CorEx, which help to steer topics in a particular direction to identify areas that might not have been recognised otherwise (Gallagher *et al.*, 2017). Finally, empirical research is encouraged to verify the identified topics in the context of dark tourism.

References

Albalawi, R., Yeap, T.H. and Benyoucef, M. (2020), "Using topic modeling methods for short-text data: a comparative analysis", *Frontiers in Artificial Intelligence*, Vol. 3, pp. 1-14.

Alnusyan, R., Almotairi, R., Almufadhi, S., Shargabi, A.A. and Alshobaili, J. (2020), "A Semi-Supervised approach for user reviews topic modeling and classification", *In 2020 International Conference on Computing and Information Technology*, pp. 1-5.

Arefieva, V., Egger, R. and Yu, J. (2021), "A machine learning approach to cluster destination image on instagram", *Tourism Management*, Vol. 85, p. 104318.

Bharadwaj, N., Ballings, M. and Naik, P.A. (2020), "Cross-media consumption: insights from super bowl advertising", *Journal of Interactive Marketing*, Vol. 50, pp. 17-31.

Blair, S.J., Bi, Y. and Mulvenna, M.D. (2020), "Aggregated topic models for increasing social media topic coherence", *Applied Intelligence*, Vol. 50 No. 1, pp. 138-156.

Cai, G., Sun, F. and Sha, Y. (2018), "Interactive visualization for topic model curation", *In IUI Workshops*.

Carter-White, R. (2018), *Death Camp Heritage 'from Below'? Instagram and the (Re) Mediation of Holocaust Heritage*, Edward Elgar Publishing.

Celuch, K. (2021), "Customers' experience of purchasing event tickets: mining online reviews based on topic modeling and sentiment analysis", *International Journal of Event and Festival Management*, Vol. 12 No. 1, pp. 36-50.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. and Blei, D.M. (2009), "Reading tea leaves: how humans interpret topic models", *Neural Information Processing Systems*, Vol. 22, pp. 288-296.

Chen, Y., Zhang, H., Liu, R., Ye, Z. and Lin, J. (2019), "Experimental explorations on short text topic mining between LDA and NMF based schemes", *Knowledge-Based Systems*, Vol. 163, pp. 1-13.

Cip, T., Münch, V., Pindzo, J. and Scharnagl, L. (2019), "Conceptual process model for instagram caption analysis in the case of the destination ischgl", *Proceedings of the International Student Conference in Tourism Research*, pp. 135-148.

Commane, G. and Potton, R. (2019), "Instagram and auschwitz: a critical assessment of the impact social media has on holocaust representation", *Holocaust Studies*, Vol. 25 Nos 1/2, pp. 158-181.

Dash, N.S. (2008), "Context and contextual word meaning", *Journal of Theoretical Linguistics*, Vol. 5 No. 2, pp. 21-31.

Dedeoğlu, B.B., Bilgihan, A., Ye, B.H., Wang, Y. and Okumus, F. (2021), "The role of elaboration likelihood routes in relationships between user-generated content and willingness to pay more", *Tourism Review*, Vol. 76 No. 3, pp. 614-638.

Filieri, R., Yen, D.A. and Yu, Q. (2021), "#ILoveLondon: an exploration of the declaration of love towards a destination on instagram", *Tourism Management*, Vol. 85, p. 104291.

Gaffar, V., Tjahjono, B., Abdullah, T. and Sukmayadi, V. (2021), "Like, tag and share: bolstering social media marketing to improve intention to visit a nature-based tourism destination", *Tourism Review*.

Gallagher, R.J., Reing, K., Kale, D. and Ver Steeg, G. (2017), "Anchored correlation explanation: topic modeling with minimal domain knowledge", *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 529-542.

Hannigan, T.R., Haans, R.F.J., Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S., Kaplan, S. and Jennings, P.D. (2019), "Topic modeling in management research: rendering new theory from textual data", *Academy of Management Annals*, Vol. 13 No. 2, pp. 586-632.

Irawan, H., Widyawati, R.S. and Alamsyah, A. (2020), "Identification of tourism destination preferences based on geotag feature on instagram using data analytics and topic modeling", *Understanding Digital Industry*, pp. 280-284, Routledge.

Islam, T. (2019), "Yoga-veganism: correlation mining of twitter health data", *arXiv preprint arXiv:1906.07668*.

Kakatkar, C., Bilgram, V. and Füller, J. (2020), "Innovation analytics: leveraging artificial intelligence in the innovation process", *Business Horizons*, Vol. 63 No. 2, pp. 171-181.

Kim, K., Park, O., Barr, J. and Yun, H. (2019), "Tourists' shifting perceptions of UNESCO heritage sites: lessons from jeju Island-South Korea", *Tourism Review*, Vol. 74 No. 1, pp. 20-29.

Kitchin, R. (2014), "Big data, new epistemologies and paradigm shifts", *Big Data & Society*, Vol. 1 No. 1, doi: 2053951714528481.

Köseoglu, M.A., Mehraliyev, F., Altin, M. and Okumus, F. (2020), "Competitor intelligence and analysis (CIA) model and online reviews: integrating big data text mining with network analysis for strategic analysis", *Tourism Review*,

Kujiraoka, T., Saitoh, F. and Ishizu, S. (2017), "Extraction of customer satisfaction topics regarding product delivery using non-negative matrix factorization", *In 2017 IEEE International Conference on Industrial Engineering and Engineering Management*, *IEEE*, pp. 225-229.

Leung, X.Y., Sun, J. and Bai, B. (2019), "Thematic framework of social media research: state of the art", *Tourism Review*, Vol. 74 No. 3, pp. 517-531.

Light, D. (2017), "Progress in dark tourism and thanatourism research: an uneasy relationship with heritage tourism", *Tourism Management*, Vol. 61, pp. 275-301.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., *et al.* (2018), "Applying LDA topic modeling in communication research: toward a valid and reliable methodology", *Communication Methods and Measures*, Vol. 12 Nos 2/3, pp. 93-118.

Mariani, M. (2019), "Big data and analytics in tourism and hospitality: a perspective article", *Tourism Review*, Vol. 75 No. 1, pp. 299-303.

Mariani, M., Baggio, R., Fuchs, M. and Höepken, W. (2018a), "Business intelligence and big data in hospitality and tourism: a systematic literature review", *International Journal of Contemporary Hospitality Management*, Vol. 30 No. 12, pp. 3514-3554.

Mariani, M., Di Fatta, G. and Di Felice, M. (2018b), "Understanding customer satisfaction with services by leveraging big data: the role of services attributes and consumers' cultural background", *IEEE Access*, Vol. 7, pp. 8195-8208.

Mariani, M.M. and Wamba, S.F. (2020), "Exploring how consumer goods companies innovate in the digital age: the role of big data analytics companies", *Journal of Business Research*, Vol. 121, pp. 338-352.

Martini, A. and Buda, D.M. (2020), "Dark tourism and affect: framing places of death and disaster", *Current Issues in Tourism*, Vol. 23 No. 6, pp. 679-692.

Möhring, M., Keller, B., Schmidt, R. and Dacko, S. (2021), "Google popular times: towards a better understanding of tourist customer patronage behavior", *Tourism Review*, Vol. 76 No. 3, pp. 533-569.

Munoz, C.L. and Towner, T.L. (2017), "The image is the message: instagram marketing and the 2016 presidential primary season", *Journal of Political Marketing*, Vol. 16 Nos 3/4, pp. 290-318.

Passos, A., Wallach, H.M. and McCallum, A. (2011), "Correlations and anticorrelations in lda inference", *Proceedings of the Challenges in Learning Hierarchical Models: Transfer Learning and Optimization*, pp. 1-5.

Peco-Torres, F., Polo-Peña, A.I. and Frías-Jamilena, D.M. (2021), "Brand personality in cultural tourism through social media", *Tourism Review*, Vol. 76 No. 1, pp. 164-183.

Reisenbichler, M. and Reutterer, T. (2019), "Topic modeling in marketing: recent advances and research opportunities", *Journal of Business Economics*, Vol. 89 No. 3, pp. 327-356.

Rizvi, R.F., Wang, Y., Nguyen, T., Vasilakes, J., Bian, J., He, Z. and Zhang, R. (2019), "Analyzing social media data to understand consumers' information needs on dietary supplements", *arXiv:1906.03171*.

Shawky, S., Kubacki, K., Dietrich, T. and Weaven, S. (2019), "Using social media to create engagement: a social marketing review", *Journal of Social Marketing*, Vol. 9 No. 2, pp. 204-224.

Soares, A.L.V., Mendes-Filho, L. and Gretzel, U. (2021), "Technology adoption in hotels: applying institutional theory to tourism", *Tourism Review*, Vol. 76 No. 3, pp. 669-680.

Sthapit, E., Björk, P. and Jiménez Barreto, J. (2020), "Negative memorable experience: North American and British airbnb guests' perspectives", *Tourism Review*, Vol. 76 No. 3, *Tourism Review*.

Wang, J. and Zhang, X.L. (2021), "Deep NMF topic modeling", *arXiv preprint arXiv:2102.12998*.

Yu, J. and Egger, R. (2021), "Tourist experiences at overcrowded attractions: a text analytics approach", *In Information and Communication Technologies in Tourism 2021*, (pp. 231-243). Springer, Cham.

Zhou, S., Bian, J., Zhao, Y., Haynos, A.F., Rizvi, R. and Zhang, R. (2019), "Analysis of twitter to identify topics related to eating disorder symptoms", *2019 IEEE International Conference on Healthcare Informatics*, pp. 1-4.

About the authors

Joanne Yu is a Researcher and Lecturer at the Department of Tourism and Service Management at Modul University Vienna. Her research interests focus on human-robot interaction, tourism experience, social media analytics and emerging technologies in tourism research. Joanne Yu can be contacted at: joanne.yu@modul.ac.at

Roman Egger is a Professor at the Department of Innovation and Management in Tourism, Head of eTourism, and Head of Key Competencies and Research Methods at the Salzburg University of Applied Sciences. His research interests focus on new technologies in tourism and their adoption from a user-centric perspective, as well as on methodological issues in tourism research. Roman Egger can be contacted at: roman.egger@fh-salzburg.ac.at