What does the media tell us about a destination? The Cancun case, seen from the USA, Canada, and Mexico

Elizabeth Olmos-Martínez, Miguel Á. Álvarez-Carmona, Ramón Aranda and Angel Díaz-Pacheco

Abstract

Purpose – *This study aims to present a framework for automatically collecting, cleaning and analyzing text (news articles, in this case) to provide valuable decision-making information to destination management organizations. Keeping a record of certain aspects of the projected destination image of an attraction (Cancun in this study) will grant the design of better strategies for the promotion and administration of destinations without the time-consuming effort of manually evaluating high quantities of textual information.*

Design/methodology/approach – *Using Web scraping, news articles were collected from the USA, Mexico and Canada over an interval of one year. The documents were analyzed using an automatic topic modeling method known as Latent Dirichlet Allocation and a coherence analysis to determine the number of themes present in each collection. With the data provided, the authors were able to extract valuable information to understand how Cancun is presented to the countries.*

Findings – *It was found that in all countries, Cancun is an important destination to travel and vacation; however, given the period defined for this study (from July 2021 to July 2022), an important part of the articles analyzed was concerned with the sanitary measures derived from the COVID-19 pandemic. Besides, given the rise of violence and the threat of organized crime, many articles from the three countries are focused on warning potential tourists about the risks of traveling to Cancun.*

Originality/value – *The examination of the relevant literature revealed that similar analyses are manually performed by the experts on a set of predefined categories. Although those approaches are methodologically sound, the logistic effort and the time used could become prohibitively expensive, precluding carrying out this analysis frequently. Additionally, the preestablished categories to be studied in press articles may distort the results. For these reasons, the proposed framework automatically allows for gathering valuable information for decision-making in an unbiased manner.*

Keywords *Destination image, Big data, NLP, Natural language processing*

Paper type *Research paper*

1. Introduction

Mexico is a renowned tourist destination due to its diversity, unique attractions and significant cultural heritage. Despite the pandemic, such attributes have allowed Mexico to be part of the most visited countries in the world (SECTUR, 2022a). The city of Cancun is one of the main destinations of the Mexican tourist offer, only in 2021, it received about 6,426,554 international travelers (by air), with the USA and Canada being the major countries of origin (SECTUR, 2022b). Cancun is a famous sun and beach destination known for its Mayan World Heritage ruins and was the first developed city (the 1970s) under the group Riviera Maya.

Given the importance of this industry, destination management organizations (DMOs), service providers and governments must continually assess the sector's performance to design strategies and make decisions. Surveys have traditionally been used to collect information mainly

Elizabeth Olmos-Martínez is based at the Department of Economic and Administrative Sciences, Universidad Autónoma de Occidente, Mazatlán, Mexico.
Miguel Á. Álvarez-Carmona and Ramón Aranda are both based at the The Center for Research in Mathematics, Monterrey, Mexico.
Angel Díaz-Pacheco is based at the Department of Electronics, Engineering Division, Universidad de Guanajuato – Campus Irapuato-Salamanca, Yuriria, Mexico.

focused on tourists' perceptions, opinions and satisfaction. Although the soundness of this instrument is out of the question, its logistical and economic costs sometimes make it unfeasible.

Given the availability of tourist data through social media and Web platforms, many tourism researchers have been interested in using computational techniques and methods such as big data and artificial intelligence (AI) as affordable alternatives to conducting studies. There is a plethora of examples of the use of AI methods, particularly in the field of natural language processing (NLP), to gain exciting knowledge for DMOs, such as the annual Rest-Mex evaluation (Álvarez-Carmona *et al.*, 2021; Álvarez-Carmona *et al.*, 2022a, Álvarez-Carmona *et al.*, 2022b, Álvarez-Carmona *et al.*, 2022c) forum and other relevant studies in Mexican cities (Álvarez-Carmona *et al.*, 2022a, 2022b, 2022c; Guerrero-Rodriguez *et al.*, 2021).

However, the study of the destination image construct has been poorly examined for Mexican destinations through the analysis of NLP techniques. This study is focused on knowing how Cancun is presented in news articles in the two most important countries of origin of potential visitors (the USA and Canada). The perspective presented in news articles to domestic tourists was also examined. A considerable amount of information (about 6,592 news articles) was collected and analyzed by the proposed method, which frees experts and managers from the comprehensive review of each document and the logistical and economic efforts required to conduct on-the-spot surveys.

This study considers news articles as an important factor to be considered in tourism research, either as a factor influencing tourists' decisions or in the image formation process. This is why the *main contribution* of this study is the proposal for a new technology-based methodology for analyzing textual information, open to textual data of any domain.

Our proposal presents solutions to collect automatically, clean and analyze data, providing valuable information for decision-making. Unlike other similar studies, the proposed approach offers an agnostic perspective that tries to know the main themes present in the data, avoiding bias toward what the researcher wants to find.

This work is organized as follows. Section 2 is a review of this work's most influential studies work is presented, and Section 3 describes the proposed methodology. Section 4 is the data analysis, and the discussion is in Section 5. Section 6 is for conclusions and future work.

2. Literature review

2.1 Destination image

Destination image is one of the most researched topics in tourism literature. In the past decades, there have been several studies that took a novel approach to conceptualize this construct [for a recent comprehensive overview, see Afshardoost and Eshaghi (2020)]. Despite the considerable number of existing studies, a consensus has not been reached about the concept or how to quantify it (Pike, 2002). Destination image in tourism (DIT) is a social construction that is related to a set of ideas, beliefs, or impressions that a person or group shares about a given place and is based on long-term information accumulated from different sources, leading to the construction of a positive/negative image (Choi *et al.*, 2007; Li *et al.*, 2015).

The DIT construct can be classified into different categories, such as projected and perceived images. The *projected image* can be defined as the ideas and impressions about a place that is available for the consideration of potential tourists. These ideas are presented to the tourist through secondary sources aiming to promote the destination (made by the DMOs) and by the official media (made by national tourism organizations) (Marine-Roig and Ferrer-Rosell, 2018). Depending on their type, secondary sources are classified as organic (individuals), induced (made by destination promoters) and

autonomous (independently produced) as news, travel magazines, digital information, cinema, TV series, etc. Most of this information is clearly outside the control of DMOs; however, it helps to inform potential travelers about the characteristics (physical and/or symbolic) of certain destinations (Bramwell and Rawding, 1996).

On the other hand, the *perceived images* are derived mainly from organic sources of two main types: people's perceptions shared with friends and family as well as user-generated content (UGC) available on different social media. Perceived images can be understood as the product of interaction between projected images and the characteristics of potential visitors, but there is no doubt the most influential source of information for DIT is the personal experience of visiting the tourist destination (Gartner, 1994).

A powerful aspect that has a significant impact on travelers' decisions is related to news stories. Even though there exist many different online news sources available, some of which are tabloids, readers weigh and consider all the material to some extent (Stepchenkova and Eales, 2010; Gabbioneta and Carlo, 2019; Nazir *et al.*, 2021). Moutinho (1987) pointed out already that travelers use search strategies of routine, limited, or extended information when looking for travel-related data. Perceived risk is one of the elements that influence the selection strategy of a destination. For instance, a greater number of information sources are checked, and more time is spent on finding relevant information in high-risk circumstances involving large sums of money (Sirakaya and Woodside, 2005).

The destination image of the Mexican tourist offer has been carefully examined in the past. To provide some context to the present work, some of the most influential studies in this matter are examined below.

The work of Crompton (1979) is one of the first and most influential studies on the destination image of Mexico. He wanted to examine the relationship between the descriptive image attributes of the respondents and the attributes that they considered important in any decision to visit or not visit the country for a pleasant vacation. The study found that two factors stood out in importance and were related to sanitation and safety, while the most positive factors were the climate and the low cost.

Hernández-Lobato *et al.* (2006) analyzed the destination image of Ixtapa–Zihuatanejo by surveying 140 North American travelers. They wanted to examine the relationship between the destination image, satisfaction and tourists' loyalty. They found that to assure tourists' loyalty, the infrastructures, services, natural resources and entertainment must be adapted to tourists' expectations. They also concluded that the sample must be more significant to reduce sampling error and extend the analysis to other destinations.

Esper and Rateike (2010) examined Mexico's destination image by investigating travelers from Spain. Their proposed model explained the relationships between the cognitive and affective components in the formation of the global image and the influence of the cognitive component on the affective component. Other variables showed importance, such as education level and age. For these reasons, they stressed the importance of developing a wide multisectorial tourist offer for the different segments of consumers.

Palau-Saumell *et al.* (2015) aimed to develop an integrated model of behavioral intentions regarding sun and beach destinations: Lloret de Mar and Cancun. In Cancun, they noted that the influence of country image is more significant for developing countries because it is believed that developed countries have no deficiencies regarding democracy, development, or international visibility. Economic impairments and corruption indices are perceived by tourists in Cancun, and such factors have a significant impact on the construct.

News articles are an essential factor that influences tourists' decisions. Although there is a wide range of online information media available, some of which are tabloids, all information is weighted by the reader and taken into account to some extent (Moutinho, 1987; Sirakaya and Woodside, 2005; Stepchenkova and Eales, 2010; Farina and Arslan, 2016).

According to González-Rodríguez *et al.* (2016), the choice of destination is affected by a more favorable and well-defined DIT, which comprises various subordinate factors, such as the internet media (Baloglu and McCleary, 1999; Beerli and Martín, 2004) and news articles (Lu and Atadil, 2021; Stepchenkova and Eales, 2010), among others.

Considering the above, a couple of studies were analyzed because they used news articles to examine the destination image. Even though the methodological approaches and technological instruments used in the studies vary greatly, the following articles are the most influential to this study:

Stepchenkova and Eales (2010) analyzed newspapers to test the viability of media messages to derive a quantification of the destination image. They analyzed newspaper articles from the UK containing articles about Russia, Moscow, or Petersburg. From a sample of 2,493 articles, approximately 1,280 were analyzed manually, assigning them to seven broad categories such as culture and history, economic development and social issues. The study concluded that news could be quantified as a proxy of a destination image index, and there is a correlation with the arrivals of tourists. However, they still wanted to test their approach with different sources of information.

Farina and Arslan (2016) carried out a content analysis of British newspapers related to Turkey. They examined around 1,081 news and manually classified them into six categories: economy, political, sociocultural, etc. They found that the destination image of a country is significantly affected by the tone and volume of the media coverage.

Bernkopf and Nixon (2018) wanted to determine whether visual social media can change the travel perception of a tourist destination. They surveyed three groups exposed to different sets of images (divided by their source). They found that UGC Instagram images and Google random images are more effective at improving the target image than images shared by DMOs.

Gabbioneta and Carlo (2019) analyzed the effect of newspapers in the formation process of the destination image. They performed online surveys before the destination experience of Milan, collected in the UK, France and Germany (about 2,095 questionnaires). On the other hand, data on media coverage and favorability were collected, such as the preferred newspapers and magazines of the respondents. All articles on Milan that were published a year before the survey took place were gathered and coded as positive, negative, or neutral. Their analysis found, paraphrasing the author: "the number of news articles individuals read is positively associated with their image of the destination and that their news involvement enhances the influence of the news articles on destination image formation" (Gabbioneta and Carlo, 2019).

Nazir *et al.* (2021) conducted an online survey of international tourists traveling to Pakistan to understand, among other things, how media exposure affects the image of potential travelers. It was concluded that the relationship between destination image and behavioral intention is always positive and becomes stronger when media exposure is lower.

2.2 Computational-based methods for the analysis

Tourism research with computer science methods interacting is not a brand-new subject. Several studies have used AI approaches (Díaz-Pacheco *et al.*, 2022) to examine the destination image construct, while many have used NLP techniques (Álvarez-Carmona *et al.*, 2022a, 2022b, 2022c).

Tourists' opinions of a destination were frequently studied using social media, particularly text data. The use of tweets connected to halal tourism was examined in the study by Feizollah *et al.* (2021). To identify the main topics in the corpus, they first obtained and preprocessed the data before using a combination of frequency analysis, concordance graphs and semantic network analysis. A lexical technique was then used to find emotions

in documents. They were able to determine the most significant countries (through geotagging) and the main topics (based on frequency) as a result.

Researchers Lu and Atadil (2021) studied how the COVID-19 outbreak and media coverage of it affected foreign visitors' perceptions of China. To determine the ratios of news articles that were good or negative about the location, a sample of 841 news articles from CNN were gathered and subjected to sentiment analysis. To investigate the variations in respondents' perceptions of their intention to visit China and their exposure to online media, statistical analyses were conducted. They discovered that the majority of the destination safety and security category items had lower marks than those who thought about news outlet websites. The punctuation for criteria like *cultural richness*, *ecological richness* and *wallet friendliness* was given higher punctuation for those who had less access to news outlet websites.

Before and during the COVID-19 pandemic, Nadeau *et al.* (2021) used sentiment analysis on Tweets in Toronto and New York City to examine temporal variations in the DIT. They gathered tweets from April 2019 to April 2020 since springtime travel was more popular in those cities. The data was cleaned using a variety of NLP techniques, and the tweets' Lexicon affectation strength of New York City was used to identify their sentiment. As judged from the perspective of the emotions exhibited on social media, they showed that DIT has some resilience.

A study on the Zhenjiang portion of the Grand Canal was conducted by Yang *et al.* (2022) using articles from the Chinese social network WeChat. The most crucial words were identified using frequency analysis, and topic segmentation was carried out manually by subject-matter specialists. By building associative semantic networks based on the frequency of nouns and adjectives, the cognitive and affective components of the corpus were examined. Finally, they produced a Sunburst chart using the Echarts visualization tool and an ArcGIS tool to determine the properties of the cultural heritage's spatial distribution. They discovered that DMOs were marketing shipping traffic more aggressively than the local literature and artwork.

From the above, we would like to emphasize some details that we consider relevant and which have not been addressed in the literature. Although there is a sound methodological framework, the analysis performed by the experts is expensive (funds for accommodation, travel and food, among others, of the team in charge of the survey) in time and human resources. These situations do not allow for an up-to-date record of changes to the DIT over time. On the other hand, it is common sense to think that despite expert knowledge being invaluable and irreplaceable, human beings are by nature fallible, so the exhaustive process of tracking the changes of the DIT will eventually fail.

Regarding computational approaches, based on the literature surveys and the studies mentioned above, it is indispensable to note that most of them aim to identify the tone of the analyzed content rather than the underlying themes in the collection of documents. To overcome these challenges, we propose an automated, agnostic approach to gather, clean and analyze news articles that can provide valuable knowledge for decision-making and relieve experts of such extenuating tasks.

3. Materials and methods

For an easier understanding, this section presents information on the articles collected (Section 3.1) and information on the proposed approach (Section 3.2).

3.1 Data collected (materials)

To study how Cancun is presented to foreign tourists and national tourists, we have gathered (using the Web scraping algorithm in Section 3.2.1) all the articles published by

online newspapers in Canada, the USA and Mexico between July 2021 and July 2022. Configuring the settings of the algorithm, a query was made for all the news articles about Cancun. From the request, it was retrieved 2,725 articles from the USA, 1,120 from Canada and 2,747 from Mexico (also regarding Cancun). The news portals from which the articles were obtained are presented in Appendix. Table A1 shows the news portals from Mexico, Table A2 is for the news portals from the USA and Table A3 shows those from Canada. It is important to note that the authors did not select the news portals; all news published during the defined period was collected independently of their source. This work presents the big data approach to data analysis, which preaches the importance of analyzing all the data and not just samples. In theory, having all perspectives online, regardless of reputation, will result in a representative average of all opinions.

### 3.2 Machine learning-based method for the analysis of news articles (method)

A comprehensive algorithm was developed to automatically collect, clean and analyze text from news articles. For the sake of simplicity, machine learning-based method (MLBM) comprises four components that are described in the following subsections (See Figure 1). The procedure is as follows: The algorithm collects the news articles based on defined parameters (see Section 3.2.1), then the raw information is cleaned and preprocessed to convert the data into a suitable format for the algorithms (Section 3.2.2). A topic coherence analysis is performed to determine a starting point for segmenting the corpus (lower limit) (Section 3.2.3). Finally, the corpus is segmented with the help of a thematic modeling technique, and the keywords identifying each topic (Section 3.2.4) are determined.

3.2.1 Web scraping. Web scraping is a technique that automatically extracts unstructured data from websites and transforms it into a structured format (Zhao, 2017). For this study, a script was developed in Python (a programing language), using the popular frameworks Beautiful Soup (for the parsing of HTML) and Pandas (to structure the data). This means that an *ad hoc* tool was created from scratch using a programing language and that no off-the-shelf software was used. The description of libraries and configuration parameters will allow the repeatability of the experiment within the means and budget of other researchers.

For the search, we have used Google's well-known search engine. Using the API reference documentation (Google, 2020), we configured the query parameter *cr* to limit the results to Web pages of a specific country. The parameter *gl* was used to simulate that the search was performed from within a specific country, such as Canada. A description of the entire process approaches, and advantages of Web scraping methods are out of the scope of this work. However, we refer to (Mitchell, 2018; Zhao, 2017).

3.2.2 Data cleaning and preprocessing. Despite the usefulness of the developed Web scraping algorithm, each article's data needs some preprocessing to transform it into a

Figure 1   Workflow of the proposed approach (MLBM)

Source: Author's own creation
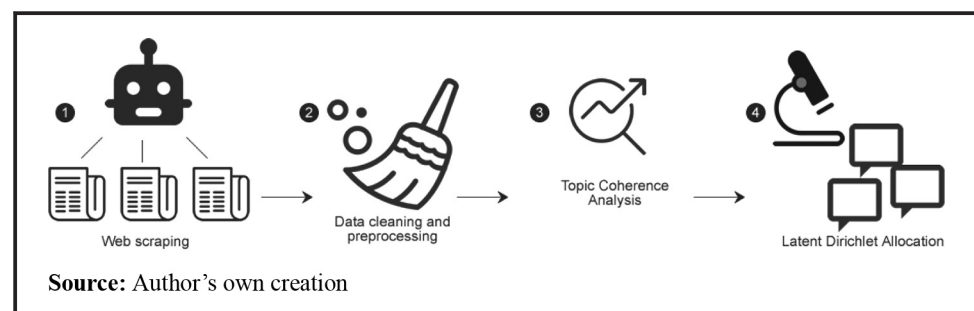
more suitable format for Latent Dirichlet Allocation (LDA). The transformations performed are as follows:

- Convert text from uppercase to lowercase.

- Elimination of *empty words*.

- Elimination of punctuation marks.

- Replacing digits with the character *d*.

- *Stemming* was used to normalize lexical variations.

- Words with fewer than ten occurrences have been deleted.

We want to clarify some concepts to provide an autonomous document as much as possible. *Empty words* (a.k.a stop words) are those words that have no significant semantic relation to the context in which they exist (Alajmi *et al.*, 2012). Since empty words do not provide meaning to a sentence, they should be deleted from the text. Al-Sughaiyer and Al-Kharashi (2004) define stemming as a method of word normalization used to match some morphologically related words. In other words, stemming algorithms gather all the words that share the same stem and have some semantic relation to reducing all variations to a single token (as far as possible).

3.2.3 Topic coherence analysis. An essential aspect of each process involving AI algorithms is the correct configuration of their hyperparameters. Hyperparameters are those parameters of a model that the model itself cannot find and that must be manually set by the user (Díaz-Pacheco and Reyes-Garcia, 2021). Regarding the topic modeling stage, the user must choose the number of topics to find in the collection of documents. This choice is far from trivial because several different themes can be present in a single document, and if the value of the hyperparameter is small, the algorithm will not find several themes, but if the number is too large, the algorithm would find absurdities. To have a starting point, we have carried out a topic coherence analysis to determine the most likely number of topics in each document collection.

A technique for assessing topic models is *topic coherence*. It gauges how semantically close the terms that received high scores are to one another. These metrics aid in separating topics that can be understood by humans from those that are the result of statistical inference. Higher coherence ratings reflect better human interpretability (Sharma *et al.*, 2020).

We have used the well-known *UMass* topic coherence measure available in the Gensim library of Python. This score measures the degree of semantic similarity between the highest-ranking words in the topic. Such quantification helps to discriminate between collections of terms semantically related and statistical noise. For a better understanding of the metric, we refer to (Cardenas *et al.*, 2018; Huang, 2019).

3.2.4 Topic modeling. Topic modeling is a statistical method used to find the underlying semantic structures in document collections of considerable size (Kherwa and Bansal, 2018). One of the best techniques to perform topic modeling is the LDA. The basic idea behind LDA is that documents are represented as stochastic mixtures over latent topics, where each topic is represented by a distribution over words (Blei *et al.*, 2003). For a better understanding, let us consider the following. LDA assumes that a document comprises several topics, each composed of a set of words, so when some words coappear with a certain probability, LDA assumes they are part of a topic. Because of the above, we use the LDA model to classify documents (or news articles) according to the topics they cover. We implemented this algorithm in Python, using the framework Gensim (Alvarez-Carmona *et al.*, 2015).

4. Data analysis

It is important to mention that for this study, no out-of-the-box informatics tools were used for the analysis of data. All the different modules for collecting and analyzing the articles were coded by the authors. However, the information presented in Section 3 is adequate to allow replicability with the informatics resources available for interested users. Although it implies the development of the scripts needed to analyze the data, the analysis of the information is automatic and the codes are reusable for new and similar tasks.

4.1 Results

The analysis with the topic modeling metric UMass obtained the following optimal values for the number of topics in each collection of articles. For the USA *3 topics*, for Canada *4* and *2 for* Mexico. Although this measure (UMass) is an excellent guide to investigating a corpus of documents, the result does not always bring the best explanation possible. In addition, the output of the LDA algorithm consists of word sets that are related according to the analysis. Using their knowledge of semantics, or common sense, the researchers should provide a name for each topic, according to the relations found among those words. Considering the above, we conducted the analysis based on the number of topics indicated by the coherence analysis. Then, we increased this number (1 unit each time), observing the changes until we found the number of topics that best explain the underlying thematic in each collection.

In other words, the Umass metric provides a lower threshold to investigate a corpus. From this minimum number, we increase the number of topics, and the resulting partitions are examined. By studying the semantic relationships of terms, subjects are evaluated. For example, the terms "dog," "dog food," "toy" and "pet" are related as terms used in a "pet shop." On the other hand, the terms "spaceship," "dog," "keyboard" and "fish" have no semantic relationship or at least not an obvious one. If all the new partitions make sense, the number of topics will increase and be re-examined. When one or more topics do not have an obvious meaning, the last partition with clear themes is chosen as the correct partition of the corpus.

In Figure 2, the topics found for the articles in Mexico are presented. The figure shows the analysis in incremental order from topics = 2 to the best partition found by the authors (4).
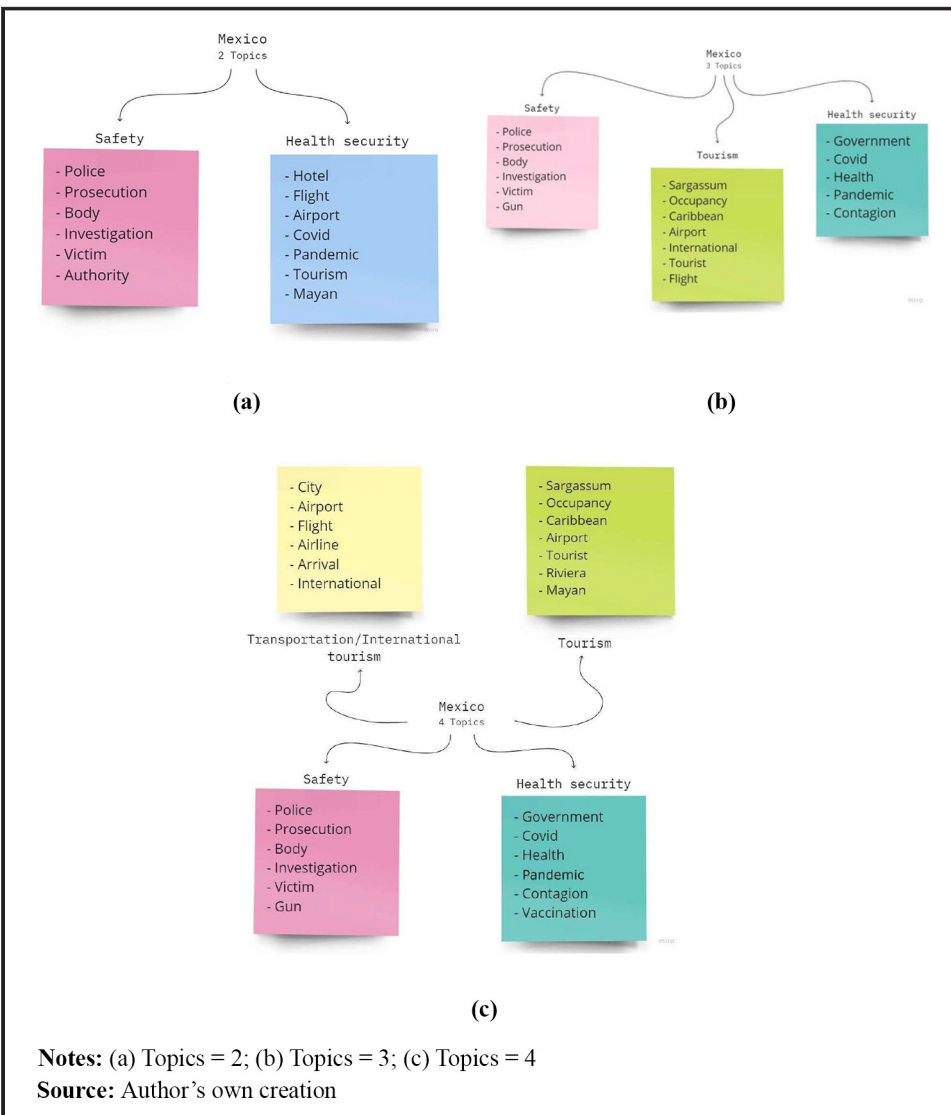
Observing the diagram of the most relevant words for the first analysis [Figure 2(a)], there are two main groups of semantic fields, one concerning individual safety, identified by the words police, body, victim, etc., and another in terms of health security (COVID-19, pandemic, contagion, etc.).

Although this seems correct, we have increased the number of topics to 3 [Figure 2(b)]. A new well-defined topic arose from this change, the one related to *tourism*. Some of the most relevant words in this topic were Sargassum (a type of algae that worsens the beauty of beaches), occupation, Caribbean, flight, etc.

We increased the number of topics to 4, and another precise topic appeared. It was named *international transport/tourism* because all their keywords share the same semantic properties related to transportation by airplanes or sea cruisers. An in-depth investigation of the news articles revealed that a substantial number of them were concerned with new air routes, economic recovery, revitalization of tourism and discount plans proposed by various airlines to promote international and domestic tourism. The main difference between the topics of tourism and international transport/tourism is that the first talks about tourist attractions, the tourism industry and travelers in general terms. International transport/tourism is related to new routes (air or sea), transport, travel policies and travel alerts.

Subsequent increments in the *Number of topics* parameter found no useful topics because not all keyword groups shared a semantic relationship; at least, for the authors, it was not obvious.
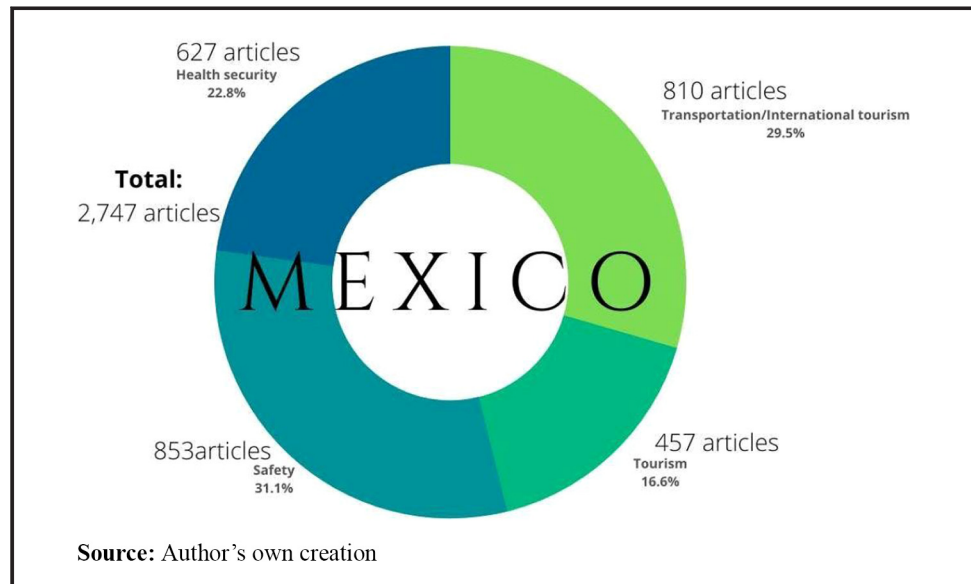
Notes: (a) Topics = 2; (b) Topics = 3; (c) Topics = 4
Source: Author's own creation

To quantify the relevance of the topics found, we classified the news articles in the predominant topics they contain. Figure 3 shows the distribution of topics in a pie chart. According to the figure, the media analyzed raises two main concerns. On the one hand, we have all the articles concerned with tourism and its economic impact divided into the classes *tourism* and *transport/international tourism*, which have 46% of the predominant topics in the articles. On the other hand, there is a significant concern about security, which is about health or about violence/crime, which contributes to 54% of the news.

Following the same analysis method, we have examined the articles from the USA and Canada. Due to space restrictions, only the best theme partition (according to the authors) is shown for each document collection. In this stage of the analysis, we have identified an important aspect to consider in this task. In both countries (the USA and Canada), we found articles written in English, French and Spanish; thus, the deletion of stop words must consider all the languages involved.

Figure 3 Distribution of the topics in the collection of news articles from Mexico regarding Cancun

627 articles
Health security
22.8%

810 articles
Transportation/International tourism
29.5%

Total:
2,747 articles

MEXICO

853 articles
Safety
31.1%

457 articles
Tourism
16.6%

Source: Author's own creation

Because of the authors' familiarity with English and Spanish, we first analyzed the corpus of documents from the USA. Figure 4 shows the final list of topics and their keywords for the USA articles (regarding Cancun).

The figure shows that we can divide the themes into two major groups *Tourism* or related to it, and *Safety*. The theme, called *Tourism*, collects documents dealing with the promotion of the destination and its tourist offer through journalistic articles. The theme of *General safety* addresses all the factors related to security, from the health aspects of the pandemic to the dangers of organized crime. Although a division should be made, as in the case of Mexico, other settings in the *number of topics* parameter did not separate them. A closer examination of the texts revealed that both topics share a common vocabulary. The above, in simpler words, means that similar expressions are used to refer to crime or COVID-19 in the press of that country, so these topics are shown as one.

The division on the topic of *travel* presents an interesting case. Although the division is mainly in terms of language, we got a deeper understanding of each group's differences by examining each group. The topic *Travel (Spanish)* addresses matters related to new air routes, health restrictions, travel policies and warnings about crime in the zone. However, it also promotes the experience in the Caribbean. On the other hand, *Travel (English)* is more concerned with warning tourists about the requirements to travel to Caribbean destinations (such as the vaccine application), mobility options in Mexican territory (such as Mexican Airlines and buses) and promoting new tourist developments (such as the Mayan train).

To understand the relevance of each topic in the collection, we made a pie chart of the documents in the collection, classified by predominant theme (Figure 5).

Once again, the themes are grouped into two principal sets *Tourism* (67.4%) and *Safety* (32.6%, with everything it entails). However, as we have seen above, travel topics are also concerned with warning tourists about sanitary precautions and travel alerts about violence in Caribbean destinations, especially Cancun. Despite this, weighing accurately every theme intertwined in that way is not a trivial task.

Finally, we have analyzed the collection of articles from Canada. An important part of the articles was written in French; all the articles in French are summarized under one

Figure 4 Distribution of the topics in the collection of news articles from the USA regarding Cancun
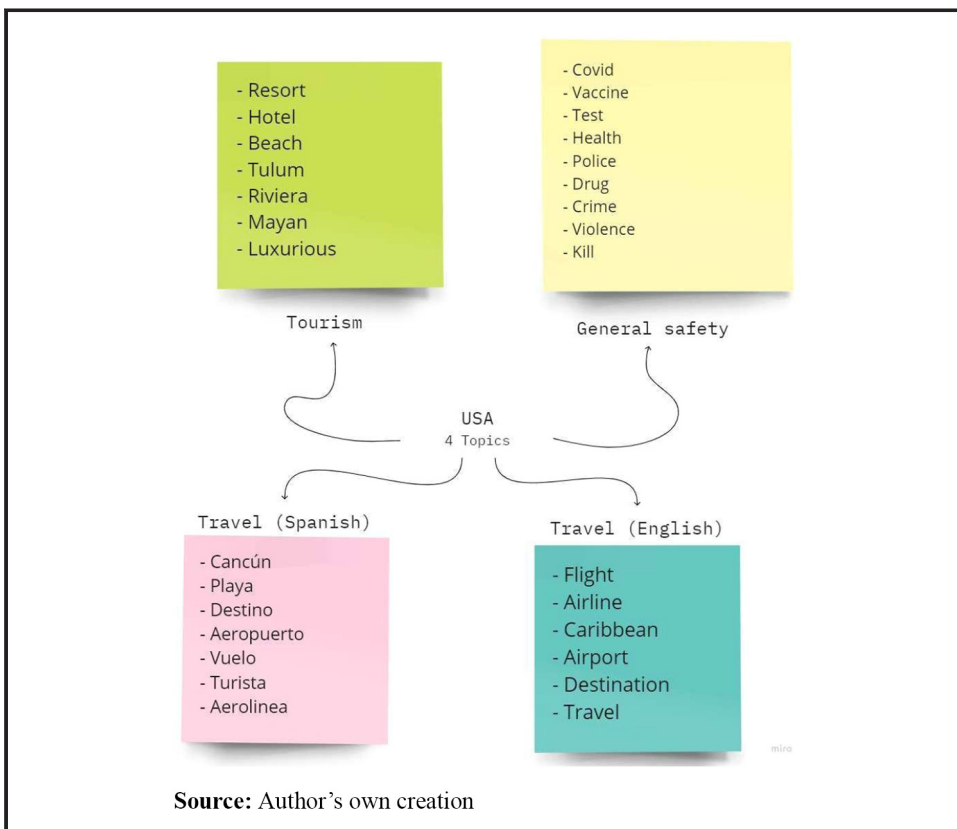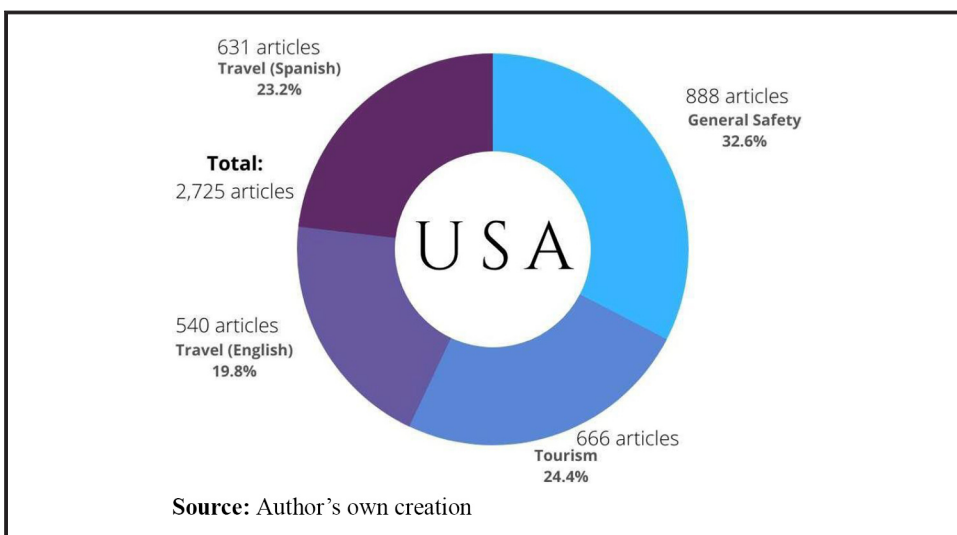
Source: Author's own creation

Figure 5 Distribution of the topics in the collection of news articles from the USA regarding Cancun

Source: Author's own creation

theme. Since this analysis had such special characteristics, we performed topic modeling in the subcollection of articles in French. Figure 6(a) shows the analysis for the complete collection, and [Figure 6(b)] shows the topics found in the subcollection in French.

As shown in [Figure 6(a)], the best partition of the thematic for Canada was of 5 topics. The topic named *Undetermined* shows an important limitation of LDA. Although it is an excellent tool for examining vast collections of documents, it is sometimes not as obvious to users and it is, therefore, difficult or even impossible to understand the collection by their

Figure 6 Diagram of topics with keywords of the entire collection (a), and for the subcollection of articles in French (b) from Canada regarding Cancun

Notes: (a) Complete collection; (b) articles in French
Source: Author's own creation
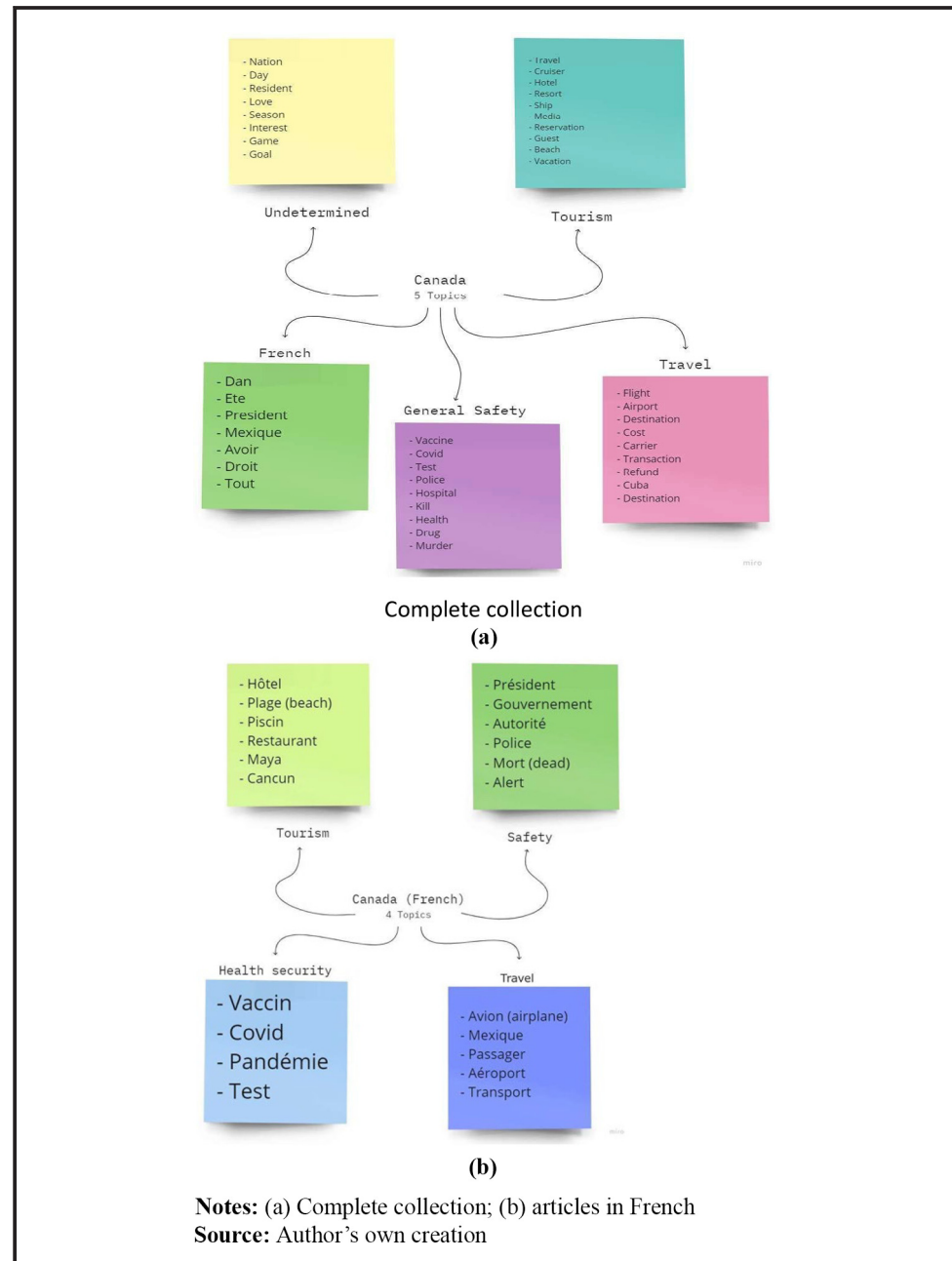
keywords. Despite the different configurations and partitions of the collection of articles on topics, no better graph was obtained. We attribute the existence of this undefined theme to the noise in the collection in the form of nonarticle-related words due to the presence of advertising on online news platforms.

From Figure 6, it may be observed that in addition to the *Undetermined* topic, the remaining themes in the two collections follow a behavior like that of other analyses. In the same way that in the collection from the USA, *Safety* is a compound theme of health and violence/crime concerns. However, the French collection of articles naturally divided health and crime. A possible explanation lies in that probably both themes share a minimal vocabulary.

It was published in French and English articles about the topic *Travel*, from the sample of articles analyzed by the authors. Articles in English reflect travel health restrictions, violence and crime warnings on different Caribbean destinations (including Cancun), travel promotions (all-inclusive packages) and connectivity delays during the holiday seasons. The French side of the topic reported on travel restrictions concerning the pandemic and new routes to enjoy warmer destinations (including Cancun).

The topic *Tourism* follows the same lines in both languages without important qualitative differences. From a closer inspection, we found the usual promotion of the destination in both languages of a collection of Caribbean destinations, with Cancun amongst them.

Since there is a mixed corpus with well-defined topics in each portion, we have presented two different pie charts (in Figure 7). It is essential to mention that the percentages in [Figure 6(b)] are calculated on the 444 articles in French.

From the chart [Figure 7(a)], we would like to point out that the presence of different languages is easily identified by the proportion that the *French* category has drawn from the graph (almost 40%). However, the category *Undetermined* also took an important number of documents from the collection (21%). Although further examinations were carried out only in that portion of the collection, no specific topic (at least for the authors) came up. Therefore, the preprocessing techniques should be improved to eliminate as much noise as possible.

The topic *Safety* took an important part of the entire collection (44.73%), considering the union of *General Safety* in combination with *Security* and *Health Security*.

Looking also at *Tourism* as a combination of *travel* and *tourism* from both collections (French and English), we get 34.5%. Considering this broad division, data shows that given the conditions of the evaluated period, *safety* is an important thematic present in the media of all the analyzed countries. However, based on the thematic discovered, the tourist industry is investing and putting all its effort into changing that panorama.

Given the topics identified among the collections, they can be classified as articles with positive connotations and articles with negative connotations. A positive topic is one focused on sharing positive information about the destination, while a negative topic has no interest in promoting the destination or at least does not encourage people to travel to that site. For this analysis stage, tourism-related topics such as "tourism," "travel" and "international tourism" made up the positive class. The undetermined topic was excluded since it could be classified in none of the classes. The union of "safety" and "health security" creates a negative class.

Although this division may seem arbitrary, we analyzed a sample of articles categorized as "safety" or "health security" in all countries. Below are some of the general ideas in the articles from each country.

Articles from Mexico:

- Increased contagion in Cancun and the region.
- Articles in French

Figure 7 Distribution of the topics in the collection of news articles from Canada regarding Cancun

113 articles
Travel
10.1%

236 articles
Undetermined
21.1%

196 articles
General safety
17.5%

Total:
1,120 articles

CANADA

131 articles
Tourism
11.7%

French
39.6%
444 articles

(a)

58 articles
Travel
13.1%

85 articles
Tourism
19.1%

Total:
444 articles

CANADA
(FRENCH)

173 articles
Health security
39%

128 articles
Safety
28.8%

(b)

Notes: (a) Complete collection; (b) articles in French
Source: Author's own creation

■ Suspension of cultural and mass events.

■ Instructions for the process of vaccination.

■ Sanitary measures for hotels, transport and supermarkets include the use of facial masks and social distancing.

Articles from the USA:

■ Requirements to travel abroad.

■ Travel warnings about crime and COVID-19.

■ Articles about crime, violence and contagions.

■ Sanitary restrictions to travel back to the USA.

Articles from Canada (English):

- Articles about crime, violence and contagions.
- Requirements to travel abroad.
- Vaccination status in Cancun and the Caribbean.
- Sanitary restrictions to travel back to Canada.

Articles from Canada (French):

- Tips for travel during pandemics.

From the above, the news articles from all countries do not encourage tourists to go to Cancun; instead, they are reminders of all the threats, therefore, were classified in the negative class.

According to the proposed classification, Mexico is the country with the most articles in the negative class, with 1,480, compared to 888 for the USA and 497 for Canada (joining topics in English and French). On the other hand, the USA was the country with more articles in the positive class (1,837), which means that despite the security issues, the American people are interested in warm places to enjoy their holidays, and Cancun is an excellent alternative for them. Unlike the remaining countries, Canada has an important portion of unclassified articles ("undefined" topic). If these articles are removed from the classification, most articles from Canada are in the negative class (497), which represents 56.22% of the total. As a result, much of Canada's news coverage warns travelers about crime and health problems in Cancun, Mexico.

The above shows a significant trend in security, projected by a big part of online news portals. Although this does not reflect the opinion of travelers, the projected information can influence Cancun's destination image for many potential travelers. The security problem in Mexico, and particularly in Cancun, is not just a bad promotion management issue; it is an important concern that is far from being solved. According to information from the INEGI (National Institute of Statistics and Geography) in 2021, Quintana Roo (the Mexican state where Cancun is located) was the 10th state with the most criminal activity, and both in 2020 and 2021, its crime rate exceeded the national rate (INEGI, 2021).

Regarding the COVID-19 pandemic, data from the Mexican Government show that Quintana Roo was the 22nd out of 32 states by confirmed cases (110,483) CONACyT (2022). Despite this being a considerably high number, it is important to mention that being one of the most important tourist entities in Mexico, the number of positive cases could have been higher, and this is an important success for the DMOs and local governments.

The data showed that these major deficiencies have not gone unnoticed by foreign travelers who have been warned by a wide range of online means. For locals, this is a painful reality that is constantly denounced in news articles and social media. While COVID-19 is decreasing, it is important to note that, at least for Cancun, these health concerns have somehow been addressed. On the other hand, crime needs a joint effort by the government and the DMOs to take strong action to reduce those rates and develop massive marketing campaigns to reshape Cancun's projected destination image.

5. Discussion

The results showed a clear pattern of concern in the three countries regarding crime, violence and the pandemic in Cancun. This concern is not new among North American travelers, as was pointed out by Crompton in the late 1970s. Despite this, there were a significant number of articles that supported or promoted Cancun. Cancun is one of the most developed destinations in the Mexican tourism offering, with considerable experience

in promoting to a wide audience. The adaptation of the offer to ensure tourist loyalty has been pointed out as a major factor in the studies of Hernández-Lobato et al. (2006) and Esper and Rateike (2010).

Statistics of hotel occupancy percentages in Cancun during the time interval examined in this study were obtained from the Mexican Secretary of Tourism (see Table 1). A T-test was used to compare the media from that period to a similar prepandemic interval. Both distributions were normal in a Shapiro–Wilk test, and no significant differences were found between the averages ($p = 0.1839$). Considering that, there is evidence that there were no important repercussions from the negative portion of news articles.

According to the study by Gabbioneta and Carlo (2019), the number of positive articles read by a tourist positively influences DIT's expectations, which could explain the hotel occupancy rates.

In contrast to studies such as those of Stepchenkova and Eales (2010), Lu and Atadil (2021), Farina and Arslan (2016) and Nazir et al. (2021), the proposed method frees experts from the tedious task of collecting, sifting and manually analyzing a large collection of news articles. In addition, the method provides a means of knowing (generally speaking) what is mentioned about a destination without predisposing what is to be found.

Our contribution has profound *practical implications* in the research of urban tourism. Since this type of destination attracts great attention from the media, DMOs need to pay close attention to the projected image to design strategies to project the aspects they want to emphasize and mitigate the effect of those perceived as defects. Since this involves a significant investment of resources, the use of automated approaches can provide the means to survey a bigger sample of the analyzed material and enable a periodical examination. Since the proposed approach does not presume the existence of topics among the documents analyzed, the information provided is unbiased and ready to collect valuable information for the experts. The data analysis results and the subsequent statistical analysis showed that the proposed methodology examined a considerable collection of documents, presenting useful information for decision-making. Since the proposed method has been developed to analyze destinations of all typologies, it can be applied to explore destinations as diverse as Cancun, Rome, or Aspen without performing profound modifications.

6. Conclusions and future work

The study of the destination image based on the analysis of NLP techniques of the city of Cancun from the main news topics read by potential visitors from the EE.UU. and Canada are closely related to tourism and international transportation/tourism since they account for 46% of the predominant topics in the articles read. At the same time, there is a concern for security related to health or violence/crime, with more than 50% of the predominant themes. It is assumed that the themes read in the articles directly influence the decisions of future tourists to visit the destination since the news (external communication) is part of the process of image formation based on the senses in an induced, complex and stimulating way. Since the image of the destination is formed before, during and after the visit from a combination of positive or negative perceptions that represent reality, the results contribute

Table 1	Hotel occupancy percentages in Cancun during the period examined and a similar prepandemic period retrieved from (INEGI, 2022)										
Interval/month	July (%)	Aug. (%)	Sept. (%)	Oct. (%)	Nov. (%)	Dec. (%)	Jan. (%)	Feb. (%)	Mar. (%)	Apr. (%)	May (%)	June (%)
2018–2019	77.8	76.0	58.9	58.9	69.4	71.7	78.2	81.4	80.2	73.8	70.5	73.8
2021–2022	69.0	58.9	48.1	52.8	71.7	75.7	68.0	71.7	81.3	77.8	77.0	79.3

Source: Author's own creation

to more accurate decision-making by DMOs, given the need for information for the design of strategies aimed at tourism services related to the periodic evaluation of the image of the tourist destination and visitor expectations.

On the other hand, recalling that the main contribution of this study is the proposal of a new technology-based methodology for the analysis of textual information, open to textual data from any domain. It is also concluded that the source of tourism data for decision-making is versatile and dynamic. The use of NLP leads to robust results from sources available on various digital platforms. It is highlighted that such results point out precise information where from a semantic relation of the most repeated words in the news, relevant topics and semantic fields are discovered where the DIT is formed under the expectation of future visitors or tourists. This information can be analyzed and valued by decision-makers. It is worth mentioning that studies which analyze information based on specific themes or categories in response to an observed problem are not ignored but rather point out that these studies can be complemented with NLP techniques to have a broader view of the tourism industry. In several studies, DIT-related categories have been proposed for the classification of news articles. Although this is due to solid tourism theories in DIT, it cannot be ensured that a topic of interest will be part of news articles such as "Tourist Infrastructure." Given the above, this study proposes an automated approach based on robust NLP techniques applied to big data collection. The proposed method will be able to automatically track the changes in the DIT, which will help the experts and DMOs define strategies and make decisions. Although the proposed method is not error-free, its ability to deal with larger data quantities will provide more representative samples and mitigate the effect of such disturbances.

Given the authors' experience with text analysis using NLP techniques, several opportunity areas and limitations were identified in this work, so the following elements will be included in future work:

Although the information collected by the Web-scraping script is very useful for obtaining an adequate source of data to analyze, an intelligent preprocessing step could improve the quality of the information by filtering the data (using AI methods) to keep just the information related to the destination, leaving out advertisements and similar noise.

This study followed big data's data analysis perspective and analyzed all the information collected to obtain the entire landscape presented by the news portals. However, filtering information to retain only those articles from the news portals that were in the interest of the analyst could be useful for decision-makers and tourism researchers.

Since the output of a topic modeling task is the segmentation of the set of documents and a group of representative words from those groups, naming the topic lies with the analyst. This task is far from trivial, and the semantic relations among the terms are not always obvious. Therefore, it is necessary to include an AI method to analyze such semantic relationships to name the topics.

The results show the image of the tourist destination constructed from external and induced communication, and it is recommended to make a comparative analysis of the journalistic publications read by potential visitors versus the advertising or marketing of the DMOs or government sector for the promotion of a tourist destination based on a classification of the stimulus variables.

To strengthen the differentiated decision-making of DMOs and governments, it is necessary to carry out an analysis of external stimuli (news) for the construction of the image destination differentiated by time of the year that allows the comparison of the temporal themes that influence this image.

References

Afshardoost, M. & Eshaghi, M. S. (2020). Destination image and tourist behavioural intentions: A meta-analysis. *Tourism Management*, *81*, 104154, doi: https://doi.org/10.1016/j.tourman.2020.104154.

Alajmi, A., Saad, E. M., & Darwish, R. R. (2012). Article: toward an Arabic stop-words list generation. *International Journal of Computer Applications*, *46*(8), 8–13. Retrieved from www.ijcaonline.org/archives/volume46/number8/6926-9341 (Full text available).

Al-Sughaiyer, I. A. & Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, *55*(3), 189–213, doi: https://doi.org/10.1002/asi.10368.

Álvarez-Carmona, M. Á., Aranda, R., Arce-Cardenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A. P., . . . Rodríguez-González, A. Y. (2021). Overview of rest-mex at iberlef 2021: Recommendation system for text Mexican tourism. Proce- samiento del Lenguaje Natural. Retrieved from http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6386

Álvarez-Carmona, M. A., Aranda, R., Rodríguez-González, A. Y., Pellegrin, L., & Car-los, H. (2022a). Classifying the Mexican epidemiological semaphore colour from the covid-19 text Spanish news. *Journal of Information Science*, (16555152211009), doi: https://doi.org/10.1177/01655515221100952.

Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-Gonzalez, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., . . . Díaz-Pacheco, Á. (2022b). Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University – Computer and Information Sciences*, *34*(10), doi: https://doi.org/10.1016/j.jksuci.2022.10.010.

Álvarez-Carmona, M. Á., Díaz-Pacheco, Á., Aranda, R., Rodríguez-González, A. Y., Fajardo-Delgado, D., Guerrero-Rodríguez, R., & Bustio-Martínez, L. (2022c). Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for Mexican tourist texts. *Procesamiento Del Lenguaje Natural*, *69*, 289–299.

Alvarez-Carmona, M. A., López-Monroy, A. P., Montes-y-Gómez, M., Villasenor-Pineda, L., & Jair-Escalante, H. (2015). INAOE's participation at PAN'15: Author profiling task. Working Notes Papers of the CLEF, 103.

Baloglu, S. & McCleary, K. W. (1999). A model of destination image formation. *Annals of Tourism Research*, *26*(4), 868–897, doi: https://doi.org/10.1016/s0160-7383(99)00030-4.

Beerli, A. & Martín, J. D. (2004). Factors influencing destination image. *Annals of Tourism Research*, *31*(3), 657–681, doi: https://doi.org/10.1016/j.annals.2004.01.010.

Bernkopf, D. & Nixon, L. (2018). The impact of visual social media on the projected image of a destination: The case of Mexico City on Instagram. *In Information and Communication Technologies in Tourism 2019*, (145–157), doi: https://doi.org/10.1007/978-3-030-05940-8 12.

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

Bramwell, B. & Rawding, L. (1996). Tourism marketing images of industrial cities. *Annals of Tourism Research*, *23*(1), 201–221, doi: https://doi.org/10.1016/0160-7383(95)00061-5.

Cardenas, R., Bello, K., Coronado, A., & Villota, E. (2018). Improving topic coherence using entity extraction denoising. *The Prague Bulletin of Mathematical Linguistics*, *110*(1), 85–101, doi: https://doi.org/10.2478/pralin-2018-0004.

Choi, S., Lehto, X. Y., & Morrison, A. M. (2007). Destination image representation on the web: Content analysis of Macau travel related websites. *Tourism Management*, *28*(1), 118–129.

CONACyT. (2022). COVID-19 tablero méxico. Retrieved from https://datos.covid-19.conacyt.mx/#DownZCSV (accessed 22. November 2022).

Crompton, J. L. (1979). An assessment of the image of Mexico as a vacation destination and the influence of geographical location upon that image. *Journal of Travel Research*, *17*(4), 18–23, doi: https://doi.org/10.1177/004728757901700404.

Díaz-Pacheco, A. & Reyes-Garcia, C. A. (2021). A classification-based fuzzy-rules proxy model to assist in the full model selection problem in high volume datasets. *Journal of Experimental & Theoretical Artificial Intelligence*, *34*(5), 1–30, doi: https://doi.org/10.1080/0952813x.2021.1925972.

Díaz-Pacheco, A., Álvarez-Carmona, M. Á., Guerrero-Rodríguez, R., Chávez, L. A. C., Rodríguez-González, A. Y., Ramírez-Silva, J. P., & Aranda, R. (2022). Artificial intelligence methods to support the

research of destination image in tourism. a systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 1–31, doi: https://doi.org/10.1080/0952813x.2022.2153276.

Esper, F. S. & Rateike, J. Á. (2010). Tourism destination image and motivations: The Spanish perspective of Mexico. *Journal of Travel & Tourism Marketing*, *27*(4), 349–360, doi: https://doi.org/10.1080/10548408.2010.481567.

Farina, I A. & Arslan, B. (2016). Managing country brand image: a content analysis of the UK newspapers on the image of Turkey. *Pazarlama ve Pazarlama Ara,Stırmaları Dergisi*, *9*(17), 65–86. Retrieved from https://dergipark.org.tr/en/pub/ppad/issue/61004/906059

Feizollah, A., Mostafa, M. M., Sulaiman, A., Zakaria, Z., & Firdaus, A. (2021). Exploring halal tourism tweets on social media. *Journal of Big Data*, *8*(1), doi: https://doi.org/10.1186/s40537-021-00463-5.

Gabbioneta, C. & Carlo, M. D. (2019). The role of news articles, prior destination experience, and news involvement in destination image formation. *International Journal of Tourism Research*, *21*(3), 291–301, doi: https://doi.org/10.1002/jtr.2251.

Gartner, W. C. (1994). Image formation process. *Journal of Travel & Tourism Marketing*, *2*(2/3), 191–216, doi: https://doi.org/10.1300/j073v02n02\12.

González-Rodríguez, M. R., Martínez-Torres, R., & Toral, S. (2016). Post-visit and pre-visit tourist destination image through eWOM sentiment analysis and perceived helpfulness. *International Journal of Contemporary Hospitality Management*, *28*(11), 2609–2627, doi: https://doi.org/10.1108/ijchm-02-2015-0057.

Google. (2020). XML API reference appendices | programmable search engine | Google developers. Retrieved from https://developers.google.com/custom-search/docs/xml\results\appendices (accessed 25 August 2022).

Guerrero-Rodriguez, R., Álvarez-Carmona, M. Á., Aranda, R., & López-Monroy, A. P. (2021). Studying online travel reviews related to tourist attractions using NLP methods: the case of Guanajuato, Mexico. *Current Issues in Tourism*, 1–16. doi: https://doi.org/10.1080/13683500.2021.2007227.

Hernández-Lobato, L., Solis-Radilla, M. M., Moliner-Tena, M. A., & Sánchez-García, J. (2006). Tourism destination image, satisfaction and loyalty: A study in ixtapa-zihuatanejo, Mexico. *Tourism Geographies*, *8*(4), 343–358, doi: https://doi.org/10.1080/14616680600922039.

Huang, C.-M. (2019). Incorporating prior knowledge by selective context features to enhance topic coherence. *Communications in computer and information science*, pp. 310–318. Springer Singapore. doi: https://doi.org/10.1007/978-981-13-9190-3\32.

INEGI. (2021). Incidencia delictiva. Instituto Nacional de Estadística y Geografía. INEGI. Retrieved from www.inegi.org.mx/temas/incidencia

INEGI. (2022). Datatur3 – Tableros. Retrieved from https://datatur.sectur.gob.mx/SitePages/Tableros.aspx (accessed 24 November 2022).

Kherwa, P. & Bansal, P. (2018). Topic modeling: A comprehensive review. *ICST Transactions on Scalable Information Systems*, *0*(0), 159623, doi: https://doi.org/10.4108/eai.13-7-2018.159623.

Li, J., Ali, F., & Kim, W. (2015). Reexamination of the role of destination image in tourism: An updated literature review. *e-Review of Tourism Research*, *12*, 191–209. Retrieved from http://agrilife.org/ertr/files/2015/09/4.-eRTR\ARN\Vol.12-No3.4\Li-Faizan-and-Kim.pdf

Lu, Q. & Atadil, H. A. (2021). Do you dare to travel to China? an examination of China's destination image amid the covid-19. *Tourism Management Perspectives*, *40*, 100881, doi: https://doi.org/10.1016/j.tmp.2021.100881.

Marine-Roig, E. & Ferrer-Rosell, B. (2018). Measuring the gap between projected and perceived destination images of Catalonia using compositional analysis. *Tourism Management*, *68*, 236–249, doi: https://doi.org/10.1016/j.tourman.2018.03.020.

Mitchell, R. (2018). Web scraping with Python: collecting more data from the modern web. O'Reilly Media. Retrieved from https://books.google.de/books?id=TYtSDwAAQBAJ

Moutinho, L. (1987). Consumer behaviour in tourism. *European Journal of Marketing*, *21*(10), 5–44, doi: https://doi.org/10.1108/eum0000000004718.

Nadeau, J., Wardley, L. J., & Rajabi, E. (2021). Tourism destination image resiliency during a pandemic as portrayed through emotions on Twitter. *Tourism and Hospitality Research*, *22*(1), 60–70, doi: https://doi.org/10.1177/14673584211038317.

Nazir, M. U., Yasin, I., Tat, H. H., Khalique, M., & Mehmood, S. A. (2021), The influence of international tourists' destination image of Pakistan on behavioral intention: The roles of travel experience and media exposure. *International Journal of Hospitality & Tourism Administration*, *23*(6), 1–25, doi: https://doi.org/10.1080/15256480.2021.1938782.

Palau-Saumell, R., Forgas-Coll, S., Amaya-Molinar, C. M., & Sánchez-García, J. (2015). Examining how country image influences destination image in a behavioral intentions model: The cases of Lloret de mar (Spain) and Cancun (Mexico). *Journal of Travel & Tourism Marketing*, *33*(7), 949–965, doi: https://doi.org/10.1080/10548408.2015.1075456.

Pike, S. (2002). Destination image analysis—a review of 142 papers from 1973 to 2000. *23*(5), 541–549, doi: https://doi.org/10.1016/s0261-5177(02)00005-5.

SECTUR. (2022a). Datatur3 – rankingOMT. Retrieved from www.datatur.sectur.gob.mx/SitePages/RankingOMT.aspx (accessed 24 August 2022).

SECTUR. (2022b). Resultados de la actividad turística diciembre 2021. Retrieved from www.datatur.sectur.gob.mx/RAT/RAT-2021-12(ES).pdf (accessed 24 August 2022).

Sharma, S. K., Dwivedi, Y. K., Metri, B., & Rana, N. P. (2020). Re-imagining diffusion and adoption of information technology and systems: a continuing conversation: IFIP WG 8.6 international conference on transfer and diffusion of IT. *Tdit 2020*, *Tiruchirappalli, India*: Springer Nature. December 18–19, 2020, proceedings, part II. Cham, Switzerland.

Sirakaya, E. & Woodside, A. G. (2005). Building and testing theories of decision making by travellers. *Tourism Management*, *26*(6), 815–832, doi: https://doi.org/10.1016/j.tourman.2004.05.004.

Stepchenkova, S. & Eales, J. S. (2010). Destination image as quantified media messages: The effect of news on tourism demand. *Journal of Travel Research*, *50*(2), 198–212, doi: https://doi.org/10.1177/0047287510362780.

Yang, Y., Sha, C., Su, W., & Donkor, E. K. N. (2022). Research on online destination image of Zhenjiang section of the grand canal based on network content analysis. *Sustainability*, *14*(5), 2731, doi: https://doi.org/10.3390/su14052731.

Zhao, B. (2017). Web scraping. *Encyclopedia of Big Data*, Springer International Publishing. 1–3, doi: https://doi.org/10.1007/978-3-319-32001-4\483-1.

Appendix

Table A1	The 169 sites from which the news articles from Mexico were collected		
www.poresto.net	sipse.com	www.reportur.com	lucesdelsiglo.com
www.lajornadamaya.com	cancun.gob.mx	www.eleconomista.com.mx	www.milenio.com
radioformulaqr.com	www.elsoldemexico.com.mx	www.informador.mx	www.yucatan.com.mx
www.forbes.com.mx	www.debate.com.mx	www.eluniversal.com.mx	www.elpuntosobrelai.com
a21.com.mx	quintanaroo.quadratin.com.mx	www.proceso.com.mx	aristeguinoticias.com
www.excelsior.com.mx	www.24-horas.mx	www.elfinanciero.com.mx	www.marca.com
caribempresarial.com	www.jornada.com.mx	politica.expansion.mx	expansion.mx
www.reporteindigo.com	lasillarota.com	www.unotv.com	elpais.com
mexicotravelchannel.com.mx	www.record.com.mx	www.yucatanalamano.com	heraldodemexico.com.mx
www.publimetro.com.mx	realestatemarket.com.mx	www.animalpolitico.com	noticieros.televisa.com
noticias.canal10.tv	www.razon.com.mx	www.vavel.com	www.sdpnoticias.com
www.mexicodesconocido.com.mx	www.espn.com.mx	t21.com.mx	almomento.mx
www.transponder1200.com	www.radioformula.com.mx	www.elsiglodetorreon.com.mx	inmobiliare.com
www.tvazteca.com	yucatanahora.com	vamoscruzazul.bolavip.com	www.viveusa.mx
amnistia.org.mx	businessinsider.mx	lopezdoriga.com	obras.expansion.mx
robbreport.mx	www.elsoldepuebla.com.mx	mexico.as.com	www.tyt.com.mx
www.palcoquintanarroense.com.mx	wradio.com.mx	lopezobrador.org.mx	mvsnoticias.com
elcomentario.ucol.mx	www.futboltotal.com.mx	www.la-prensa.com.mx	www.elsoldetampico.com.mx
regeneracion.mx	www.revistacentral.com.mx	www.cronica.com.mx	www.etcetera.com.mx
www.alcaldesdemexico.com	www.lacapital.com.mx	amqueretaro.com	centralelectoral.ine.mx
rotativo.com.mx	chivaspasion.bolavip.com	www.oinkoink.com.mx	www.diariodemexico.com
www.altonivel.com.mx	www.merca20.com	mexico.quadratin.com.mx	www.vivaaerobus.com
tiempo.com.mx	www.sopitas.com	www.am.com.mx	www.sinembargo.mx
www.masnoticias.mx	www.yotambien.mx	los40.com.mx	www.gq.com.mx
hotbook.mx	plumaslibres.com.mx	www.eldictamen.com	sanluis.eluniversal.com.mx
elle.mx	leonesnegrosudg.mx	elcapitalino.mx	m.arenapublica.com
www.diariodequeretaro.com.mx	contralinea.com.mx	thehappening.com	piedepagina.mx
imagendeveracruz.mx	www.jorgecastronoriega.com	www.tudn.mx	energiahoy.com
www.meteored.mx	www.elsoldezacatecas.com.mx	oaxaca.eluniversal.com.mx	www.hidrocalidodigital.com
saludcolima.gob.mx	guanajuato.lasillarota.com	foodandtravel.mx	www.glamour.mx
lifeandstyle.expansion.mx	www.leones.mx	www.sensacine.com.mx	www.elsoldelcentro.com.mx
m.imagenradio.com.mx	americamonumental.bolavip.com	www.bigfish.mx	www.elimparcial.com
www.ejecentral.com.mx	www.meganews.mx	www.nytimes.com	tribunadeloscabos.com.mx
www.elheraldodesaltillo.mx	www.quien.com	www.tribuna.com.mx	plazadearmas.com.mx
criteriohidalgo.com	www.dcubanos.com	www.elfutbolero.com.mx	www.larevista.com.mx
todup.news	diariodechiapas.com	puebla.gob.mx	www.poblanerias.com
turquesanews.mx	www.primeralinea.mx	tribunadelabahia.com.mx	news.culturacolectiva.com
laverdadnoticias.com	www.elsudcaliforniano.com.mx	vanguardia.com.mx	www.lavozdemichoacan.com.mx
intoleranciadiario.com	mx.hola.com	www.adn40.mx	www.elsoldeleon.com.mx
diario.mx	www.eldiariodechihuahua.mx	www.endondecorrer.com	www.e-veracruz.mx
mexicorutamagica.mx	24horaspuebla.com	instyle.mx	www.angulo7.com.mx
mexicoaeroespacial.com.mx	www.digitalmex.mx	www.elsiglodedurango.com.mx	www.diariocambio.com.mx
www.capitalmexico.com.mx			

Source: Author's own creation

Table A2	The first 200 sites (from 729) of which the US articles were collected. The selection was based on the number of items obtained from the site		

www.caribjournal.com	www.insider.com	entretenimento.r7.com	www.popsugar.com
www.travelpulse.com	people.com	www.si.com	www.borderreport.com
www.traveloffpath.com	www.periodicocubano.com	english.elpais.com	www.billboard.com
www.aviacionline.com	la-lista.com	www.youredm.com	musicfestnews.com
www.infobae.com	www.newsweek.com	www.chicagotribune.com	oncubanews.com
www.theyucatantimes.com	www.fodors.com	www.liderempresarial.com	www.insideedition.com
mexiconewsdaily.com	aeroin.net	cultmtl.com	www.costar.com
simpleflying.com	www.nytimes.com	www.mtlblog.com	pressperu.com
www.soynomada.news	www.prnewswire.com	www.independentespanol.com	thehill.com
thepointsguy.com	es-us.noticias.yahoo.com	www.americanbazaaronline.com	musicrow.com
www.hospitalitynet.org	en.mercopress.com	themusicuniverse.com	www.detroitnews.com
airlinegeeks.com	liveforlivemusic.com	www.narcity.com	www.sacurrent.com
www.reforma.com	www.thedailybeast.com	www.star-telegram.com	revistaespejo.com
www.edmtunes.com	www.laprensalatina.com	www.digitaljournal.com	noticiasya.com
www.riviera-maya-news.com	www.cbsnews.com	www.timesonline.com	www.abc4.com
www.washingtonpost.com	fadeawayworld.net	deadline.com	www.financecolombia.com
edm.com	www.houstonchronicle.com	montrealgazette.com	www.ellitoral.com
www.forbes.com	www.rayados.com	volleyballmag.com	www.khou.com
www.latimes.com	www.vallartadaily.com	www.distractify.com	retro1025.com
www.dallasnews.com	vanguardiaveracruz.com	www.nbcnews.com	www.bostonglobe.com
www.thetravel.com	www.expressnews.com	www.euronews.com	www.crawfishboxes.com
www.directoriocubano.info	www.theguardian.com	www.themanual.com	www.homosensual.com
www.travelweekly.com	news.yahoo.com	mixmag.net	www.firstcoastnews.com
latinus.us	www.univision.com	www.southbendtribune.com	www.kilgorenewsherald.com
www.reuters.com	www.informadorqr.com	www.channelfutures.com	www.click2houston.com
periodicoviaje.com	www.soyfutbol.com	www.breakingbelizenews.com	www.ladevi.info
www.businesswire.com	travelnoire.com	www.telemundo.com	www.thedrive.com
www.efe.com	dancingastronaut.com	liveandletsfly.com	runwaygirlnetwork.com
www.mercurynews.com	www.travelmarketreport.com	www.tmz.com	www.nbcsports.com
jambands.com	www.guatemala.com	www.primerahora.com	conocedores.com
www.usatoday.com	cubanosporelmundo.com	www.nbcdfw.com	patch.com
abcnews.go.com	edmidentity.com	www.khq.com	passageirodeprimeira.com
www.milb.com	www.businesstraveller.com	wifinowglobal.com	quepasamedia.com
www.travelandleisure.com	www.staradvertiser.com	www.newyorker.com	calgaryherald.com
www.cibercuba.com	www.luxurytravelmagazine.com	www.arizonafoothillsmagazine.com	gatopardo.com
nypost.com	www.elnorte.com	www.cheatsheet.com	fightnews.com
www.prweb.com	www.insidehook.com	goblueraiders.com	datebook.sfchronicle.com
www.thrillist.com	www.cnbc.com	mymmanews.com	noticiasenlamira.com
www.travelawaits.com	www.chron.com	www.brooklynvegan.com	tasteofcountry.com
www.lonelyplanet.com	www.pasionaguila.com	m.vanguardia.com	www.vogue.com
www.mediotiempo.com	www.xevt.com	www.yardbarker.com	www.stereogum.com
www.bloomberg.com	loyaltylobby.com	blueskypit.com	www.travelmole.com
www.businessinsider.com	volavi.co	www.kxan.com	www.metroweekly.com
www.the-sun.com	www.jambase.com	www.gratefulweb.com	centreforaviation.com
www.foxnews.com	www.rollingstone.com	elceo.com	www.nj.com
eldiariony.com	finance.yahoo.com	www.bizjournals.com	www.lofficielusa.com
gomastodons.com	www.vice.com	www.huffpost.com	peopleenespanol.com
www.traveldailynews.com	usa.inquirer.net	www.peta.org	communityimpact.com
www.yahoo.com	es.travel2latam.com	insightcrime.org	www.mibolsillo.com
ftnnews.com	dailycaller.com	pursuitist.com	www.goiowaawesome.com

Source: Author's own creation

Table A3	The 149 sites of which the Canada articles were collected	
ca.travelpulse.com	www.billie.ca	www.psychologytoday.com
ca.news.yahoo.com	mtltimes.ca	olympic.ca
www.ledevoir.com	www.clindoeil.ca	www.noovo.ca
www.travelweek.ca	o.canada.com	www.ecosdelcombeima.com
www.lapresse.ca	www.sudbury.com	www.guelphmercury.com
www.thestar.com	www.cheknews.ca	www.lanouvelle.net
www.tvanouvelles.ca	ca.sports.yahoo.com	thenelsonpost.ca
www.cbc.ca	www.todocanada.ca	www.uvic.ca
www.ctvnews.ca	www.therecord.com	www.hcamag.com
www.theglobeandmail.com	www.latribune.ca	toronto.citynews.ca
www.newswire.ca	www.townoflaronge.ca	kawarthanow.com
www.journaldequebec.com	bravewords.com	www.universitymagazine.ca
www.journaldemontreal.com	infotel.ca	www.lecitoyenvaldoramos.com
www.narcity.com	www.acadienouvelle.com	lethbridgenewsnow.com
ici.radio-canada.ca	skiesmag.com	www.journalexpress.ca
globalnews.ca	hollywoodpq.com	triathlonmagazine.ca
www.lesoleil.com	winnipeg.ctvnews.ca	www.nationalobserver.com
www.eonline.com	www.toronto.com	indie88.com
journalmetro.com	www.985fm.ca	www.canadiancattlemen.ca
laoferta.com	www.winnipegfreepress.com	www.iheartradio.ca
ca.movies.yahoo.com	www.silo57.ca	montreal.citynews.ca
montreal.ctvnews.ca	nightlife.ca	www.fightbookmma.com
milesopedia.com	canucksarmy.com	www.chch.com
nationalpost.com	nugget.ca	www.castanet.net
ca.yahoo.com	edmonton.citynews.ca	www.lavantage.qc.ca
www.vancouverisawesome.com	www.tbnewswatch.com	ygknews.ca
www.noovomoi.ca	www.sasktoday.ca	ca.finance.yahoo.com
www.bnnbloomberg.ca	1075koolfm.com	urbania.ca
www.insauga.com	torontolife.com	www.straight.com
www.macleans.ca	www.lavoixdusud.com	www.redbull.com
senderoartesmarciales.com	chatnewstoday.ca	www.cp24.com
cimtchau.ca	www.bradfordtoday.ca	www.telemundoarizona.com
tubarco.news	www.thepost.on.ca	showbizz.net
exclaim.ca	elcentronews.net	theconcordian.com
thepostmillennial.com	www.lhebdojournal.com	thetyee.ca
www.thebeaverton.com	www.thelawyersdaily.ca	atlantic.ctvnews.ca
veroniquecloutier.com	www.meteomedia.com	cannabislifenetwork.com
kitchener.ctvnews.ca	www.netnewsledger.com	www.electriccitymagazine.ca
www.lequotidien.com	www.slice.ca	www.timescolonist.com
ottawa.ctvnews.ca	thereview.ca	www.thereminder.ca
www.droit-inc.com	519magazine.com	oilersnation.com
www.kelownanow.com	canadiantravelnews.ca	viewthevibe.com
lfpress.com	bc.ctvnews.ca	zone911.com
www.vaughantoday.ca	edmonton.ctvnews.ca	www.thenewswire.com
etcanada.com	informefracto.com	passages.winnipegfreepress.com
ottawasun.com	www.lavantagegaspesien.com	thekit.ca
osoyoostoday.ca	www.refinery29.com	info.51.ca
www.thespec.com	www.kelownadailycourier.ca	www.yorkregion.com
www.msss.gouv.qc.ca	www.technologynetworks.com	www.shorelinebeacon.com
www.rds.ca	www.timminspress.com	

Source: Author's own creation

Corresponding author

Angel Díaz-Pacheco can be contacted at: angel.diaz@ugto.mx