No selfies, please: An analysis of travel texts and images from Reddit

Article · March 2022

CITATIONS
0

READS
226

1 author:

Rafael Almeida de Oliveira
Federal University of Minas Gerais
22 PUBLICATIONS **44** CITATIONS

SEE PROFILE

EJABM European Journal of Applied Business and Management

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

Research Paper

No selfies, please: An analysis of travel texts and images from Reddit

Submitted in 24, May 2021
Accepted in 20, March 2022
Evaluated by a double-blind review system

RAFAEL ALMEIDA DE OLIVEIRA[1*]

ABSTRACT

Purpose: This research will analyse the main posts from Reddit, more specifically the subreddit dedicated to travel known as r/travel.

Methodology: The 946 most relevant posts and their 891 images were collected by web scraping and their titles were analysed and gathered using the image classification technique.

Findings: Results show that most Reddit posts do not have people in the pictures, drifting away from the popular selfies that are ubiquitous in other social networks. It is believed that users prefer to post images on Reddit without people due to its anonymous appeal. Moreover, there is an evident focus on images associated with natural environments and less so with urban environments; also, images taken at or around cities and countries out of the beaten track are favoured and it supports the idea that Reddit users search for less-known destinations to guarantee exclusivity during the tourist's experience.

Originality/Value: Even though there is a high number of tourism research connected to social networks, there is none that tries to understand Reddit users' approach to tourism. It is also important to highlight that there are few studies using the image classification technique towards the identification and quantification of objects included in travel photography.

Keywords: Tourism, Web scraping, Image classification.

1. Introduction

The explosion of the use of smartphones and the evolution of the internet capacity of interconnection reshaped the culture in which we currently live. We left the reality of mass transmission technologies such as television, radio, and computers, to mobile devices that allow a huge information traffic that affects all aspects of our daily life. Regarding tourism, there was an expansion of platforms and apps specialized in navigation, accommodation, evaluation, and image sharing, creating new kinds of interaction among suppliers and tourists (Jansson, 2018; Sotiriadis, 2017).

Many of these interactions take place on social networks, characterized by virtual spaces where users can evaluate and exchange information about services and tourism products. These assessments help other users to decide on planning the next trips, making social networks become essential channels for the promotion of tourism (Sigala, 2016).

Much of the existing content on social networks is produced by users themselves, calling the concept of user-generated contents (UGC). In the case of tourism, UGC has become

[1]*Autor correspondente. Federal University of Minas Gerais, Brazil. E-mail: rafalolbh@hotmail.com

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

an important part of the trip, as it allows tourists to plan their itineraries based on information posted by tourists who have already experienced that destination (Narangajavana Kaosiri et al., 2019; Ukpabi & Karjaluoto, 2018). This context makes UGCs more reliable than the information found on official tourist destination websites (Han et al., 2018; Lam et al., 2020; Ukpabi % Karjaluoto, 2018). The ease of access for tourists to search for information and evaluate services and destinations from UGCs makes the market need to readapt its strategies for publicising destinations (Zhang et al., 2021).

Social networks can be divided in two ways based on the form of UGC available. The first, known as relationship-based environments is formed by social networking sites (SNSs), i.e., spaces that facilitate information exchanges between people who share social connections. The second, known as topic/interest-based environments is formed by review sites (RSs) that allow public access to information posted by users and that encourage anyone to distribute and comment on topics of interest. These two channels differ in that while SNS demands a social connection for the information to be accessed by the user, SRs allow the information to be viewed by anyone (Zhang et al., 2021).

Of the most popular SNSs, both Facebook and Instagram are often used by users to post travel photos in order to share them with friends and relatives, as well as serving as an album that can be viewed years later (Douglas, 2014). Among the RSs available, Reddit has had exponential growth in the past few years, becoming one of the biggest virtual communities in the world (Proferes et al., 2021). Unlike Facebook and Instagram, Reddit is composed of thematic discussion forums, where tourism is present through the publication of images, videos, advice, discussions, and questions about destinations. In addition, many profiles are created anonymously, facilitating discussions of more sensitive topics or uninhibited feelings within the platform (De Choundhury & De, 2014).

The proliferation of social media makes the form of distribution and the style of information different in each one, being fundamental that the managers of tourist destinations understand their particularity for a good marketing management (Han et al., 2018). Although they recognise the importance of using UGC to publicise destinations, many managers do not know how to work them in a correct way to take advantage of them for tourism (Zhang et al., 2021). The UGC platforms have the potential to directly impact the image of a destination and thus, managers should understand the profile of these communication channels and encourage users to post information that promote local tours and businesses. In addition, they must ensure that the image disclosed in each of the networks is the desired image, increasing its reputation in the digital environment (Lam et al., 2020). However, in the case of Reddit, despite its great growth, studies on the behaviour of its users and the content published on the network are practically nonexistent in the academic tourism sector.

It is also observed that although there are several works of analysis of texts and comments extracted on social networks, there is no research in the tourism sector that contemplates the classification of images posted by users. Image classification is defined as an automated system that can process data from an image by categorizing it into predefined classes to generate useful information for analysis (Rawat & Wang, 2017; Maulik & Chakraborty, 2018).

Thus, this work aims to analyse the profile of Reddit users' posts from a specific travel forum known as r / travel. To this end, text and images were collected from 998 posts, considered the most relevant to the forum from its foundation in 2008 until April 2020,

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

using the data extraction technique known as web scraping. The content of the titles and images in each post were analysed using text mining and image classification tools.

It is hoped that this article can be a pioneer in the analysis of content in tourism on Reddit, not only from texts but mainly from the classification of images, facilitating the understanding of which content is more accepted by the users of the platform. Within the domain of computer vision, image classification is considered the area of greatest research interest, achieving important results in computational competitions around the world (Pathak et al., 2018) and can be used constantly in the future of tourism research. Understanding the acceptance of users by certain types of post facilitates the creation of more relevant content within the platform by tourism agents, improving the promotion of their destinations within Reddit.

This work is divided into five stages. The first, characterized by the literature review, will address debates on tourism and social networks, the Reddit social network and web scraping and image classification techniques. The second stage will show the data collection and analysis methodology that resulted in the third stage of presenting the results. After the fourth stage of discussion of the results, we present the conclusion to elucidate the debates shown in this paper and a possible future research agenda.

2. Literature Review

2.1. Tourism and social networks

All different kinds of social networks had the potential to transform tourists, previously seen as mere consumers, into actors with a fundamental role in the co-creation of the tourism product and in the act of sharing experiences from comments, reactions, or image publication. Now, they prefer to interact and learn from contact with other tourists about tourism destinations instead of being a mere observer. Thus, tourists search, throughout the several social networks, the information they need regarding their trips; about the creation of a touristic experience from other people's behaviour on a shared digital society; and, ultimately, they also influence the travelling desires on the same network (Sigala, 2016). Taking Instagram as an example, photographs can be geotagged and commented and this contributes towards the creation of cultural identity for tourist attractions (Jansson, 2018).

Comments on social networks are shown as an important element while searching for travel information as future tourists (Ukpabi & Karjaluoto, 2018). This tool is considered as an electronic version of the word-of-month: eWOM. Such comments reflect travellers' opinions regarding their trips and they also contain product and service recommendations in a qualitative way. However, it is often difficult for those who are searching for information and have little time to read all the comments, so the first phrase of the comment becomes fundamental in the delivering of the information. Working as a newspaper headline, the titles of the comments help to create the tourist product's first impressions and to shape the author's general opinion towards his or her trip experience (De Ascaniis & Gretzel, 2013).

On the one hand, the fact that there are several information sources allows destinations to create and promote alternative information regarding an attraction. They may suggest new gazes for the tourist and new ways to interpret spaces, altering the format of conventional tourism. On the other hand, however, such exposition may increase the visits demand, converting trips into less exclusive ones. This propels people into searching for new

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

authentic experiences, without staging or middlemen from destinations' promotion departments, as a means to differentiate themselves from other people in their community. So, the search for distinct places becomes a way to increase the individual's symbolic and cultural capital towards everybody else (Jansson, 2018).

Social networks have never had so many visual elements in their composition and marketing professionals know well the persuasive potential in images. The advance of digital cameras and the convenience of communication connection allow photographs taken during a trip to be shared in real-time from social networks, encouraging visual content production and sharing (Gretzel, 2017). While information is shared from textual elements, such as comments and posts, the sharing of experiences is done through audiovisual elements (Munar & Jacobsen, 2014).

Photographs work as an auto expression tool and it is common to share 'selfies' as a way of empowering the tourist's online identity and to enhance a feeling of belonging to the online society. So, trips are photographed not only for the travellers' eyes but for everybody else's (Sigala, 2016). We can understand the 'selfie' as the person's desire to register him or herself in a photograph to be shared live with an online audience, converting the lenses into a mirror of the audience itself. So, the person aims to put and present him or herself in the visual focus, instead of focusing on what is told by the image, capturing not only what is extraordinary during the trip, but capturing what is extraordinary in him or herself as well. As the destination becomes the background for the photograph, the tourist is elevated as the tourist product, eg, the tourist travels and consumes so that the tourist can have a visual record of the self, enhancing his or her virtual identity (Dinhopl & Gretzel, 2016). On social networks, such an enhancement is materialized from the number of 'likes' and 'shares' that the tourist gets from sharing his or her experience (Sigala, 2016).

This context forces tourism marketing professionals to understand, ever so much, the relationship between the creation and the consumption of visual media created and shared by tourists so that there can be new promotion strategies to the intended market. Studies and research that aim to define the consumer's profile on the several platforms available become essential to manage and influence tourist behaviour that, in turn, benefit tour operators and tourism destinations. In general, such management is done from the use of hashtags, number of photographs, or mentions on specific content. Nevertheless, tourists are increasingly sharing visual content without textual elements, making it more difficult for brands to assess their digital reputations. Therefore, it is fundamental to create new ways of image analysis and tracking such as the use of machine learning software that allows the identification of objects, logos, products, and people to help the interpretation of visual content to ease promotion strategies (Gretzel, 2017).

2.2. Reddit

It is known for being a platform where users may send, comment, and evaluate posts and links shared by other users (Singer et al., 2014). All the content posted on the social network may be up- or downvoted and it gets moved higher or lower on the site in a direct reference to the difference between positive and negative votes (Medvedev et al., 2019; Ovadia, 2015; Weninger et al., 2013).

According to data surveyed by Statista (2022) in June 2021, 49% of Reddit users originated from the United States, followed by the United Kingdom (7.7%) and Canada (7.5%). In October 2021, 63% of visitors to the platform were male. In the United States, in February 2021, 36% of users were aged 18-29. Additionally, people who attended

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

college were more likely to use Reddit, when compared to people with lower levels of education.

Reddit is divided into several thematic forums called 'subreddits'. They may be created by any member and they generally have an address in the form of "r/subject-to-be-discussed". In this way, a politics subreddit may have an address like r/politics and the one dedicated to travel may have the address r/travel (Massanari, 2013). Each subreddit is independent, it is monitored by volunteers, and it has its own rules regarding posts – it allows, for example, a subreddit where only photographs may be posted (Singer et al., 2014; Weninger et al., 2013).

The votes on content within each subreddit are weighed in karma points. When the content is upvoted, the user receives karma points and the content goes up on the forum list; nevertheless, if there are downvotes, the karma goes down and the content goes down the list as well. The more karma points a content has, the more attractive it is shown to users, allowing a higher engagement with that forum's community (Massanari, 2013; Medvedev et al., 2019; Ovadia, 2015).

Reddit's community can, through the points system, create a high sense of organization of the content within the platform and, because of that, the website itself considers it "the front page of the internet" (Jamnik & Lane, 2017; Weninger et al., 2013). In the past few years, celebrities such as actors, presidents, and Nobel laureates have made their presence known on the website, increasing its global influence (Weninger et al., 2013).

Any person on the Internet may access Reddit's content anonymously without a proper account. However, only registered users may post, vote, or create communities on the website. Registration requires only a username and a password, allowing users to be anonymous even after registration (Weninger et al., 2013).

Because of the easy access, and quantity and quality of the information, Reddit was a source of interest for scientific researchers (Medvedev et al., 2019). Even though the website itself is not academic grounded, researchers can find content and conversations on Reddit that are not easily found elsewhere (Ovadia, 2015). Understanding how people behave on online forums may bring to light individual and collective thoughts, helping to deepen debates in several themes (Medvedev et al., 2019).

2.3. Web scraping

The large volume of data found in digital media added to the variety of available formats and the speed of capture and dissemination of information ended up culminating in the definition of the term Big Data (Mcafee & Brynjolfsson, 2012). In other words, it is characterized by a set of gigantic and very complex data that requires differentiated servers and communication technologies that can effectively organize the content produced for management analysis (Chen et al., 2012). From then on, digital data are collected and correlated autonomously, making it possible to interpret information that would have previously been impossible to be analysed only by individuals (Andrejevic, 2014).

Information shared on social media and the Internet as a whole has made the use of digital data extraction techniques a great tool to support managers of organizations and companies during the decision-making process (Chen et al., 2012; Devika & Surendran, 2013). These tools, known as web scraping or scrapers, made it possible to expand the sociological potential of the Internet, from the moment that data inserted in large online platforms such as Facebook, Twitter and Wikipedia could be analysed (Marres & Weltevrede, 2013).

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

In general, the web scraping technique allows unstructured data available on pages on the Internet - usually encoded in HTML format - to be transformed into a structured and manageable database that can be analysed by any researcher (Landers et al., 2016; Vargiu & Urru, 2012). Websites are created, in most cases, to assist in the visualization of information and not to display data in a structured way. Extracting this information manually can be very time consuming and susceptible to extraction errors (Vargiu & Urru, 2012) and in this context, the extraction of information in an automated way has a fundamental role in the analysis of the exposed data (Devika & Surendran, 2013).

The pages of a website usually have templates or layouts that look like each other, similarly disseminating information in the architecture of each page. In this way, the web scraping tool searches for the essential information of a page within the site, replicating the same process for the following pages, enabling the extraction of data in an organized manner (Marres & Weltevrede, 2013). That is, web scraping can be considered as the reverse process of creating a website. When creating, the site is made from a code to the final template for the user. Web scraping starts from the search for information contained in a template for the extraction of data that fed the site (Devika & Surendran, 2013). Thus, the tool transforms unstructured data into structured banks, eliminating elements that are irrelevant and formatted to produce information in an orderly and usable manner (Marres & Weltevrede, 2013)

The extraction of web data enables analysis of information, including in real-time, from digital platforms powered by the users themselves and understanding the behaviour of this audience can help answer various questions and problems faced by researchers and organizations (Landers et al., 2016; Oliveira & Baracho, 2018). They also assist in the extraction of large-scale data from content in the historical line of events, facilitating longitudinal analyses, minimizing contamination risks by the researcher of the sample being observed, in addition to obtaining data from regions and countries that are difficult to access, which could increase the efforts of a survey (Landers et al., 2016).

2.4. Image classification

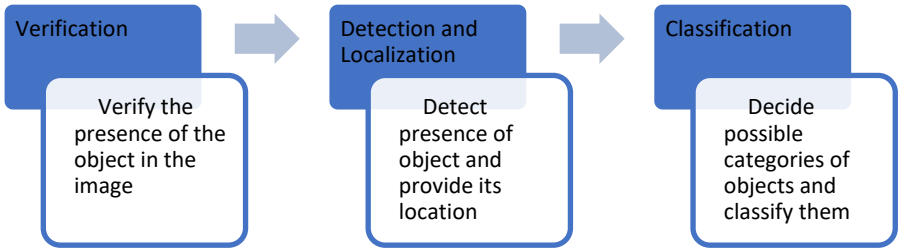
With the advance of information systems in the last decade, the availability of large databases and the power of graphic processing units have increased, culminating in a greater interest in the development of solutions for image classification (Chen et al., 2015; Pathak et al., 2018). Although image classification is considered natural for humans, it is a major challenge for automated systems (Rawat & Wang, 2017).

Autonomous classification systems must define which class an image belongs to. This problem becomes more difficult when you have several classes available or when you have several objects of different classes within the image. Thus, an image may not belong to just one class, but several classes simultaneously (Druzhkov & Kustikova, 2016).

The image classification process comes from the concept of machine learning. This concept is defined by the ability of algorithms to search for the best parameters of a given model created from a database. For this, the algorithms are trained from the insertion of information, making a given model be constantly improved based on the observed results, minimizing errors and increasing its assertiveness. (Maulik & Chakraborty, 2018). In an image classification model, the algorithms receive an image database, identify, and extract the objects from each image by using probabilities to distinguish the objects from the background view (Chang et al., 2020). After that, the algorithms classify them from pre-established categories. The greater the number of images received, the greater the ability of the algorithm to find patterns among objects by classifying them assertively. In summary, the steps for image classification are illustrated in Figure 1.

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

Figure 1. Image classification process steps



Source: Adapted from Pathak et al. (2018)

Algorithms learning can be classified as supervised or unsupervised (Sathya & Abraham, 2013). The supervised machine learning algorithms are the ones that need external help to learn. The main dataset is divided into train and test datasets. First, the algorithm learns some patterns from the training dataset and applies them to the test dataset for classification. The unsupervised learning algorithms learn when new data is inputted, so it uses predictions previously learned to recognise what class the data belongs to (Dey, 2016; Sathya & Abraham, 2013).

One of the image classification systems currently available was created by Amazon. The Amazon SageMaker image classification algorithm is a supervised learning algorithm that supports multi-label classification. It can take an image as it inputs and outputs one or more labels correlated with the image. It can also be trained from scratch or by transfer learning when a sufficient number of images for training is not available (Amazon, 2020). The system is known as one of the most precise of the market with an accuracy of 99.45% (Babbar et al., 2020).

In tourism, image classification can bring several benefits. The popularity of photo distribution websites and applications has considerably increased the volume of digital images. A way of classifying images by the users of these sites, such as Flickr, created great difficulty in organizing the data and, mainly, made it difficult to retrieve images from search engines. Therefore, an automated classification of images can reduce the subjectivity of the classification of users, making indexing more accurate and assists in strategies for monitoring and promoting images in the digital environment. Besides, it can facilitate the development of automated travel recommendation systems for tourists, based on the photos posted by the user on social networks (Pliakos & Kotropoulos, 2014).

3. Methodology

For data collection, 998 posts with the highest number of positive votes of all time were chosen in a subreddit specially dedicated to the theme of travel, called r / travel. The number of posts collected refers to the maximum number that the platform allows the user to see in the feed. The subreddit was created on January 25, 2008, and the data showed 4.6 million users on April 16, 2020. Using the Chrome browser extension called Data Miner, the title, type of content, number of positive votes, number of comments and photo were collected from the posts, if they existed in the post. Figure 2 presents an example of the elements collected from a post detailed by numbers: 1 - title, 2 - type of content, 3 - number of positive votes, 4 - number of comments, 5 - photo.

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

After extraction, promotional posts were identified, which were removed from the analysis, reducing the number of posts in the database to 946. The textual data were stored in an Excel spreadsheet. From the posts, 891 photos were also extracted and stored in a workbook.

The analysis of the title terms of the posts was made using a website (voyant-tools.org) that allowed counting the number of repetitions of each of the 2811 distinct terms in the database. The site enables the user to enter a full text or textual database and automatically generates a list of all existing terms together with the number of times the terms are repeated. In this research, the complete database with the titles of the 946 posts was uploaded into the site and the results appeared instantly. Afterwards, the results were separated between the terms that identified tourist destinations and the other elements, creating two distinct databases.

Figure 2. Example of Reddit post

Source: Reddit

Image classification was carried out from the clarifai.com website, which uses the Amazon SageMaker image classification algorithm to identify and classify the objects in each image. When inserting an image in the system, the site lists a ranking of objects that most closely relate to that image. In this research, when uploading each photo captured from Reddit on the website, the tool automatically identified the top ten objects that related to that image, listing them based on the relevance percentage that the object could be represented in the image. The values ranged from 0.0 to 1.0, and the higher the value, the higher the chances of the classified object being contained within the image. Obviously, depending on the content of each image, different objects could be classified.

After uploading the 891 photos collected from Reddit on clarifai.com, an automatic dashboard was formed where each photo had a different page address or Uniform Resource Locator (URL). When clicking on each photo, the algorithm automatically classified the objects contained in the photos in a table format, containing the name of the object and its relevance value, from highest to lowest. To collect the information from clarifai.com and transform it into a structured database, the WebHarvy software was used. Like the Data Miner, the software is a webcraper that enables the extraction of web data from the insertion of URLs. First, the URL of clarifai's photo dashboard was inserted into the WebHarvy tool. From that URL, WebHarvy could access the database as a common

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

web browser. Then, the webscraper was automated to collect each URL from each photos, forming a database with 891 URLs.

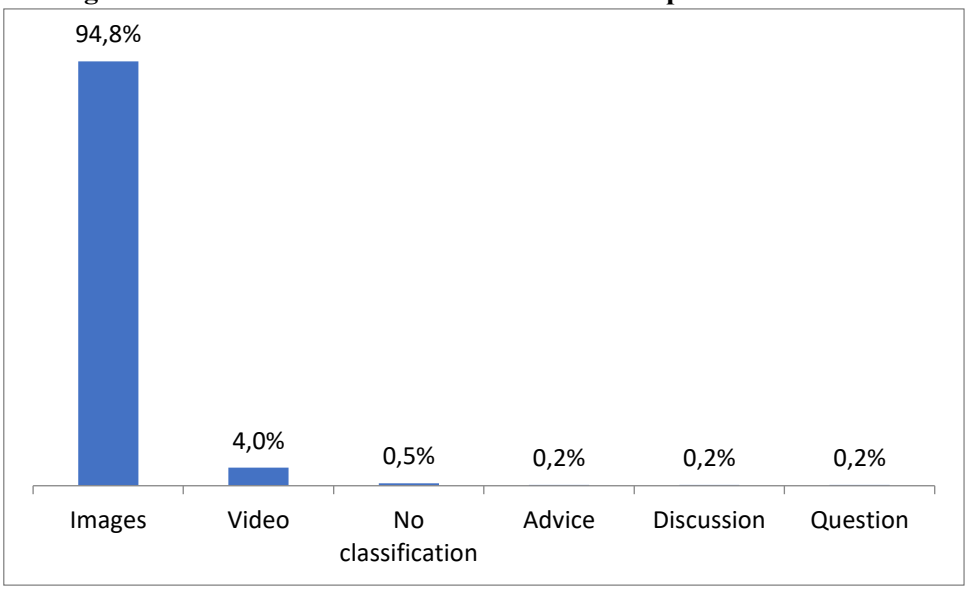
The next step was to create a new webscraper from WebHarvy and teach it to collect the object tables from each of the 891 photos. To do that, the 891 URLs of the collected photos were inserted in the webscraper allowing it to access each of the photos automatically and download each of the object tables. Finally, the software resulted in an excel database with 891 lines, representing each photo, and 10 columns, each one containing textually the object name and its relevance value for each photo. For instance, the first column could have the text water1.00, the second travel0.98 and the third nature0.98. To clean the data and separate between texts and numbers, the software PowerBi was used. The software made it possible to transform the content of the 10 columns into 20 columns, containing 10 columns only with the name of the objects and 10 with the respective relevance values. For the example presented, the column with the values water1.00 was transformed in a new column only with the text water and a new column only with the number 1.00.

Finally, for the analysis of the results, all the text columns were aggregated into only one, generating a new database with 539 different lines representing the variety of objects classified in the set of all 891 photos. Thus, two databases were analysed, the first with the titles and types of content of 946 posts and the second with the objects of the 891 photos.

4. Results

The 946 posts collected were divided according to the type of content. The content type of posts on Reddit can be classified in different ways inside the platform. Figure 2 shows that the vast majority of the most relevant posts on subrredit r/travel were classified as images (95.3%) and only a small portion stood out as videos (4.0%).

Figure 2. Distribution of the content of the main posts on Reddit



European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

When analysing the content types by the average of positive votes, it can be seen that discussion posts have the highest values, followed by video content, images, advices and questions. The total average of positive votes of the main travel posts was 7612. The data can be seen on Table 1.

Table 1. Average of positive votes by type of content

Type of content	Average of positive votes
Discussion	24900
Video	8939
Images	7514
Advice	7150
Question	6000
Average	7612

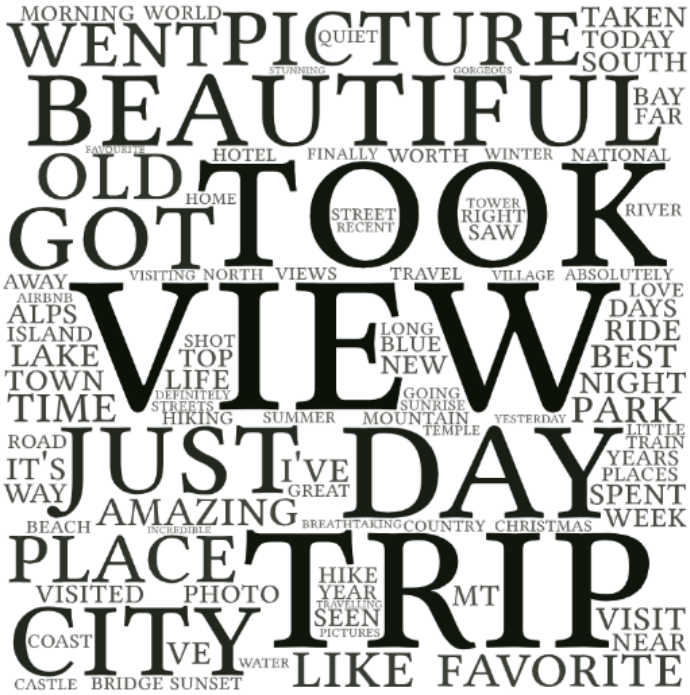
Regarding the average number of comments by type of content, it can be seen that, as well as the average number of positive votes, posts about discussions have the highest number of comments followed by posts about questions, advices, videos and images. The total average number of comments of the travel posts analysed was 164, according to Table 2.

Table 2. Average of positive votes by type of content

Type of content	Average of comments
Discussion	1217
Question	584
Advice	199
Video	196
Images	159
Average	164

The results of the most relevant terms extracted from the post titles show the predominance of words that relate to the user's act of taking the picture (picture, photo, took, taken, I've) in addition to the beauty of the places visited (beautiful, amazing, best, favorite, absolutely) and especially the landscape of the place by the tourist (view). The words also highlight in a varied way some attractions observed by the tourist such as city, alps, island, lake, town, beach, coast, castle, river, bay, village, etc. In the case of means of accommodation, similar relevance is found in the words hotel, Airbnb and home. Finally, it is important to highlight several terms that reinforce the immediacy of the post such as morning, just, today, as shown in Figure 2.

Figure 2. Most cited words in post titles



In the analysis of the 75 destinations that appear the most in the post titles, not only countries in Europe that have strong tourist demand, such as Italy, Switzerland, Germany, Spain and France, but also a range of countries around the world that stand out are not part of the more traditional tourism. Routes such as Iceland, Japan, Norway, China, Morocco, Scotland, Portugal, Vietnam, Croatia, and Turkey were cited. In the case of cities, the main highlights were Amsterdam, Venice, Prague, Paris, London, Petra, Cappadocia, Florence, Lauterbrunnen and Rome, as shown in Figure 3.

Figure 3. Most cited destinations in post titles

European Journal of Applied Business Management, 8(1), 2022, pp. 29-44 ISSN 2183-5594

Finally, the classification of objects from images posted by users shows that, according to Table 3, the algorithm used was able to identify in the set of photos that 87.9% referred to travel figures. Then, 72.2% of the photos did not have the presence of people, with only 10.4% of the photos showing this object. The main objects were manually classified between two categories - natural and urban, in order to check the predominance of the types of landscapes posted on Reddit. The results point to a greater distribution of elements related to natural landscapes (16) as opposed to the urban or built environment (7).

Table 3. Image classification terms

Objects	Frequency	% of images	Category
travel	783	87,9	-
no person	643	72,2	-
landscape	440	49,4	nature
architecture	384	43,1	urban
water	377	42,3	nature
nature	323	36,3	nature
city	285	32,0	urban
outdoors	284	31,9	nature
sky	273	30,6	nature
mountain	252	28,3	nature
building	179	20,1	urban
street	141	15,8	urban
river	125	14,0	nature
tree	124	13,9	nature
old	122	13,7	urban
snow	116	13,0	nature
house	106	11,9	urban
sea	98	11,0	nature
town	95	10,7	urban
people	93	10,4	-
wood	92	10,3	nature
lake	88	9,9	nature
rock	86	9,7	nature
beach	83	9,3	nature
seashore	83	9,3	nature
sunset	83	9,3	nature

5. Discussion

The results show that the social network Reddit.com has a great predominance of photographs in the best-rated posts, strengthening the role of the image as a way of self-expression of the tourist and acceptance by the community in which he finds himself (Gretzel, 2017). Adding the result of the posts containing photos or videos, the visual elements represent almost all the topics best evaluated by users. These posts end up influencing the construction of the network users' travel wishes (Sigala 2016), represented by the number of votes that users receive for their posts. That is, if I vote on that post, I reaffirm my desire to also visit that place and at the same time increase the symbolic

capital of the user who published the post (Jansson, 2018). On the other hand, images and videos did not have the same engagement as discussion posts, questions and advice, being commented on less frequently. Thus, destinations that want to gain relevance from engagement can publish posts that primarily encourage discussions about the destination.

The words contained in the titles of the posts reinforce the idea of the immediate use of Reddit for sharing experiences during the trip (Gretzel, 2017) as well as in other social networks. However, unlike other platforms, Reddit has no concern for the user to put himself in the place of destiny through the use of selfies (Dinhopl & Gretzel, 2016). Yes, there is a concern to share, literally, his vision of the recorded experience. In this case, the tourist's gaze is the focus of the image and not the tourist itself. The idea then is to share with the community the same image that the tourist sees, the destination is the focus of the experience. This may happen because on Reddit, the culture of anonymity is more prevalent than user exposure compared to other social networks such as Facebook or Instagram. This anonymity ends up being reflected in the non-presence of the tourist in the image, resulting in a greater concern with the content that is passed on and not on the user itself.

From the analysis of the most cited destinations in the comment titles, it is clear that the Reddit user is closer to the idea of seeking experiences authentically, from trips to places other than usual tourism as a way to differentiate themselves in the community. (Jansson, 2018). Thus, in the most relevant posts, there is a predominance of destinations that deviate from traditional routes, creating the feeling of exclusivity of the experience and making this feeling reinforced by the community itself, which values this type of novelty through positive votes. This can also be proven by the fact that most of the elements contained in the photos are linked to the natural environment, known to be more remote and exclusive when compared to urban environments with a greater movement of people.

The behavioural characteristics of Reddit users can help tourism marketing professionals to develop more focused strategies for publicizing their destinations, increasing the success of the campaigns developed (Gretzel, 2017). In the case of the results presented, it is expected that campaigns aimed at the nature segment, with more exclusive destinations and focusing on the user experience and not on the user itself, may be more successful with the target audience of the network analysed.

6. Conclusion

Social networks have shaped the way tourists experience their travels, transforming them from consumers into co-creators of the product. With each post or comment on social networks, the tourist strengthens his image before his network and at the same time influences the travel choice of other members of the community. This study aimed to analyse the most relevant posts on the Reddit social network based on the mining of the texts from the post titles and the classification of the objects contained in the posted images. Although the data mining technique known as web scraping is widely used in the academic field for analysis on social networks, image classification is still not very common. Therefore, when showing the success of using this technique for studies on social networks, it is expected that it can be used more consistently in future research.

The results demonstrate that there is a greater appreciation of destinations about the user in the posts of members in the social network. Unlike other communication channels, on Reddit, the anonymity of profiles means that the published images also do not have the presence of the tourist, distancing themselves from the phenomenon known as selfie.

Thus, the post is more valued by the community for the way the destination is portrayed and not for who portrayed it. Besides, landscapes focused on natural elements proved to be more relevant, as well as the variety of destinations outside the usual itineraries of visitation, showing that Reddit users have preferences for more exclusive locations. It is also noteworthy that posts that encourage discussions about a destination or travel, despite being seen more rarely in the best rated posts, showed a higher level of engagement than the other types of content, opening a field for new marketing strategies within the platform.

It is hoped that understanding the profile and behaviour of Reddit users can help develop more effective marketing strategies from within the social network. It is worth mentioning that despite being a widely used network, studies focused on tourism on Reddit are practically nonexistent. Therefore, this study aims to initiate a first analysis of the network and hopes to contribute to future discussions.

It is believed that, for future research, topics related to the deepening of text mining techniques and mainly image classification can be studied, as well as research aimed at comparing results collected on Reddit with other posts from other social networks. Successful and unsuccessful tourism promotion strategies carried out on Reddit can also be studied, expanding the potential of the network as a means of communication for tourism.

References

Andrejevic, M. (2014). The big data divide. *International Journal of Communication, 8*(1), 1673–1689.

Amazon (2020). *Amazon SageMaker*. https://aws.amazon.com/sagemaker/data-scientist/

Babbar, S., Dewan, N., Shangle, K., Kulshrestha, S., & Patel, S. (2020). *Cross-Age Face Recognition using Deep Residual Networks*. 257–262.

Chang, M., Xing, Y. Y., Zhang, Q. Y., Han, S. J., & Kim, M. (2020). A CNN image classification analysis for "clean-coast detector" as tourism service distribution. *Journal of Distribution Science, 18*(1), 15–26.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: from Big Data to big impact. *Mis Quarterly*, *36*(4), 1165–1188.

Chen, Q., Song, Z., Dong, J., Huang, Z., Hua, Y., & Yan, S. (2015). Contextualizing object detection and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(1), 13–27.

De Ascaniis, S., & Gretzel, U. (2013). Communicative functions of online travel review titles. A pragmatic and linguistic investigation of destination and attraction OTR titles. *Studies in Communication Sciences*, *13*(2), 156–165.

De Choundhury, M., & De, S. (2014, May). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*., 71-80.

Devika, K., & Surendran, S. (2013). An overview of web data extraction techniques. *International Journal of Scientific Engineering and Technology*, *2*(4), 278–287.

Dey, A. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179.

Dinhopl, A., & Gretzel, U. (2016). Selfie-taking as touristic looking. *Annals of Tourism Research, 57*, 126–139.

Douglas, N. (2014). It's supposed to look like shit: The Internet ugly aesthetic. *Journal of visual culture*, *13*(3), 314-339.

Druzhkov, P. N., & Kustikova, V. D. (2016). A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis*, *26*(1), 9–15.

Gretzel, U. (2017). The visual turn in social media marketing. *Tourismos: An International Multidisciplinary Journal of Tourism*, *12*(3), 01–18.

Han, W., McCabe, S., Wang, Y., & Chong, A. Y. L. (2018). Evaluating user-generated content in social media: an effective approach to encourage greater pro-environmental behavior in tourism? *Journal of Sustainable Tourism*, *26*(4), 600–614.

Jamnik, M. R., & Lane, D. J. (2017). The use of Reddit as an inexpensive source for high-quality data. *Practical Assessment, Research and Evaluation*, *22*(5), 1–10.

Jansson, A. (2018). Rethinking post-tourism in the age of social media. *Annals of Tourism Research*, *69*, 101–110.

Lam, J. M. S., Ismail, H., & Lee, S. (2020). From desktop to destination: User-generated content platforms, co-created online experiences, destination image and satisfaction. *Journal of Destination Marketing and Management*, *18*(July).

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: automatic extraction of Big Data from the internet for use in Psychological Research Richard. *Psychological Methods*, *21*(4), 475–492.

Marres, N., & Weltevrede, E. (2013). Scraping the social? *Journal of Cultural Economy*, *6*(3), 313–335.

Massanari, A. (2013). Playful participatory culture: learning from Reddit. *Selected Papers of Internet Research*, *3*, 1–7.

Maulik, U., & Chakraborty, D. (2018). Remote sensing image Classification: A survey of support-vector-machine-based advanced techniques. *IEE Geoscience and Remote Sensing Magazine*, *XLII*, 33–52.

Mcafee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, *October*, 1–9.

Medvedev, A. N., Lambiotte, R., & Delvenne, J. C. (2019). The anatomy of Reddit: an Overview of academic research. *Springer Proceedings in Complexity*, 183–204.

Munar, A. M., & Jacobsen, J. K. S. (2014). Motivations for sharing tourism experiences through social media. *Tourism Management*, *43*, 46–54.

Narangajavana Kaosiri, Y., Callarisa Fiol, L. J., Moliner Tena, M. Á., Rodríguez Artola, R. M., & Sánchez García, J. (2019). User-generated content sources in social media: a new approach to explore tourist satisfaction. *Journal of Travel Research*, *58*(2), 253–265.

Oliveira, R. A. de, & Baracho, R. M. A. (2018). The development of tourism indicators through the use of social media data: The case of minas Gerais, Brazil. *Information Research*, *23*(4).

Ovadia, S. (2015). More than just cat pictures: reddit as a curated news source. *Behavioural and Social Sciences Librarian*, *34*(1), 37–40.

Pathak, A. R., Pandey, M. and Rautaray, S. (2018). Application of deep learning for object detection. *Procedia Computer Science*, *132*(Iccids), 1706–1717.

Pliakos, K., & Kotropoulos, C. (2014). PLSA driven image annotation, classification, and tourism recommendation. *2014 IEEE International Conference on Image Processing, ICIP 2014*, 3003–3007.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C. and Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2), 1-14.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Computation*, *29*, 2709–2733.

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38.

Sigala, M. (2016). Social media and the co-creation of tourism experiences. *The Handbook of Managing and Marketing Tourism Experiences*, 85–111.

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., & Strohmaier, M. (2014). Evolution of Reddit: From the front page of the internet to a self-referential community? *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*, 517–522.

Sotiriadis, M. D. (2017). Sharing tourism experiences in social media: A literature review and a set of suggested business strategies. *International Journal of Contemporary Hospitality Management*, *29*(1), 179–225.

Statista (2022). *Reddit - Statistics & Facts*. https://www.statista.com/topics/5672/reddit/

Ukpabi, D. C., & Karjaluoto, H. (2018). What drives travelers' adoption of user-generated content? A literature review. *Tourism Management Perspectives*, *28*, 251–273.

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research*, *2*(1), 44–54.

Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the Reddit community. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013*, 579–583.

Zhang, Y., Gao, J., Cole, S., & Ricci, P. (2021). How the spread of User-Generated Contents (UGC) shapes international tourism distribution: using Agent-Based Modeling to inform strategic UGC marketing. *Journal of Travel Research*, *60*(7), 1469–1491.