4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15098

Covid 19 and lodging places

Estefania Ruiz-Martinez, Francisco Porras-Bernardez, Georg Gartner
Technischen Universität Wien, Vienna, Austria.

Abstract

Tourism is a very important source of income for national economies all over the world. Before Covid-19, this sector contributed with 10.4% of the global GDP. Innovative tools for tourism study and promotion are very necessary for a future recovery of the industry. Thus, we have explored Airbnb data as a source of information about the lodging sector, very relevant within the tourism industry. We have analyzed these data to explore the experience of tourists before and after the pandemic. Our aims included identifying and visualizing opinion changes through semantics extracted from semi-structured data generated by the Airbnb customers. We used Natural Language Processing and techniques such as sentiment analysis combined with spatial analysis with KDE in order to characterize and spatially visualize user opinion. Results did not show significant differences in user opinion before and after the outbreak of Covid, however spatial patterns related to sentiments were made visible. Moreover, a large dataset covering 3.6M Airbnb lodging spots from 108 cities was compiled and will be made available in the future. This paper can be useful for the lodging industry, tourism organizations as well as social media researchers by providing an alternative approach that involves the role of location in the study of customer behaviour.

Keywords: *Airbnb; Sentiment Analysis; Covid-19; Kernel Density Estimation.*

1. Introduction

The current situation generated by Covid-19 has affected most of the economic sectors in the world. Tourism is one of the most impacted areas due to a fundamental dependence in mobility and safety. The travel and tourism sector lost near to 3.8€ billion in 2020, whereas its contribution to global GDP sunk by a huge 49.1% compared to the previous years. Thus, innovative tools for tourism study and promotion are even more necessary for a future recovery of the industry.

One very relevant part of the tourism sector is the lodging industry. People need a physical location to stay when visiting a new region. These locations and their surroundings are places. For the study of places and human perception and behaviour in them, traditional methods include questionnaires or travel diaries among other tools. However, traditional data collection methods are often limited by the number of participants that can be involved in collection campaigns. These tools can now be complemented or replaced by the use of big spatiotemporal data. User-Generated Content (UGC) (Wyrwoll, 2014) sources offer huge amounts of data usable for the analysis of spatially related phenomena. In particular, in the lodging industry, online reviews are widely used to explore the experience of customers (Xiang et al., 2015). According to (Li et al., 2018), only in tourism research, UGC accounted for almost half of the literature during the last decade.

We aim to study the influence of Covid-19 on the lodging industry by analyzing the experiences of tourists when being hosted in lodging places before and after the pandemic. For this purpose, we analyze Airbnb data including the users' reviews associated with places listed in the platform. We used Natural Language Processing (NLP) techniques such as sentiment analysis to classify online reviews as positive, neutral and negative and then characterize and visualize the spatial distribution of listings accordingly by using Kernel Density Estimation (KDE).

This work can contribute to a better understanding of the impact of Covid-19 and other phenomena on the lodging industry and on the perception of places very relevant for tourism. This paper can be useful for the tourism industry, property owners or the public administration. Moreover, it can provide a better insight into available Airbnb datasets and analysis methods for worldwide researchers.

2. Method

We collected the data and pre-processed it in two steps. The first step involved the collection, preparation and storage of the raw data from listings existing since 2015 until February 2021. The second step had to do with the pre-processing of the online reviews to

implement the semantic analysis using sentiment analysis then visualize the spatial distribution of listings according to the average sentiment.

2.1. Data preparation

The data used in this work was collected from the website Inside Airbnb, which uses public information compiled from the Airbnb website referred to the hosting places (a.k.a. listings). The location information of these listings is anonymized by Airbnb by introducing a geolocation error of 0-150 meters.

The preparation process involved the collection and further processing of the data. The first step consisted in the collection of all the datasets available until February 2021 at Inside Airbnb. Data accounts for 242 GB and includes the listings existing in 108 cities worldwide since 2015. In a second step, a Python script was developed for the data processing and a PostgreSQL spatial database was created for the storage. The raw files contained monthly snapshots of the listings. The temporal coverage for each city varies between 72 and only a few months in some cases. The processing parsed the monthly files selecting 110 attributes and generating a point element for each listing. In order to build our geodataset, we stored all the listings that have existed at any time in each city during the whole collected period. The final number of unique listings reached more than 3.6 Million.

2.2. Text pre-processing

In this step, reviews were prepared for the implementation of sentiment analysis, which required the removal of reviews written in a language different from english. To do so, a script was developed to use the available libraries for text pre-processing in Python. First of all, automated postings (e.g., “This is an automated posting”), non-English, duplicated, empty reviews and reviews consisting in only two characters, numbers or NaN were discarded. Non-English reviews were identified using the python library Fasttext (Joulin et al., n.d., 2016), which employs a pre-trained model to predict the language of a sentence. This model was trained using data from Wikipedia, Tatoeba and SETimes and can identify 176 languages. When used in python, it returns a tuple with an ISO code of the language recognized and a confidence value indicating the probability of the sentence belonging to that language.

2.3. Semantic Analysis

2.3.1. Sentiment Analysis

We performed sentiment analysis on the reviews in order to estimate the polarity of the texts. The sentiment analysis determines sentiment orientation and classifies the reviews into classes of polarity: positive, neutral or negative. For the analysis, we used the VADER

model (Hutto & Gilbert, 2014). This model follows an approach based on valence and considers the sentiment as well as its intensity. VADER is a lexicon and rule-based sentiment analysis tool. Its effectiveness has been compared against eleven state-of-the-art sentiment benchmarks with more favourable results in different contexts and even offering better performance than human raters.

The model relies on a list of lexical features that are labelled as positive or negative depending on their semantic orientation, i.e a sentiment lexicon. VADER also quantifies how positive or negative sentiment is. A text is analysed and its constituent words are searched in the lexicon: positive words have higher ratings whereas negative words lower. All lexicon ratings are combined in a compound score formed by the summarization of all of them and standardized between -1 and 1. A score of -1 represents a fully negative sentiment, a value of 0 denotes a neutral text, and a score of 1 represents a fully positive sentiment.

2.4. Spatial Visualization

2.4.1. Kernel Density Estimation (KDE)

Listings were categorized into positive, neutral and negative polarity, according to the average compound score of all the reviews one year before the outbreak of covid and one year after it. KDE was used to generate density surfaces and visualize the distribution of listings according to sentiment polarity and time period. A bandwidth was determined for the density surfaces of each polarity but was the same for both time periods. It was selected based on an iterative process that finished when the density surface was not either over-smoothed or under-smoothed. The values of pixels from density surfaces were normalized between 0 and 1 using *Map Algebra*. The new values indicate the density of listings in proportion to the maximum density obtained.

3. Results

As a case study, up to now, we have focused on two cities from two continents, i.e. Rio de Janeiro (Brazil) and New York (U.S). Both cities have a different tourism orientation and at the same time, are in two of the most severely affected countries by Covid-19.

To analyse the experience of users after the outbreak of covid, we took as reference the experience before the outbreak as well. To do so, only listings with reviews one year before and after the outbreak were included in the analysis. As a result a total of 26,262 reviews from 3,522 property listings in Rio de Janeiro and 486,438 reviews from 18,751 properties in New York were processed.

3.1. Sentiment Analysis

The proportion of positive, neutral, and negative reviews from both Rio de Janeiro and New York did not significantly change after the outbreak of Covid-19 (see Figure 1). Positive reviews from Rio de Janeiro decreased 1%, neutral reviews remained the same amount and negative reviews increased 1%. Similarly, positive reviews from New York decreased 2% and negative and neutral reviews increased 1%.

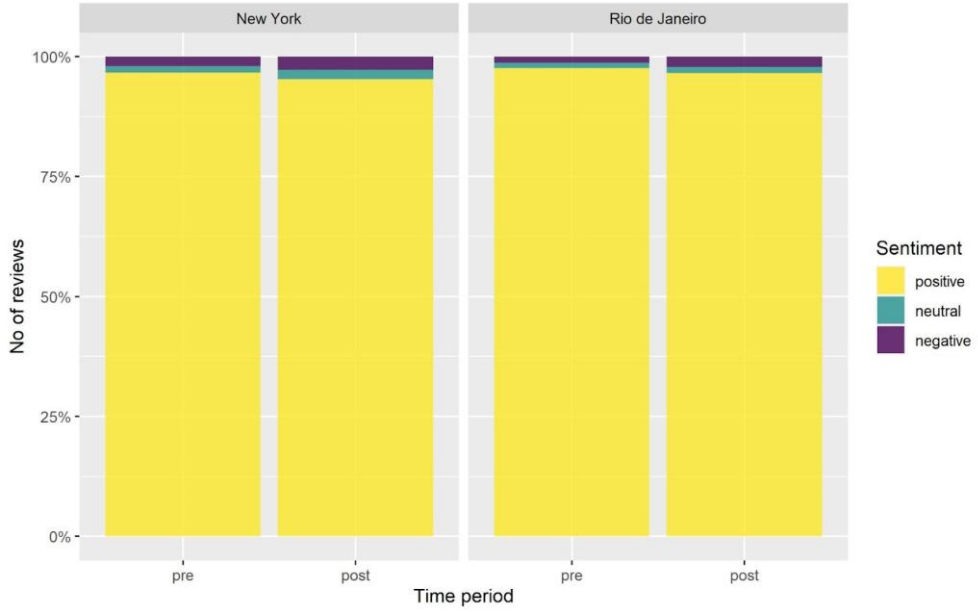

Figure 1. Percentage of positive, neutral and negative reviews from listings in NY and RJ before and after the outbreak of covid.

3.2. Spatial Visualization

Figure 2 and Figure 3 show contrasting results. While in Rio de Janeiro hotspots of overall positive, neutral, and negative listings remained basically in the same area after the pandemic, in New York, a relevant amount of listings located in Brooklyn, received mostly negative and neutral opinions which created a new hotspot on this area.

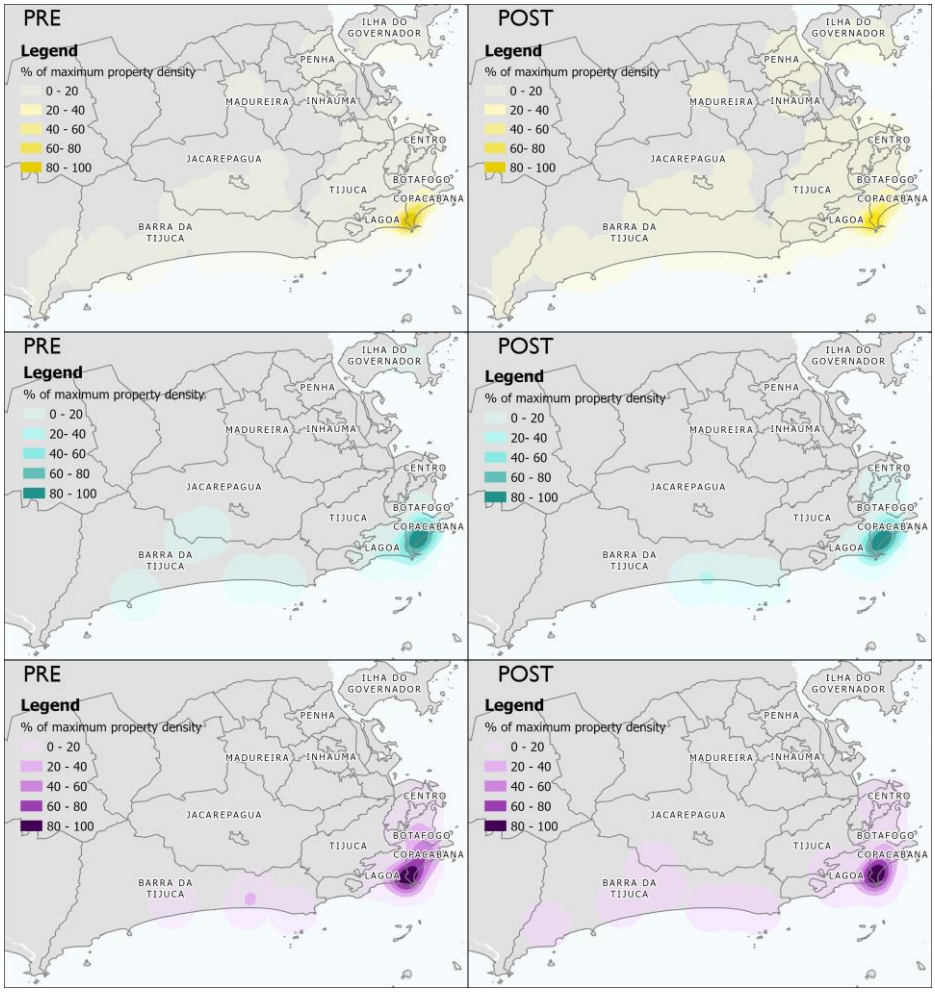

Figure 2. Spatial distribution of listings in Rio de Janeiro with overall positive (yellow), neutral (aquamarine), and negative (purple) polarity before and after the outbreak of Covid-19.

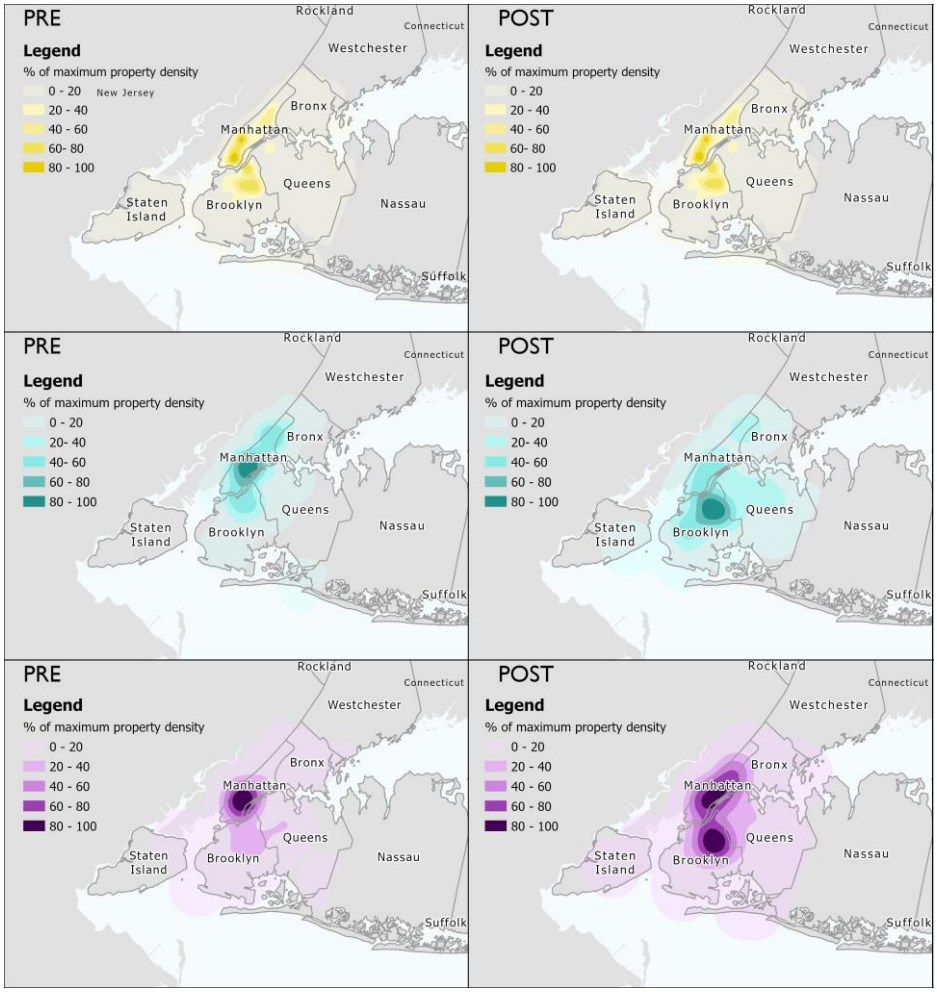

Figure 3. Spatial distribution of listings in New York with overall positive (yellow), neutral (aquamarine), and negative (purple) polarity before and after the outbreak of Covid-19.

4. Discussion and conclusions.

This research offers an alternative approach to the study of customer experience in tourism and hospitality literature. The methodology contributes by illustrating how spatial analysis can be combined with NLP techniques to visualize the role of location in customer experience during health crises. The findings show that after the pandemic, in New York, a new area of the city became a hotspot of neutral and negative reviews, which raises the question of why after the pandemic, in this area a relevant amount of listings experienced an increase in negative and neutral reviews and whether the characteristics of this area

could have negatively influenced the perception of customers. Thus, reviews from listings located in this area deserve further analysis, as they might reveal useful insights that lead to a better understanding of customer needs and expectations during health crises.

Results were contrasting, but also suggest that different areas of a city might play a new role in customer experience during health crises. However, at this stage, these findings can not be generalized, and therefore the need to extend this analysis to other cities including those that are not popular touristic destinations or that were not severely affected by the pandemic.

References.

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. 10.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism Management, 68, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Wang, S., & Chen, J. S. (2015). The influence of place identity on perceived tourism impacts. Annals of Tourism Research, 52, 16–28. https://doi.org/10/f7dgqj

Wyrwoll, C. (2014). User-Generated Content. In C. Wyrwoll (Ed.), Social Media: Fundamentals, Models, and Ranking of User-Generated Content (pp. 11–45). Springer Fachmedien. https://doi.org/10.1007/978-3-658-06984-1_2