NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

BSc THESIS

Sentiment Analysis in Tourism: Athens's vibe through online platforms

Dimitra M. Mavroforaki

Supervisor : **Maria Roussou,** Assistant Professor
Athanasia Kolovou, Instructional Lab. Personnel

ATHENS

FEBRUARY 2020

BSc THESIS

Sentiment Analysis in Tourism: Athens's vibe through online platforms

Dimitra M. Mavroforaki

S.N.: 1115201400104

SUPERVISOR: **Maria Roussou,** Assistant Professor
Athanasia Kolovou, Instructional Lab. Personnel

ABSTRACT

Tourism is a big worldwide industry and a major source of income, jobs and wealth in several countries. Particularly in Greece, which is considered to be one of the top global travel destinations, tourism is the keystone of economic growth and employment. Even during the recent economic crisis, tourism is one of the few sectors that has thrived. The local industry is currently focusing on expanding the tourist season and also improving the services provided. Therefore, more tourists will be attracted and Greece will maintain its advantageous position in the tourist market.

On an everyday basis, Greece makes significant investments on products and strategies that will make the holiday experience more personalized and appealing. The use of technology and Data Science is able to minimize the risk of such investments, to highlight and fulfill visitors' needs. Especially, by leveraging Data Mining and Big Data, scientists could analyze a huge amount of data and extract non-obvious results.

This thesis seeks to "listen to the pulse" of Greece's capital city, Athens. The main goal is to extract results from analyzing reviews and posts made by tourists on different platforms (Twitter, Google Maps, Foursquare, Airbnb). The above were collected, stored and combined into a single file that contained significant information, such as the review/post text, the origin, the location, the topic and the type of the mentioned activity. Each review/post was then processed and transformed into a vector. Afterwards, sentiment analysis algorithms and tools were used to define the sentiment of every data.

The results of the analysis are multidimensional. Diagrams and other visualization practices were used to represent the general feeling of tourists about the different districts and provided services of this city (food, accommodation, nightlife, museums/archaeological spaces, entertainment).

The outcome of the research proposes an effective way to detect both the city's inconveniences and highlighted facilities, based on visitors' opinion. In the future, it would be interesting to combine these data aspects with environmental indicators (e.g., gas emission, water consumption) during tourist periods to depict the overall impact on tourism. This way, the holidays' industry can provide more personalized, yet sustainable activities.

SUBJECT AREA: Sentiment Analysis, Data Visualization, Data Mining

KEYWORDS: big data, natural language processing, tourism, Athens, twitter, foursquare, airbnb, google maps

ΠΕΡΙΛΗΨΗ

Η βιομηχανία του τουρισμού αποτελεί παγκοσμίως μία βασική πηγή εισοδήματος, εργασιών και πλούτου. Συγκεκριμένα στην Ελλάδα, η οποία κατατάσσεται ανάμεσα στους κορυφαίους ταξιδιωτικούς προορισμούς, ο τουρισμός είναι θεμέλιος λίθος για την οικονομική ανάπτυξη και τις θέσεις εργασίας. Εν μέσω της πρόσφατης οικονομικής κρίσης, ο τουρισμός είναι ένας από τους λίγους τομείς που κατάφερε να επιβιώσει και να ευδοκιμήσει. Επί του παρόντος, η τοπική βιομηχανία εστιάζει στην επιμήκυνση των τουριστικών περιόδων και στην βελτίωση των παρεχόμενων υπηρεσιών. Κατ' αυτόν τον τρόπο η Ελλάδα θα προσελκύσει περισσότερο κοινό με αποτέλεσμα να διατηρήσει την θέση της στην αγορά.

Σε καθημερινή βάση, η Ελλάδα επενδύει σε προϊόντα και στρατηγικές που οδηγούν σε πιο εξατομικευμένη και ελκυστική εμπειρία διακοπών. Η χρήση της τεχνολογίας και της επιστήμης των υπολογιστών ελαχιστοποιεί το ρίσκο των επενδύσεων και ταυτόχρονα, επισημαίνει και ικανοποιεί τις ανάγκες των επισκεπτών. Ειδικότερα, μέσω των τομέων της Εξόρυξης Δεδομένων και των "Μεγάλων" Δεδομένων, οι επιστήμονες μπορούν να αναλύουν τεράστιο ποσό δεδομένων και να εξάγουν μη προβλέψιμα συμπεράσματα.

Η παρούσα πτυχιακή εργασία επιδιώκει να αφουγκραστεί τον "παλμό" της ελληνικής πρωτεύουσας, της Αθήνας. Ο κύριος στόχος είναι η εξαγωγή δεδομένων από την ανάλυση κριτικών/αναρτήσεων που έχουν συνταχθεί από τουρίστες σε διάφορες πλατφόρμες (Twitter, Google Maps, Foursquare, Airbnb). Τα παραπάνω στοιχεία συλλέχθηκαν, αποθηκεύτηκαν και συνδυάστηκαν σε ένα αρχείο, το οποίο περιέχει σημαντικές πληροφορίες για τις κριτικές/αναρτήσεις όπως για παράδειγμα το κείμενο, την προέλευση, την τοποθεσία, το θέμα και τον τύπο της δραστηριότητας που περιγράφεται. Κάθε κριτική/ανάρτηση μεταποιείται και μετατρέπεται σε διάνυσμα. Έπειτα αλγόριθμοι και εργαλεία ανάλυσης συναισθήματος χρησιμοποιούνται για να καθορίσουν το συναίσθημα κάθε δεδομένου.

Τα αποτελέσματα της ανάλυσης είναι πολυδιάστατα. Διαγράμματα και άλλες πρακτικές οπτικοποίησης χρησιμοποιούνται για να παρουσιάσουν τη γενική αίσθηση των τουριστών για τις διάφορες περιοχές και παροχές αυτής της πόλης (φαγητό, διαμονή, νυχτερινή ζωή, μουσεία/ αρχαιολογικοί χώροι, διασκέδαση).

Η έκβαση της έρευνας είναι ένας αποτελεσματικός και βασισμένος στην κοινή γνώμη τρόπος να ανιχνευτούν προβλήματα αλλά και διευκολύνσεις της Αθήνας. Στο μέλλον, θα ήταν ενδιαφέρον ο συνδυασμός αυτών των δεδομένων με περιβαλλοντικούς δείκτες (π.χ. εκπομπή καυσαερίου, κατανάλωση νερού) κατά τη διάρκεια τουριστικών περιόδων για να απεικονίσουμε το συνολικό αντίκτυπο του τουρισμού. Συνεπώς, η βιομηχανία των διακοπών θα παρέχει προσωποποιημένες αλλά ταυτόχρονα βιώσιμες δραστηριότητες.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ανάλυση Συναισθήματος, Οπτικοποίηση Δεδομένων, Εξόρυξη Δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: "μεγάλα" δεδομένα, επεξεργασία φυσικής γλώσσας, τουρισμός, Αθήνα, twitter, foursquare, airbnb, google maps

ΕΥΧΑΡΙΣΤΙΕΣ (ή ACKNOWLEDGMENTS)

For their valuable guidance, feedback and motivation I would like to express my gratitude to my supervisors, Dr. Maria Roussou and Ms. Athanasia Kolovou. They both helped me to overcome the obstacles and to implement this thesis. I would also like to say a special thanks to my ex colleagues (Sociality - Cooperative for Digital Communication). Their interesting projects and use of technology with a social oriented approach constitutes my personal inspiration.

CONTENTS

LIST OF FIGURES

LIST OF IMAGES

LIST OF TABLES

PREFACE

The basis for this research originally stemmed from my will to combine data science with everyday issues and to develop methods, tools, and applications that make a social impact. As the world moves further into the digital age, society is under continuous transformation. Social media and online platforms determine the status quo and can form public opinion. From my point of view, studying and processing the online information provided by people are effective and interactive ways to improve the standard of living. The current thesis was conducted at the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens from May 2019 to February 2020 and was supervised by Assistant Professor Maria Roussou and Instructional Lab. Personnel Ms. Athanasia Kolovou.

1. INTRODUCTION

This section summarizes the topic and objective of the current thesis, the motivation, the methodology of the research carried out as well as the overall structure.

1.1 Topic

The main objective of this thesis is to gather the general sentiment of visitors in Athens by exploiting tourism content shared on online platforms. The idea comprises the study of real data about tourism, the discovery of sentiment and insights from them and finally, their visualization. The steps of the research include the data collection (reviews/posts), the data preprocessing and transformation and the algorithms' application (i.e. topic modeling and sentiment analysis algorithms) on the refined data. A combination of tools and algorithms is used in order to intersect the output and to approach more realistic results. Finally, the outcome is visualised with diagrams and maps. The graphs provide answers to a series of questions and depict the satisfaction of visitors in Athens.

The project is implemented on Python 3.7.3 with the use of python libraries. Google Places, python-twitter and foursquare are wrappers that provide a pure Python interface for the corresponding APIs. During the preprocessing stage, *langdetect*, *ast*, *re*, *stop_words* and *nltk* libraries eliminate non-english and redundant elements. As far as algorithms are concerned, topic modeling utilises *gensim* modules and sentiment analysis combines nltk.sentiment.vader (ready sentiment analysis tool) and *sklearn* (library with custom vectorizers and classifiers). A library of offline reverse geocoding (reverse_geocoder) and a service that uses third-party geocoders (*geopy*) are responsible for the conversion of coordinates to districts. The third-party geocoder in this case is *Nominatim*, a geocoder for open source OpenStreetMap data. The visualization process is also achieved using python libraries. A word cloud generator (word_cloud), a data visualization library (*seaborn*), a tool that creates JS Leaflet maps (*folium*) and a 2D plotting library (*matplotlib*) are used to form the diagrams and the plots.

Despite the essential components, many helping functions and modules were used. *CSV*, *JSON* and *pickle* libraries manage the generated files, while pandas, collections and operator handle the data structures. Python system libraries (sys, os, time) complete the project's environment.

1.2 Objective

The objective of this thesis is to extract the general sentiment of tourists who have visited different districts of Athens and the quality of its services (food, accommodation, nightlife, museums/archaeological spaces, entertainment). The research marks the facilities and the problems of the city and endorses the exploitation of the data from online platforms. The project aims to approach social monitoring practices and highlights an alternative way of communication with the public. It could become a reference to improve the city's structures and a motivation to initiatives oriented towards smart cities.

1.3 Motivation

The main inspiration of this thesis originates from a project[1] undertaken by the cooperative that my internship took place. The objective of this project was to design and construct an integrated system of spatial decision-making and participatory planning using strategies and technologies such as ppGIS, IoT and Social Media. The part of Social Monitoring was, at the time, in the state of blueprints. The aim was to collect data from social media, analyze them and mark the highlights and the complaints of the inhabitants of Athens.

Moreover, as a resident of Athens, I was intrigued to observe the global approach to the Greece's capital; therefore in this thesis, I have focused on the point of view of the tourists and how they express their opinion on online platforms. The data that the original project has to process are in the greek language because they are composed by locals. Due to the incompatibility of most algorithms to non-english data, work with foreign reviews, is also an advantage.

Finally, I noticed there was a lack of similar projects, namely data analysis and visualization of Athens. The existing data had not been collected or analyzed before. Through this work, I attempt to extrapolate both predictable and unpredictable results that had not emerged so far and could possibly contribute to improving the travel experience.

1.4 Thesis Structure

The followed methodology during research as well as the organization in chapters are based on global prototypes in an attempt to achieve the best distribution of information.

1.4.1 Methodology

Sentiment analysis and data/information visualization methodologies guide this research.

- Sentiment analysis methodology:

  In the first part of my research, sentiment analysis methodology is the means to convert unstructured data to information. There are specific steps to analyze sentiment data as seen in Figure 1 [33], namely:

  1. Data collection
  2. Text preparation
  3. Sentiment detection
  4. Sentiment classification
  5. Presentation of output

Figure 1: Sentiment analysis methodology [33]

[1] https://ppcity.eu/

The first step is **data collection**. This process is either based on a database or it is a real time collection from public forums, blogs and social network sites. The gathered elements are stored and then processed (text preparation) because opinions and feelings are expressed in ways that are making the data huge and disorganized. **Text preparation** is filtering the data before analysis through elimination of non-textual or irrelevant content. Typical pre-processing procedure includes tokenization, stemming, lemmatization, stop-words removal and lowering the case of the letters [30]. The third step pertains to **sentiment detection** and examines the subjectivity of each element which leads to the sentiment classification. During **sentiment classification**, each subjective sentence detected is grouped by its polarity (negative, neutral, positive). The final stage is to **present the output**. The meaningful information extracted by this process is displayed on graphs and diagrams.

The recommended programming languages to perform Sentiment Analysis is Python and R, an open source tool for statistical computing and graphics. In this case, I have chosen to work in the Python environment. Python is suitable for text processing while R is preferred when data are in other forms. Moreover, Python has a variety of libraries and small supporting tools. Most of the researchers working in sentiment analysis use Python, therefore there is an active community as well. In conclusion, Python is a familiar tool and my personal preference.

- Data/Information visualization:

  The approach of designing visualizations can be divided into six steps [27] [22]:

  1. mapping
  2. selection
  3. presentation
  4. interactivity
  5. usability
  6. evaluation

  The first step is **mapping** and it determines how to encode information into visual graphics. Mapping is the most efficient translation of data objects into visual objects. This process contributes to the understanding of the offered information and prepares the ground for the next step. **Selection** means choosing the appropriate data among every element according to the given task. It is essential to define what questions should be answered and then, select the proper data for this. After mapping and accurate selection of data it is important to present the information in the most suitable and understandable form (**presentation** stage). According to Tufte, the most important characteristics of an effective graphical form is accuracy, several levels of details, cohesion and clarity [31]. The right decision in terms of presentation allows greater **interactivity**. User friendly interactivity gives the user the opportunity to explore the information easily. The next step in the visualization process is **usability**, namely taking into consideration human factors. The graphic plots are addressed to people from different backgrounds, so the visualization must be universally comprehended. The final step is the **evaluation** of the created form to find out whether it was effective or not. In this project, evaluation forms an aspect of future work.

1.4.2 Organization

The thesis is organized in the following sections:

- Section 2 contains the theoretical background needed to implement this project. It describes the relation between data science and tourism. Specifically, the chapter includes the definitions of machine learning and sentiment analysis as well as use cases of these fields into the tourism industry.
- Section 3 contains similar projects that my thesis was based on and their impact on the community.
- Section 4 is the main part of the thesis. It describes the implementation phase. The chapter is divided into four parts, which reflect the stages of the procedure followed. The first part includes the data collection, the types and the origins of the gathering elements. In the second part, the data preprocessing is analyzed. Each step is described with technical details, so that the reader can comprehend the reason of existence. The third part contains the algorithms applied onto the transformed data. It describes the libraries used and the result of each tool and algorithm. The final part is the data visualization. Diagrams are displayed and annotated according to the correlation between statistics and social reality.
- Section 5 unit concludes the thesis and describes future additions that could potentially lead to interesting results.

2. DATA SCIENCE IN TOURISM

The study of information related to tourism is generally known as "Tourism Informatics" and include research in the field of Social Informatics that can be applied in the tourism industry, such as the number and the expenditure of travellers in each country. Information is important for tourism, however there are few businesses that are aware of the power of data and how to use it [21]. Recently, the number of tourism researchers has been increased and consequently there is an augmentation in the quantity and the quality of research. Big data analytics and machine learning have become critical tools to process big chunks of data and to possibly satisfy visitors' needs [24].

2.1 Machine Learning

2.1.1 Introduction

Machine Learning (ML) systems have the ability to automatically learn from the environment and apply that learning to make better decisions. ML algorithms study and improve data in order to foresee better outcomes. Firstly, specific patterns are observed in the unstructured data and afterwards these patterns are processed in an attempt to discover useful information. Therefore, machines can imitate the learning processes of the human brain and "learn" to be better. Machine Learning is a tool that facilitates visitor - provider relation and enriches the whole travelling experience, so marketers have started to incorporate it in the tourism market (Figure 2) [37].

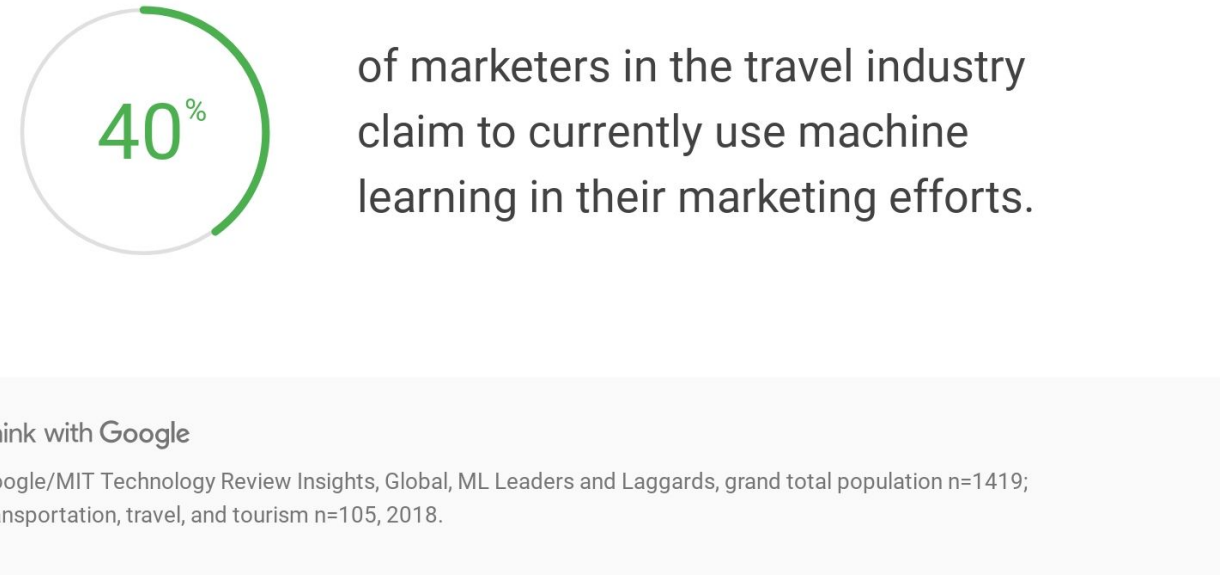


Figure 2: Incorporation of machine learning in tourism industry

2.1.2 Use cases

1. **Personalized experience**:

   Personal preferences from person to person differ, so customers' needs are almost impossible to be satisfied individually [25]. Via ML algorithms, the target group (visitors) are partitioned into subgroups that share similar characteristics, and consequently related demands and interests. Through this process, a much more specialised service can be offered. Moreover, applying ML algorithms will help travel providers to understand how customer experience should be designed. Specifically, it highlights the customers' lifecycle, their requests and their most basic needs and improves the travelling management [26].

2. **Competitive pricing**:

   Features like seasonality, events, location are analyzed through predictive models and are used to offer the best possible prices for any service, from booking a flight to renting a car [34].

3. **Custom recommendations**:

   Due to recommendation engines, the travel agencies can propose suitable travel packages to all their customers. Personal data (budget, interests, demands) are collected and processed to provide users their best experience. It is considered a very promising application of Machine Learning because it offers a variety of services (alternative travel dates, attractions, venues) based on user's search and choices [24].

4. **Route optimization**:

   Trip planning benefits from machine learning algorithms because minimum cost, time and distance can be predicted and achieved [24].

2.2 Sentiment Analysis

2.2.1 Introduction

Sentiment analysis is the process of identifying the subjective information in text and classifying each piece of data as positive, negative or neutral. Through sentiment analysis, unstructured information transforms into structured data of expressed opinions, actionable insights come to light, and hours of data processing are saved. In the tourism sector, sentiment analysis supports large-scale examination of reviews and social web posts obtained from either professional websites (e.g., TripAdvisor, Booking, Airbnb) (Image 3) or social media (Twitter) about venues and experiences by detecting the texts' polarity [34].

2.2.2 Use cases

1. **Discovery of customers' preferences**:

   Sentiment analysis examines the social content published by an individual client in order to construct a customer's profile (needs, interests, likes, dislikes). These artifacts synthesize a "persona" which helps to identify and evaluate market opportunities and design user experiences that are focused on different customer categories/personas [32].

2. **Discovery of market opinion**:

   The decisions of a traveler often depends on the reviews posted on web platforms and websites. Modern booking websites proceed to sentiment analysis as climb descent indicator of their services, so that travel agencies, hotels and hostels seek to cooperate with them [24].

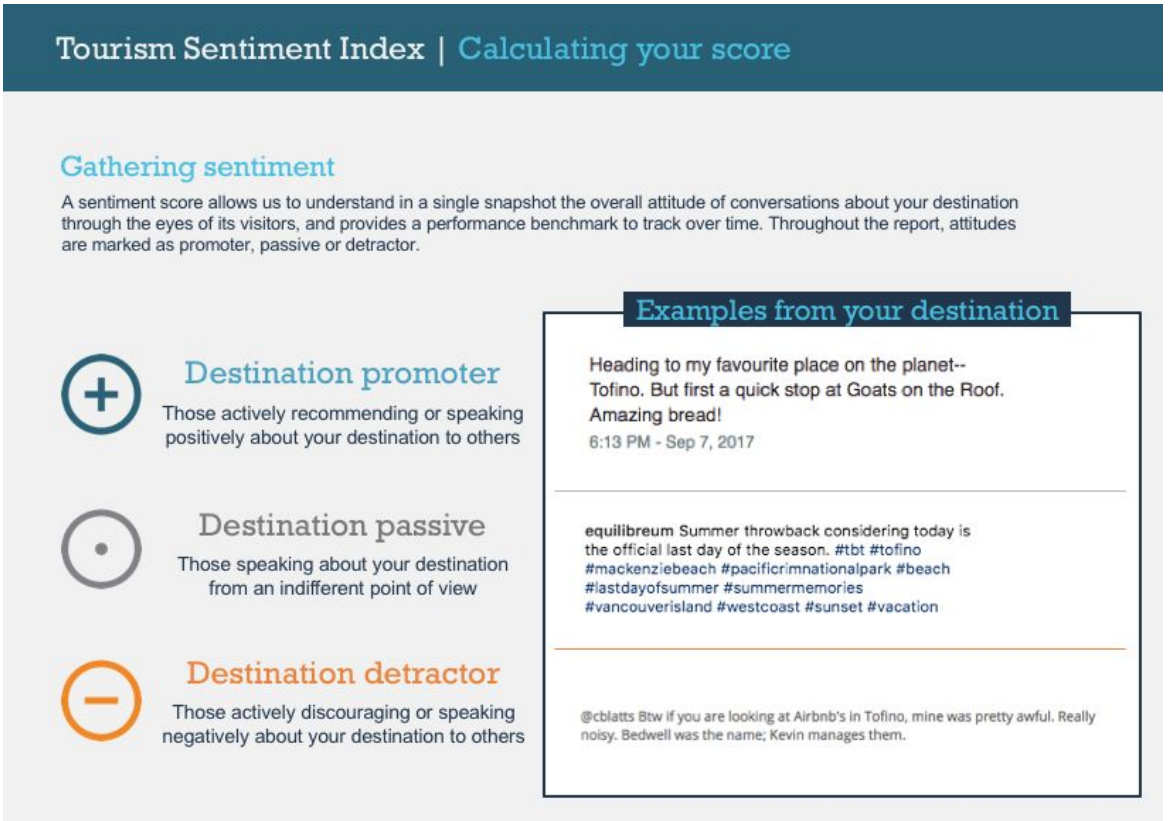
Image 1: Sentiment analysis using data from Tripadvisor platform

3. RELATED WORK

This thesis is based on both similar open source research projects and commercial products. The common ground between these projects and my work is the sentiment analysis on data collected from digital platforms, with the primary aspiration to extract results and statistics for popular tourism destinations. A mentionable fact is that the city of Athens has only been included in a single resemblant project.

1. **Tourism sentiment index**[2] (commercial product) :

   The Tourism Sentiment Index (TSI) captures the sentiment of word-of-mouth in an attempt to monitor the success of a destination. With the aid of numerous online peer-to-peer communication platforms, including all major social network and reviews sites, TSI examines conversations happening around a certain destination, a view that cannot be provided by any survey. Firstly, the words and images related to the desired destination are collected and classified into tourism categories. Next, the content is categorized into positive, neutral or negative. Depending on the results, the destination's score is calculated with the methodology shown in Image 2. The methodology behind the Tourism Sentiment Score excludes the not relevant to tourism conversations from the end result. Finally, the data are summarized into a turnkey report that reveals how people feel about different experiences and services in the wanted destination. This product is an example of the potential commercial opportunities of this thesis, in case specific coordination and resources are given.



[2] https://destinationthink.com/about-tsi/ (Last accessed January 23, 2020)

Image 2: Tourism Sentiment Score calculation

2. **Natural Language Processing (NLP) on Berlin Airbnb Data** (open source project) [38]:

   According to this project, Berlin is one of the hottest markets for Airbnb in Europe. The collected features from detailed Berlin datasets, sourced from the Inside Airbnb website, answers the main question of this research, namely "what do visitors like and dislike". The data are preprocessed and then visualised using Word clouds. Despite the words' frequency, the project focuses on discovering the sentiment score of each review and investigating the positive and the negative comments. In this project, even though accommodation is one of the examined parameters, the extracted data answers more generic questions that also apply to the rest tourist dimensions.

3. **Airbnb: The Amsterdam story with interactive maps** (open source project) [23]:

   In this case, Amsterdam constitutes the point of interest. Using Airbnb reviews and data, the researcher is able to identify the neighbourhoods that hold most listings (Figure 3), the most popular type of properties (Figure 4) and the most frequent number of people accommodated. No other city in the world is cracking down so radically on mass tourism. Consequently, it is interesting to observe the effect of the city's restrictions depicted on the gathered data. Moreover, this research contributes to the municipality as it facilitates the discovery of possibly illegal hotels as well as the detection of hosts with numerous estates. The project includes information useful to potential visitors, such as the average daily price per neighbourhood, the neighbourhoods' safety and the hosts' scores. Interactive maps and detailed description of the current law in the Dutch capital completes the overall analysis. Its overall approach is a guideline for this thesis implementation.

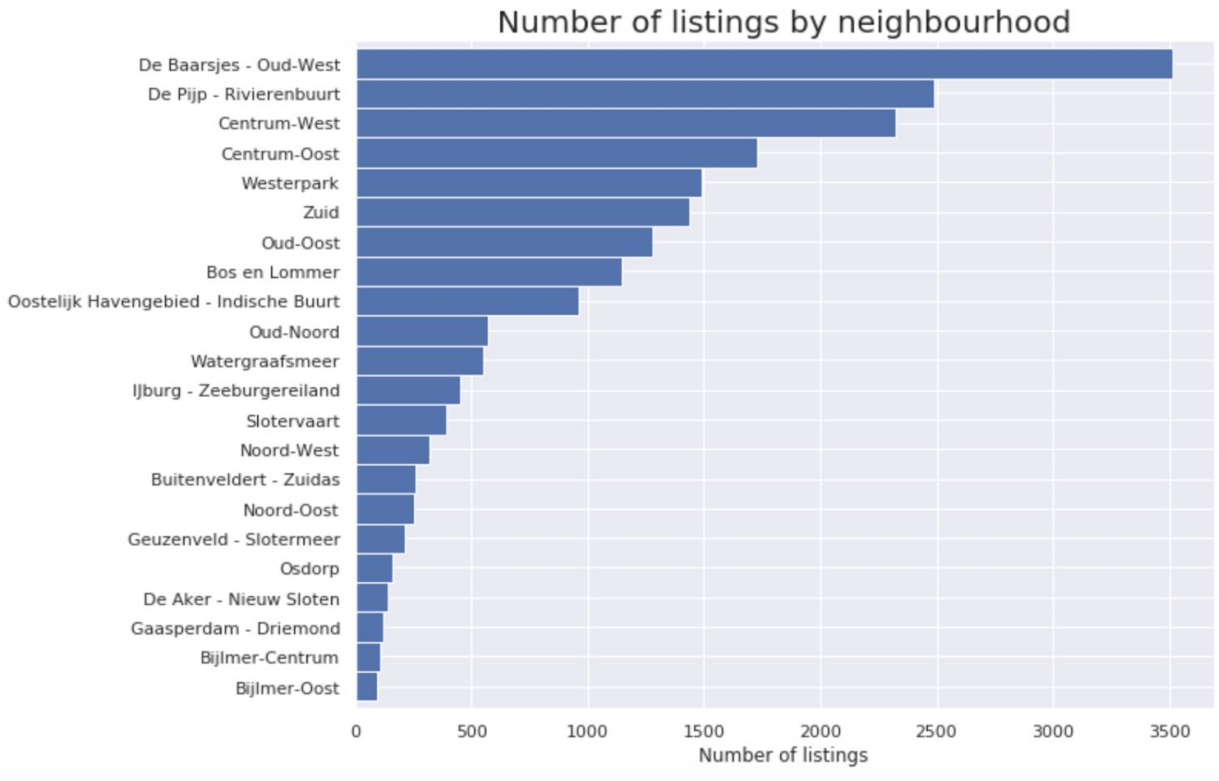


Figure 3: Number of listing per neighbourhood in Amsterdam city

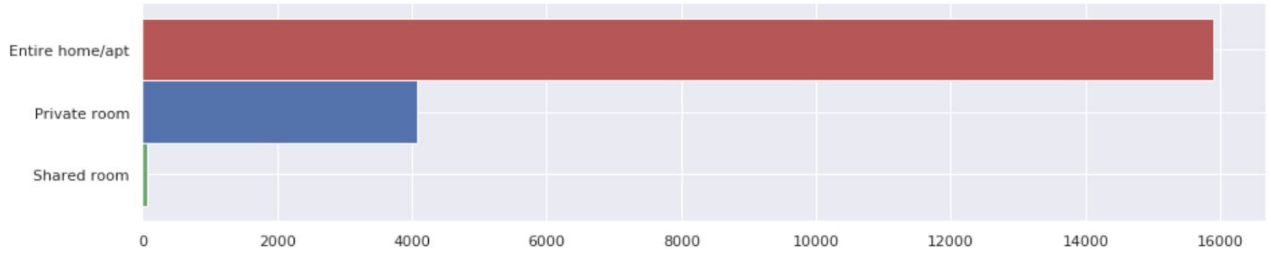


Figure 4: Most popular type of properties in Amsterdam

4. **NLP - Twitter Sentiment Analysis Project** (open source project) [28]:

   This project applies sentiment analysis on Twitter to identify racist/sexist content. The tweets acquire binary value, where 1 represents racist/sexist content and 0 represents non racist/sexist content (Figure 5). Twitter consists a main data resource in this research as well. Consequently, similar projects contribute to form an idea of the way that tweets should be handled during sentiment analysis process.

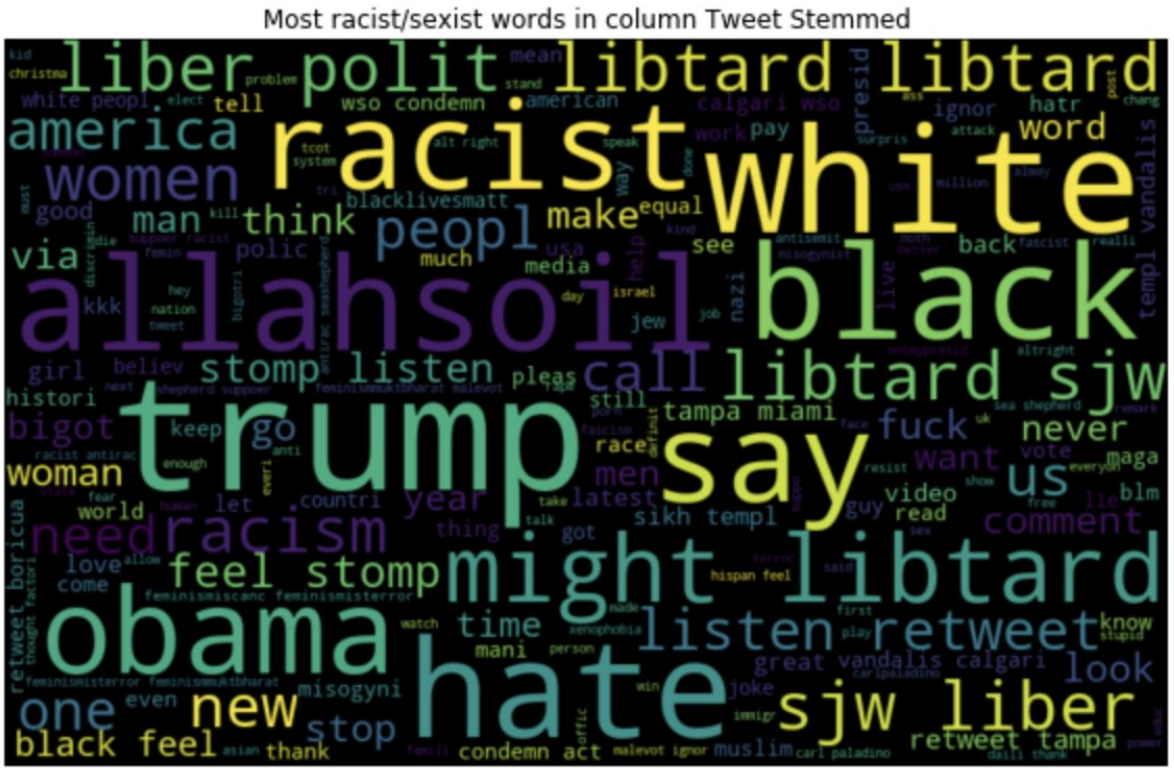


Figure 5: Most racist/sexist words from NLP Twitter Sentiment Analysis Kaggle project

5. **TripAdvisor Reviews - Data Analysis** (open source project) [29]:

   This research comprises an example of the possible results produced by the combination of TripAdvisor data. The input includes information about different types of restaurants in European countries. Diagrams, plots, pies and maps represent the scope of European food and cuisines (pricing, rating per city, quantity,

cuisine styles). The specific project not only handles the data effectively, it simultaneously visualizes them in the most user friendly way. As a result, it constitutes the personal inspiration of specific diagrams and data visualization techniques.

4. IMPLEMENTATION

This chapter describes the implementation of sentiment analysis on the collected data set, as well as the visualization approach taken. In particular, the process of implementation presented below includes data collection, data preprocessing, the application of algorithms onto the transformed data, and data visualization.

4.1 Data Collection

The first step in the process of sentiment analysis is the "data collection". In my case, the collected data pertain to travelling experiences and tourist reviews pertaining to Athens. The main concerns are the origins of the data and the types of entities.

4.1.1 Origins of data

The criteria which determined the selected websites from which data would be drawn were the content and the policy of each online platform. Consequently, data from Twitter[3], Airbnb[4], Foursquare[5] and Google Maps[6] have been included (Image 3). A posts' collector has been coded for each of the above online services.

Image 3: Platforms' API logos

- **Twitter:**

  People send tweets for all sorts of reasons; mainly as a way to express their feelings regarding a situation. A growing number of Twitter users send out useful

[3] https://twitter.com/?lang=el
[4] https://www.airbnb.gr/
[5] https://foursquare.com/city-guide
[6] https://www.google.gr/maps

content, and that's the real value of Twitter. However, there is also a lot of drivel, thus extracting results from Twitter's data is a questionable process [18]. Twitter's posts are approximate and contribute to the general sentiment of Athens as a travel destination, and are gathered using *twitter*[7] Python library. The body of the selected and gathered tweets contains words as districts of Athens or hashtags (#athenstravel, #loveathens) etc. "Twitter's developer platform"[8] provides many API tools and resources that enables the researchers to harness the open and real-time communication network. The endpoints of the standard API allow to retrieve and engage with Tweets and timelines without any special access requirements. However, it requires authentication. Any potential developer needs to create a Twitter developer account and afterwards, a new project. The approval of this process will generate an app API key that is an essential part of the request. Rate limits are divided into 15 minute intervals.

- **Airbnb:**

  Airbnb has successfully disrupted the traditional hospitality industry as more and more travelers decide to use Airbnb as their primary accommodation provider.

  Insideairbnb.com[9] is a website on which web scraped datasets of "snapshots" of cities are published. It provides listings and reviews from June 2019. The combination of these datasets may render the accommodation' s rating in Athens.

- **Google maps:**

  Google is the most used resource to search for places, businesses, and products. It would be a deficiency to disregard the reviews from the platform where the majority of people are actively looking for information. Google maps disposes Places API[10]. The process to acquire an API key is free and the service provides a maximum of 1000 requests per day. The goal is to collect reviews from points of interest in Athens using *googleplaces*[11] Python library. Firstly, "places" have been defined in order to accumulate "places ids" [19]. Subsequently, it is possible to collect the details of each place; reviews are part of the details [20]. Google places API returns the five most helpful reviews for each place.

- **Foursquare:**

  the Foursquare application focuses on finding new places for a user to visit, based on searches or personalized recommendations. The main website's content comprises of a list of events/establishments on a location (venues). An amount of reviews (tips) corresponds to each venue and contains the information needed. Library *foursquare*[12] is a Python wrapper that collects the reviews by exploiting Foursquare's API. Foursquare has at its disposal a free API[13] that allows maximum 950 calls per day. Foursquare API provides the two most popular/liked reviews for each venue.

[7] https://pypi.org/project/twitter/
[8] https://developer.twitter.com/en.html
[9] http://insideairbnb.com/
[10] https://developers.google.com/places/web-service/intro
[11] https://github.com/slimkrazy/python-google-places
[12] https://github.com/mLewisLogic/foursquare
[13] https://developer.foursquare.com/

Taking into account the platforms' API policy and the irrelevant data and gibberish that users post on social networks, data collection is considered a difficult and time consuming procedure. In our case, these difficulties were confirmed. The daily data gathering was minimal and in some platforms could not be automated. In the case of Foursquare and Google maps, the most effective way to collect data was through a particular radiant with given coordinates. The developer must then either change the coordinates manually or find an efficient algorithm to generate random coordinates that could successfully scout the entire district of Athens.

4.1.2 Type of entities

The selection of the entities to collect comments/reviews took place according to the criteria that define whether a city has been touristically developed. The received information is divided into categories (entertainment, museum, food, nightlife), despite Twitter posts which has no category and Airbnb data that are classified as accommodation.

Reviews/posts from each platform have been stored in different local folders in the form of JSON files. When the amount of data was deemed as sufficient, the data collection process stopped and the generated JSON files were combined in a single CSV file. The CSV file has been preprocessed and contains the fields listed in Table 4.1:

Table 4.1: Type of entities

Field	Explanation
id	Original post's identifier. It operates as a reference to the json file that contains the particular post; each filename consists of the retrieved post's id.
origin	The origin of the post. It is essential for statistical reasons.
date	The date that the post was written, if it is included (optional).
review_text	The main information of the posts, which is processed and classified according to its sentiment.
clean_text	The review's text after the data preprocessing.
rating	The original text's rating, if it is included (optional).
location_lat	The location's latitude of the related

	review.
location_long	The location's longitude of the related review.
venue_category	The specific category of the venue (i.e. apartment, theater, bar) (optional).
type	The general type of the described venue. (i.e. entertainment, food) (optional).
topic	Three keywords that specify the text's topic
address	The location of the related review.
sentiment	The sentiment (negative = -1.0 , positive = 1.0, neutral = 0.0) as occurred from the sentiment analysis algorithm or the given rating.

The choice of the above fields is not random and serves particular reasons. The fields defined synthesize the data of each platform successfully, leading to an integrated whole. Moreover, they present appropriately the collected information, thus contributing to the process of visualization.

4.2 Data preprocessing

A significant part of the collected data are gibberish or not relevant to the topic of the project but cannot be eliminated during data collection. Data preprocessing takes place in order to eliminate the irrelevant elements that could possibly alter the result of the research. By the end of this process, the data become compatible input for the data analysis algorithms.

4.2.1 Language detection

The Airbnb dataset contained an amount of non-english reviews which needed to be excluded from the preprocessed data. Using *detect* function from the *langdetect*[14] library, each review that was not exclusively english text was deleted from the set.

[14] https://pypi.org/project/langdetect/

Image 4: Language detection in animation

4.2.2 Cleaning reviews/posts

The reviews' and posts' body includes many words and characters that are impossible to analyze in algorithmic level and, consequently, should be removed. Links, stopwords, punctuation symbols, emojis are ignored and the remaining text is transformed into lowercase. For example, after the cleaning procedure, the review's text "Amazing places to visit !!! http://… " is converted into "amazing places to visit".

4.2.3 Tokenization

Tokenization is the process of demarcating and possibly classifying sections of a string of input characters [17]. Tokenization constitutes an essential step in lexical analysis. Texts need to be converted from uniform strings into tokens because each word will be transformed into a vector in order to become an acceptable input for machine learning algorithms (Image 5). The division into multiple tokens is, moreover, appropriate for lexical statistics (topic modeling, calculation of words' frequency, creation of word clouds etc.). In Python, tokenization is implemented via modules of *nltk*[15] library. Using the above example, "amazing places to visit" is transformed into an array with content ["amazing", "places", "to", "visit"].

[15] https://www.nltk.org/

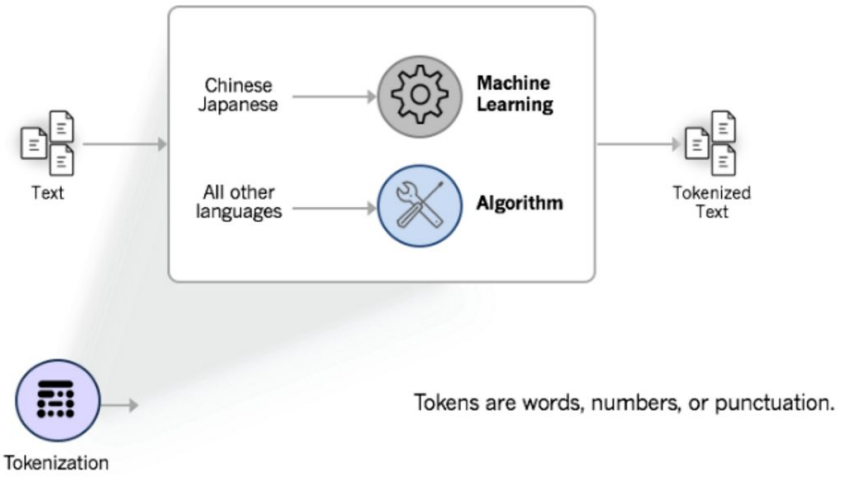


Image 5: Tokenization process

4.2.4 Lemmatization

In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning [17]. Lemmatization for English language is relatively simple but not a manual process, because hand-written lemmatization rules can be troublesome. *Nltk* library uses machine learning techniques to help developers with the stemming process. The algorithms' input does not include only words. Additionally, it includes metadata such as POS tag [12]. POS tagging labels the words according to a set of tags, known as tagset, that usually contains parts of speech (noun, adverb, verb etc.). This procedure helps to determine the correct lemma for ambiguous forms [10][15][16].

Lemmatization contributes to topic modeling, due to the fact that it depends on the distribution of content words that need to be consistent across documents. Furthermore, it is important for training word vectors. In the same example, ["amazing", "places", "to", "visit"] contains ["amazing", "place", "to", "visit"] after lemmatization. An additional example is depicted in Image 6.

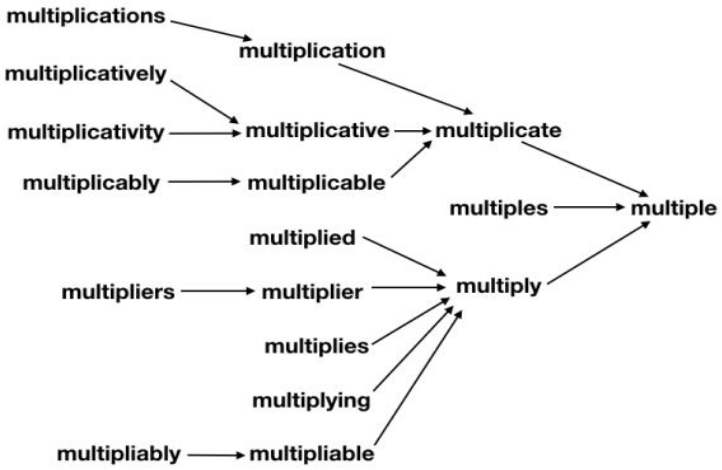


Image 6: Example of lemmatization that ends up to the word "multiple"

4.2.5 Converting coordinates to districts

Following lexical analysis, the final step of data preprocessing is to convert coordinates into matching districts. The coordinates of the reviews are part of the response in every online platform except from twitter. Tweets does not refer to a particular venue or location consequently they do not provide specific latitude and longitude. Due to data consistency, each tweet has been combined with the location mentioned in the tweet's body. If no location is mentioned in tweet' s text, a default set of coordinates is assigned.

Having assured that every post contains a single position on map, the coordinates are converted into the location name. Location names are used during data visualization as a factor in diagrams and graphs. Due to the large amount of data (approximately 250 thousands) and restrictive API policies, the exclusive use of a library that depends on user requests to an online geocoding platform was impossible. As a solution, coordinates should be initially processed with an offline tool.

Reverse_geocoder[16] which is a Python library for offline reverse geocoding, linked coordinates with general locations. Parallelised implementation of K-D trees improved performance especially for large inputs, so the procedure was fast. However, the library classified the majority of the posts (approximately 150 thousands) under a generic designation (municipality of Athens). On account of that, a complementary library was needed.

Geopy[17] enables the users to find the coordinates of addresses, cities, countries and landmarks across the world using third-party geocoder and other data sources. This library particularized the district of Athens of the 150 thousands general located posts. Nominatim [18] was the chosen third party geocoder because it forms a search engine for OpenStreetMap data. OpenStreetMap[19] is an open source map so the requests' policy is lenient (1 request per second).

The last step of converting coordinates is to transform the results of the geolibraries to the official districts' names. Data visualization of city of Athens requires the corresponding geospatial data. Geoadata.gov.gr[20] is providing open geospatial data and services for Greece, serving as a national open data catalogue, as well as a powerful foundation for enabling value added services from open data. The most appropriate dataset [12] in case of this project, includes the districts' boundaries of the Greek capital. In fact, it consists of the municipal boundaries of the residential structure, based on the general urban plan of Athens.

[16] https://github.com/thampiman/reverse-geocoder
[17] https://pypi.org/project/geopy/
[18] http://nominatim.org/
[19] https://www.openstreetmap.org/#map=17/37.97560/23.73322
[20] http://geodata.gov.gr/content/about/

Image 7: Geocoding and reverse geocoding process

4.3 Algorithms application

When the data preprocessing procedure is completed, data are ready to be transformed into an appropriate format for algorithms application. Topic modeling and sentiment analysis algorithms contribute to the automatic extraction of information, taking into consideration both the diction and the content of each post/review.

4.3.1 Topic Modeling

Topic modeling is an unsupervised machine learning technique for abstracting topics from collections of documents or, in this case, for identifying which topic is being discussed in a review. It constitutes a usual technique in machine learning and natural language processing, as topic modeling is a frequently used text-mining tool based on probabilities. A document typically concerns multiple topics in different proportions. Statistical algorithms are responsible for discovering the most frequent words used and, deductively, the most used topics. The "topics" produced by topic modeling techniques are clusters of similar words. Based on the statistics of the words in each cluster, the topics and their balance in a specific document are clarified. Topic models can help to organize and offer insights. Furthermore, they are commonly used to understand large collections of unstructured text bodies [11]. In image 8, there is an example of how LDA is used on a real world example, specifically on New York Times articles.

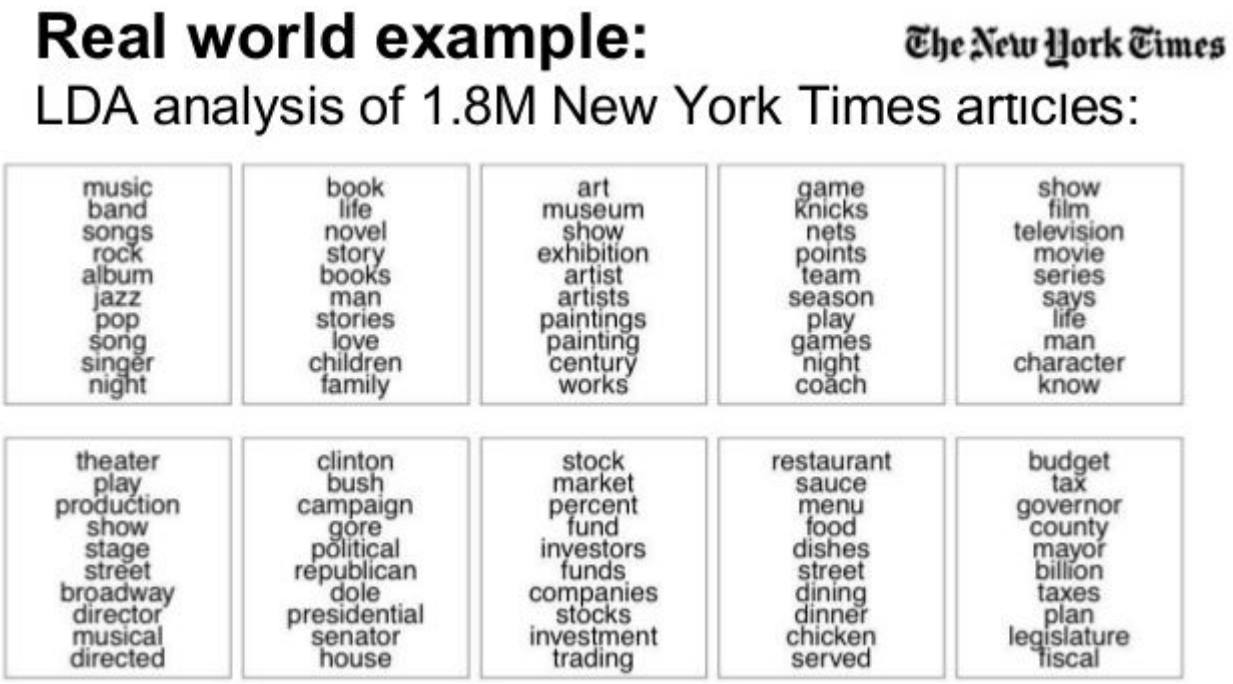


Image 8: Real world example of LDA on New York Times articles

LDA is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. LDA algorithm is part of *Gensim*[21] package. LDA's approach to topic modeling is it considers each document as a collection of topics in a certain proportion. And each topic as a collection of keywords, again, in a certain proportion. The algorithm rearranges both topics and keywords distribution to acquire the most optimized form. Finally, the topic is the collection of dominant and representative keywords.

The algorithm requires clean data. The two main inputs to the LDA topic model are the dictionary (id2word) and the corpus (the preprocessed reviews). Gensim creates a unique id for each word in the document and by the end of the procedure, a mapping of (word_id, word_frequency) has been created. This corpus constitutes the input of the LDA model. The LDA model is built by different amount of topics. Each keyword has a certain weight that depends on the topics. Eventually, the words that are chosen have the most weight. Due to the limited amount of words that each post/review includes, the number of topics is bounded, as well.

By the end of topic modeling, each review is related to the three most important included words. Given the airbnb review ['stelios', 'amaze', 'host', 'beside', 'fact', 'apartment', 'area', 'town', 'close', 'evreything', 'evrething', 'need', 'include', 'phone', 'stelios', 'make', 'sure', 'fine', 'give', 'recommendation', 'day', 'strike', 'athens', 'come', 'pick', 'airport', 'make', 'sure', 'wont', 'take', 'trxi', 'give', 'tour', 'around', 'city', 'find', 'evreything', 'evreything', 'need', 'apartment', 'amaze', 'host', 'amaze', 'place'], the application of LDA algorithm will result on ['amaze, evreything, apartment'] [14].

[21] https://radimrehurek.com/gensim/models/ldamodel.html

4.3.2 Sentiment Analysis

Sentiment analysis stage is implemented with the use of both custom classifiers and off-the-shelf tools. Consequently, a hybrid approach was followed for the posts' classification. A custom classifier using the result of two different vectorizers and an automatic tool, synthesized a set of three different sentiment analysis tools , one for each process. The dominant sentiment is chosen as the label of the particular post/review.

Custom classifiers

The sentiment analysis task is usually modeled as a classification problem where a classifier is fed with a text and returns the corresponding category, e.g. positive, negative, or neutral (in case polarity analysis is being performed). The first step is the training process. During this stage, the model associates a particular input to a specific output based on the test samples. The training is followed by the prediction process. In the prediction process, the feature extractor transforms unpredicted text inputs into feature vectors. These vectors are reprocessed and generate the predicted tags (positive, negative, neutral).

The model's input requires the text to be transformed into a numerical representation (vector). This process is known as 'feature extraction' or 'text vectorization'. There are many different algorithms that are suitable for this procedure e.g. *BOW*, *TF-IDF*, *bag-of-ngrams*. New feature extraction techniques include *word embeddings*, known as *word vectors*. This kind of representation makes it possible for words with similar meaning to have a similar representation, which could generally improve the performance of classifiers.

- **BOW**

  A bag-of-words is a representation of a text that describes the occurrence of words within a document. The main involved notions are a vocabulary of known words and a measure of the presence of known words. As implied by its name, bag-of-words does not take into account any information about the order or structure of words. The only concern is the word itself and not its location in the document. As a result, documents are similar if they have similar content.

  Once a vocabulary has been chosen, the occurrence of words needs to be scored. This procedure will eventually lead to the creation of a vector. Bag-of-words vectors are binary. Zero value is applied if a word is absent in the vocabulary and unit value is applied if the word is present. As the vocabulary size increases, so does the vector representation of documents. Generally, the length of the document vector is equal to the number of known words [6].

  In this project, bag-of-words vectorizer was implemented via CountVectorizer module of *sklearn* package. BOW vectorizer classified most data as neutral, so in terms of classification it had mediocre results. As it was expected, discarding word order ignores the context and semantics, which could offer a lot to the model.

- **TF-IDF**

  TF-IDF is a metric that represents how 'important' a word is to a document. Scoring word frequency has an occupational hazard. Highly frequent words start to

dominate in the document (e.g. larger score), but may not contain as much "informational value" to the model as rarer words. The vectorizer's concept is to monitor the number of times a word appears in a document taking into account the overall appearance in every document. Words that are generally common rank low, even though it may occur many times, since it is implied that they do not have importance to that particular document [5]. TF-IDF means Term Frequency - Inverse Document Frequency. Thus, the idf of a rare term is high, whereas the idf of a frequent term is low.

In this project, tf-idf vectorizer was implemented via TfidfVectorizer module of *sklearn* package. Tf-idf vectorizer lead to a better classification than BOW vectorizer, because the number of neutral identified data was not so high and better distributed.

- **Word embeddings**

  Word embeddings is another way to depict each word of a document as a vector. The main idea is essentially to represent words with similar meaning as similar vectors. In order to achieve this, the distributed representation is learned based on the usage of words.The vector values are learned in a way that resembles a neural network, and hence the technique belongs into the field of deep learning. There are multiple techniques used to produce a word embedding from text data [9].

  In my case I used *Word2Vec* and *Doc2Vec*. Word2Vec uses a simple neural network with a single hidden layer to learn the weights. The only important point is the weights of the hidden layer because they are responsible for the vector's form.

  Doc2vec model is based on Word2Vec with an additional vector. The word vector is a vector with a dimension 1xV. The document vector has a dimension of 1xC, where C is the number of total documents [7]. However, both vectorizers did not manage to classify the data successfully. A reason might be the data that were given as input to the vectorizers were not suitable for training them.

The classification stage usually involves a statistical model like *Naïve Bayes*, *Logistic Regression, Support Vector Machines*, or *Neural Networks*. Unlike traditional probabilistic and non-probabilistic algorithms, neural networks are associated with deep learning and consists of a diverse set of algorithms that attempts to imitate the brain's functionality by employing artificial networks to process information.

- **SVM**

  Support vector machines are widely used in classification tasks. Their main data structure is a hyperplane. Hyperplanes are decision boundaries that help classify the data points. Depending on the side of the hyperplane that the data points fall, it is decided in which class they will be assigned [8].

  The input of SVM is called sample features. It is a training set of data with predefined sentiment. Training data are vectorized and afterwards, is inserted into SVM paired with the corresponding sentiment label. In this project, training set for classifying the posts of each online platform differed. Moreover, the lookup for well

trained open source data were a difficult task. Trained data for Twitter[22] are 28061 tweets that are obtained from a university research within the framework of SemEval content in 2017. Regarding foursquare, the dataset[23] is composed of tips referring to the localities of the city of São Paulo/Brazil. The tips belong to the foursquare's categories: Food, Shop & Service and Nightlife Spot and contain 179181 tips. Before being inserted into the vectorizer, the tips were translated into english to achieve more accurate vectors. In general, a training set is preferred to have the same content as the test set and to contain a larger amount of information. Twitter and Foursquare data were processed via an SVM classifier, because datasets for Airbnb and Google Maps were unavailable. Google Maps reviews' sentiment has occurred via rating. Airbnb sentiment analysis was only handled by an automatic data analysis tool which is discussed below.

The output of the SVM includes a set of weights, one for each feature, whose linear combination predicts the value of the test set. The test set in this project is formed by the vectorized collected posts and reviews from online platforms. After the end of the classification stage, each post is matched with a predicted sentiment (-1 if negative, 0 if neutral, 1 if positive). The modules needed for SVM classification were imported from *sklearn* package.

During the evaluation stage the classifier's performance metrics are obtained, in an attempt to comprehend the accuracy of the sentiment analysis model. The most frequent evaluation method is cross-validation. The training dataset is divided into a certain number of training folds and testing folds. The training folds are used to train the classifier and the testing folds are used to test the classifier and obtain performance metrics. The process is repeated multiple times, until an average of each metric is calculated. The standard metrics to evaluate a classifier are *precision*, *recall* and *accuracy*. Precision measures the percentage of the correctly predicted texts of a particular category. Recall measures the percentage of the correctly predicted texts of a particular category, given the texts that should have been predicted in the specific category. This metric will be improved if the classifiers' input data are increased. Accuracy measures the percentage of the correctly predicted texts out of all the texts in the corpus. For a difficult task like analyzing sentiment, precision and recall levels are likely to be low at first and will increase as more data are given.

If testing set is always the same, it might lead to overfitting. Overfitting means over adjusting the analysis to a given dataset and consequently failing into analyzing any other dataset. Cross-validation contributes to prevent that phenomenon. In this project, classifier' s evaluation was not a main desideratum, however the results proved that overfitting was avoided. In case of overfitting the predicted values would be tampered, hence useless [34].

Automatic tools

In case there is no deep learning or NLP expertise there are some readily available pretrained tools. During sentiment analysis stage of the current thesis, a ready tool contributed to the intersection of the produced results from the custom classifiers. The final outcome is more enriched, hence accurate.

[22] https://github.com/cbaziotis/datastories-semeval2017-task4
[23] https://www.kaggle.com/thaisalmeida/tips-foursquare

- **VADER**[24]

  VADER is a lexicon and sentiment analysis tool that is specialized in sentiments expressed in social media. VADER uses a sentiment lexicon which is a list of features which are generally identified based on their semantic orientation as positive, negative or neutral. VADER accumulates a lot of advantages over traditional methods of Sentiment Analysis and other similar libraries. Firstly, it is considered successful when dealing with social media texts, editorials, movie and product reviews, due to the fact that it provides sentiment percentage. Sentiment percentage defines how positive or negative a review is. VADER does not require any trained data but is constructed from a generalizable, valence-based, human-curated standard sentiment lexicon. Moreover, VADER is fully open-sourced [13].

From a practical point of view, VADER tool is imported in a *nltk* module, thus can be integrated in the project's Python code. The function polarity_scores() obtain the polarity indices for the given sentence. Given the review, "The apartment was kind of good.", the method returns : **compound: 0.3832, neg: 0.0, neu: 0.657, pos: 0.343**.

The positive, negative and neutral scores represent the proportion of text that falls in these categories. In this case, the sentence is 67% neutral, 34% positive and 0% negative. The sum of the polarities must be 100%.

The compound score is a metric that denotes the accuracy of the result. A higher the compound score indicates a more precise vader prediction. Given the example above, the compound means that the neutral impact of this sentence is vague. In the reviews dataset of this project, compound score was very high, hence it was not taken into account.

By the end of VADER sentiment analysis in current project, most data were labelled as neutral. In some cases, the divergence between neutral and positive/negative sentiment was insignificant. Therefore, if the difference between the neutral and non-neutral percentage was less than 20%, the review was identified by the following dominant non-neutral sentiment.

4.4 Data Visualization

By the end of the algorithms' application stage, the extracted results should be visualized. Data Visualization is essential to depict the conclusions of the project and to answer the research' s question with a user friendly perspective. The main objective of this thesis is presented with the use of the appropriate diagrams, charts and maps.

4.4.1 Most frequent words in reviews/posts

Word clouds are a collection of words in various sizes. The more a specific word appears in a source of textual data, the bigger and bolder it appears in the word cloud. Consequently, this representation type is ideal ways to extract the most pertinent parts of huge chunks of textual data. Word clouds are used to depict the most frequent words in

[24] https://www.nltk.org/_modules/nltk/sentiment/vader.html

reviews/posts. In Python code, word clouds are generated using the *wordcloud*[25] package. Depending on the different platform, the most common words differ.

As it is noticeable in the below word clouds, most words have a positive or neutral complexion, which means that the general tourist vibe is pleasant. The most common words are connected with the accommodation service i.e. apartment, stay, place, host because airbnb data constitute the cardinal majority. Figure 6 represents the most frequent words in the overall reviews.

Figure 6: Most frequent words in the overall reviews

As a result, the Airbnb word cloud (Figure 7) is similar to the general word cloud. Many neutral words that are related to the hosting industry are noticed, i.e. apartment, place, stay, location. In addition, many words with positive sentiment exist, i.e. recommend, nice, good, perfect, lovely and some facilities, i.e. metro, restaurants are mentioned as well. Airbnb word cloud displays the positive opinion of the visitors in Athens. Moreover, it indicates that the city's structures are well-appreciated, so they should be maintained and improved.

Figure 7: Most frequent words in Airbnb reviews

[25] https://github.com/amueller/word_cloud

In Foursquare reviews, the overall feeling remains positive. As in any word cloud, words with general content, i.e. food, place, service are extremely frequent. Particularly, the Foursquare word cloud (Figure 8) includes many pleasant words, i.e best, good, friendly, delicious, which indicates that the users highly resort to Foursquare in case they want to recommend and not to criticize a venue.

Figure 8: Most frequent words in Foursquare tips

As in the Foursquare word cloud, the Google maps (Figure 9) word cloud includes many positive words, i.e. nice, excellent, great, amaze and the same high frequency content, i.e. food, place, coffee. The similarity of the Foursquare and Google maps words' collections makes sense because of the same nature of the reviews on the two platforms.

Figure 9: Most frequent words in Google maps reviews

The Twitter word cloud (Figure 10) diverges considerably from the previous word clouds. Twitter is a social medium with a different direction. Users through this platform express their opinions and comments on everyday incidents. Most common words in tweets include the key words for posts' collection, i.e. acropolis, athens, exarcheia. However, high frequency words are connected with the greek daily life and affairs, i.e. transport, deport, refugees, men, german. Taking into account that most greek users and citizens choose to

express themselves in the english language, it is expected to encounter topics that concern the society of Athens.

Figure 10: Most frequent words in tweets

4.4.2 Most common topics of discussion

Despite the most frequent words on posts from each online platform, an interesting question that occurs concerns the most common topics of discussion depending on the sentiment. The most suitable diagram to display the quantitative dominance of each object is bar plots, hence *seaborn*[26] corresponding structs are chosen as the means of the representation.

- **Airbnb most common topics of discussion**

The most common neutral topics of Airbnb (Figure 11) about Athens include general words that are relevant to the accommodation services, i.e. apartment, place, stay, athens, host. However, some positive words appear, i.e. great, nice. In many cases, sentiment analysis could not comprehend whether a post has a positive or a neutral connotation. Consequently, some positive posts have been labeled as neutral, and interchangeably some neutral posts have been labeled as positive, which explains the existence of positive words as most common neutral topics.

[26] https://seaborn.pydata.org/

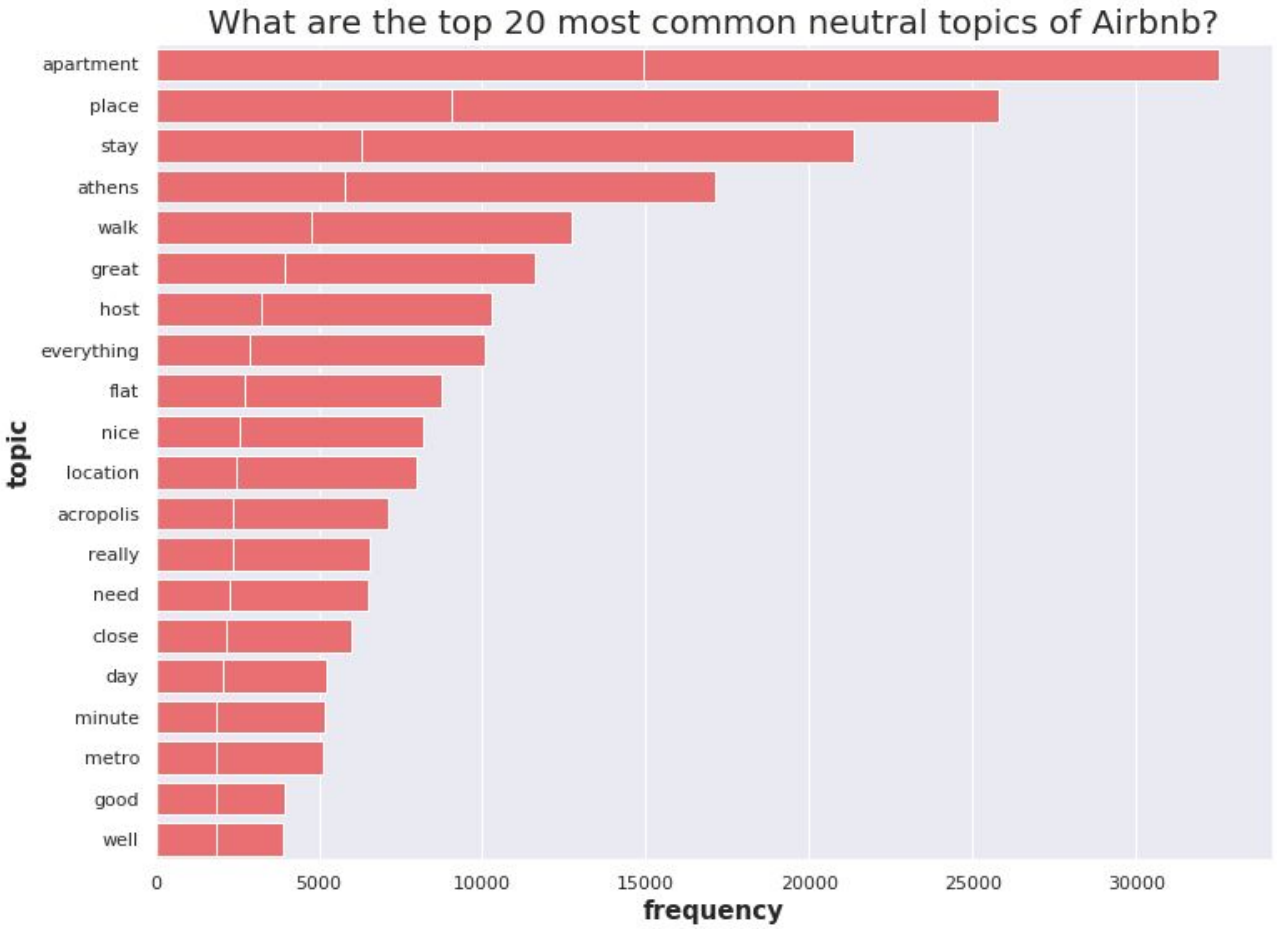


Figure 11: Top 20 most common neutral topics of Airbnb

In the diagram that contains the most common positive topics of Airbnb (Figure 12) it is noticeable that the adjective "great" has surpassed the word "apartment" which is the most common word in the hosting industry. This mentionable fact could be an indicator that the visitors/users of Airbnb platform are highly pleased by the accommodation in Athens. The word "recommend" is frequent as well, which means that the hospitality in Athens is well appreciated.

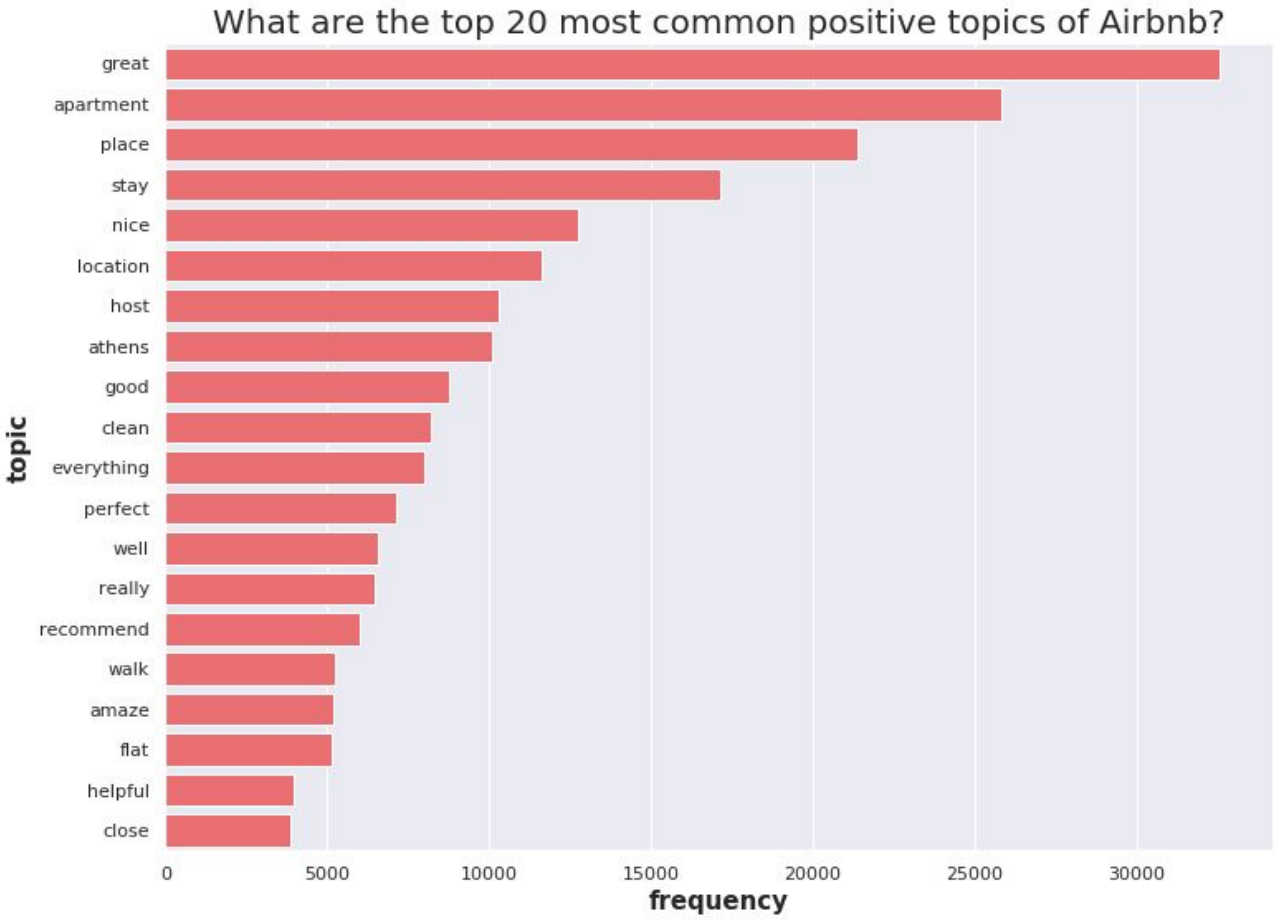


Figure 12: Top 20 most common positive topics of Airbnb

While observing the most common negative Airbnb topics (Figure 13), it is noteworthy that the frequency of unpleasant words is visibly lower, compared to the previous diagrams. "Noisy" and "unclean" apartments are elements that should be improved in order to eliminate every negative review. However, the "non recommended" factor (≈8) is a submultiple of the "recommended" factor (≈7000).

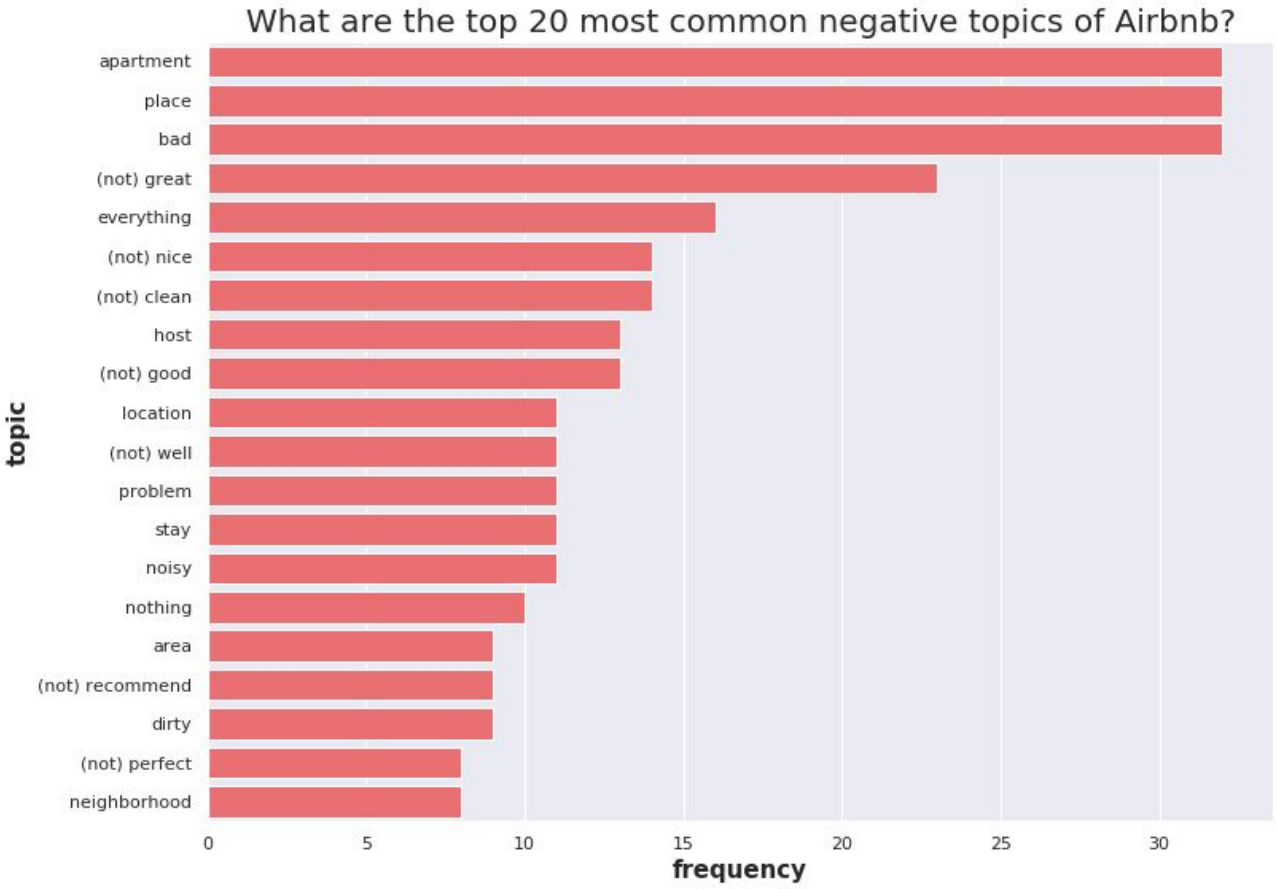


Figure 13: Top 20 most common negative topics of Airbnb

- **Foursquare most common topics of discussion**

Foursquare reviews are focused on nightlife, entertainment, food and coffee places of Athens. As a result, the most common neutral topics (Figure 14) include words such as "place", "food", "cocktail", "music", "dance", "night" etc. The bar plot also includes very positive words for reasons that has been discussed above. An interesting fact is that "wifi" facility has become so essential that it is mostly found in neutral reviews, meaning that this feature is no longer considered a luxury.

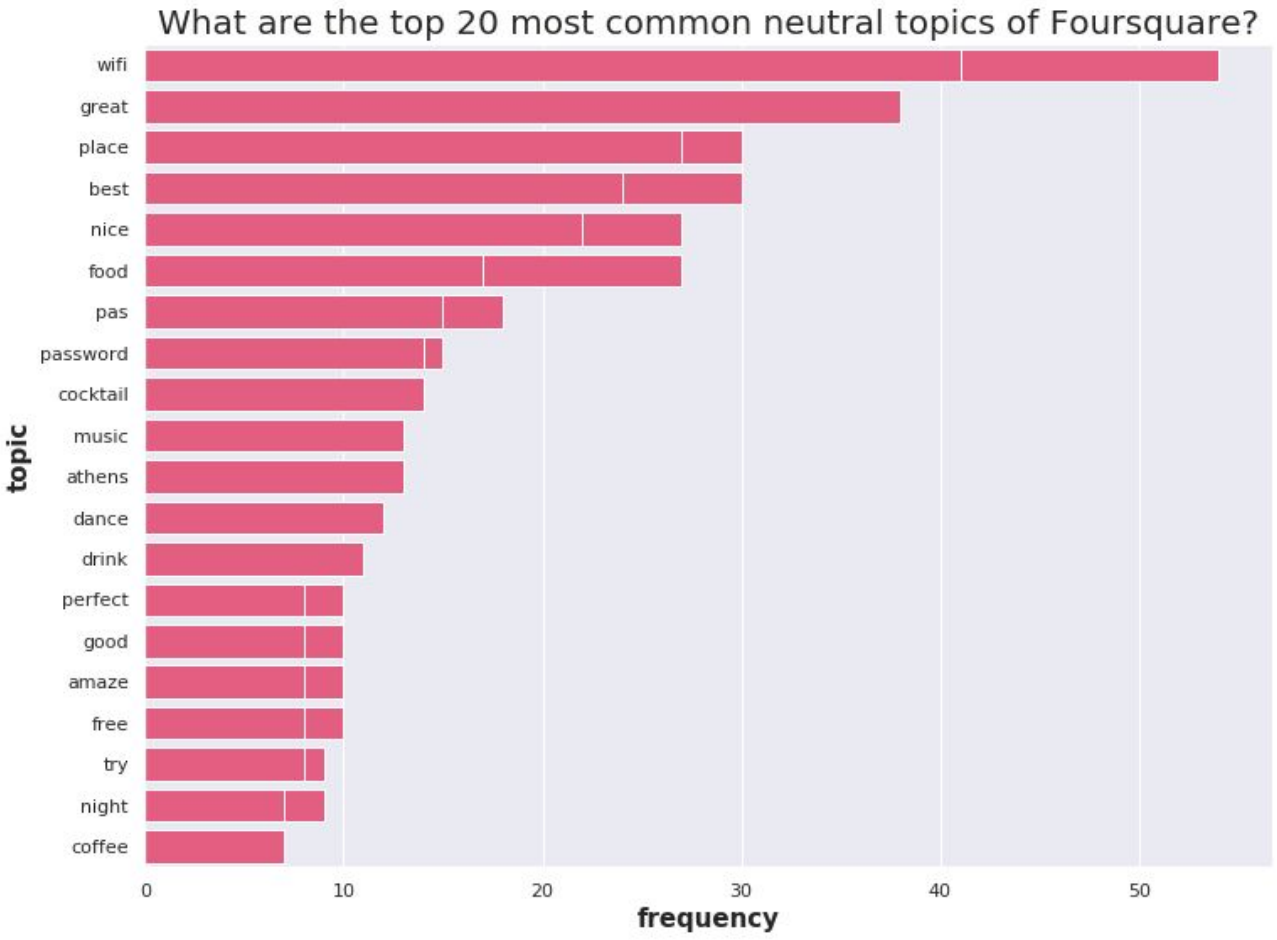


Figure 14: Top 20 most common neutral topics of Foursquare

The range of topics which has positive content is based on food, drinks and coffee. Topic modeling algorithm has mostly detected pleasant adjectives, i.e. great, best, good, but "wine", "beer" and "music" are highly mentioned and imply that tourists are generally satisfied from greek restaurants and bars (Figure 15).

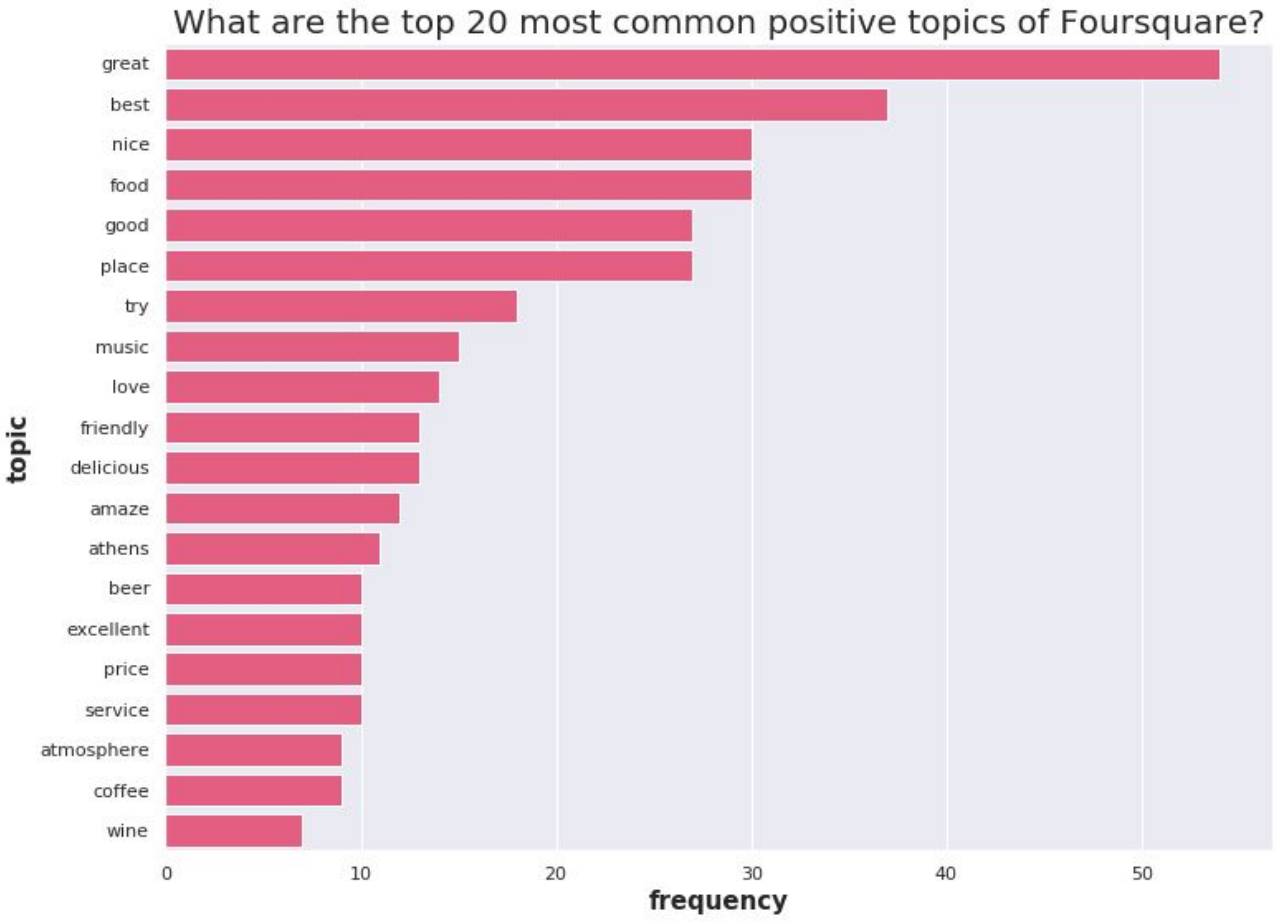


Figure 15: Top 20 most common positive topics of Foursquare

In Foursquare, the number of reviews with negative polarity (Figure 16) is likewise limited. However, "food" and "coffee" topics are highly mentioned and in particular, they are rated as "mediocre". Judging from the bar plot, visitors are not happy with the "service". Considering the numeric difference between the positive and the negative common topics, the majority of the restaurants and bars are decent except from individual cases.

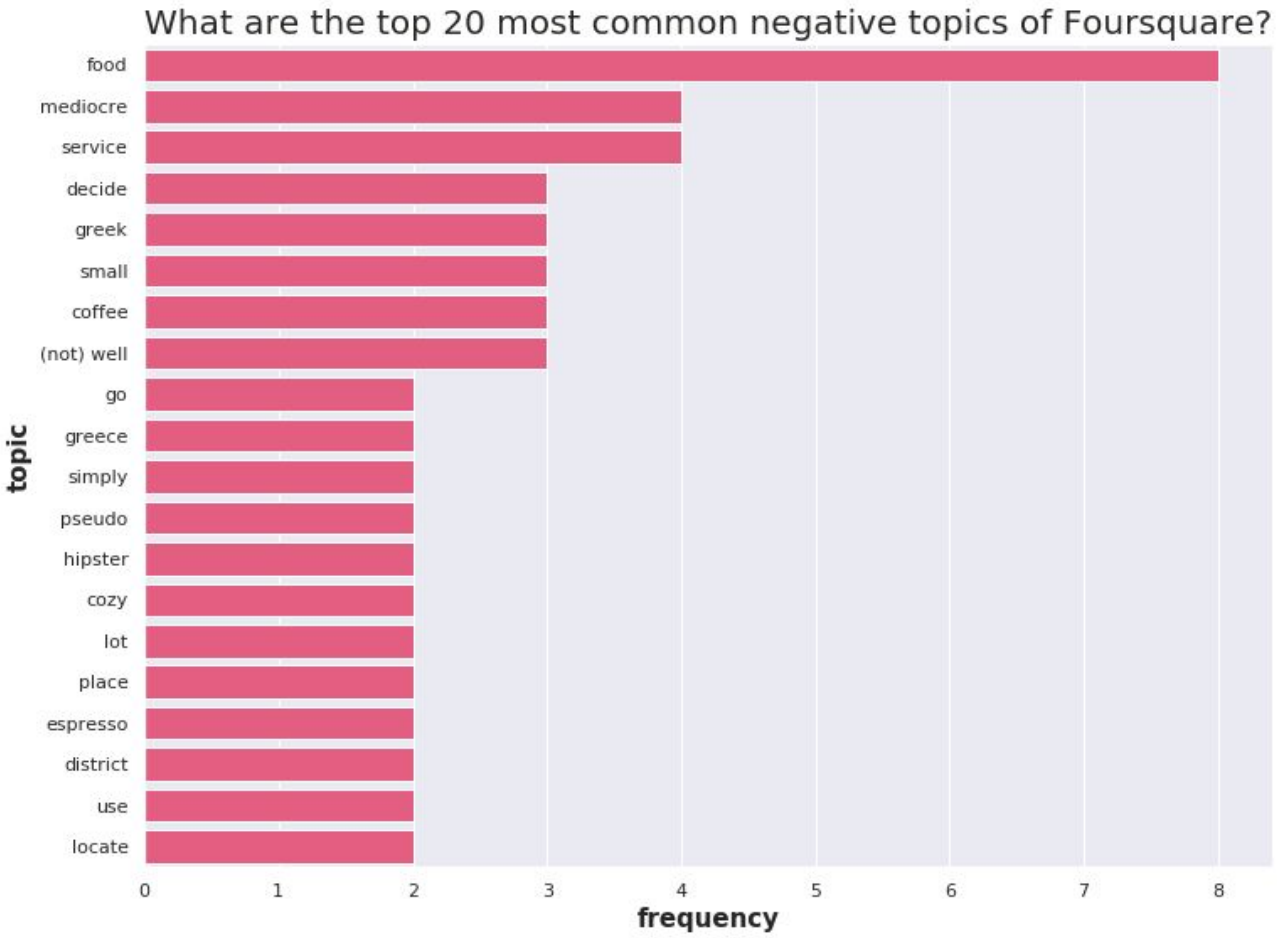


Figure 16: Top 20 most common negative topics of Foursquare

- **Google Maps most common topics of discussion**

Google maps diagram that contains the most common neutral topics (Figure 17) resembles the corresponding Foursquare diagram. Reviews on both platforms are similarly structured. Google Maps range of topics include places for "coffee", "food", "music" and "dance". In the diagram emotional words i.e. "bad", "nice", "good" have appeared, though they could not affect the neutrality of the review.

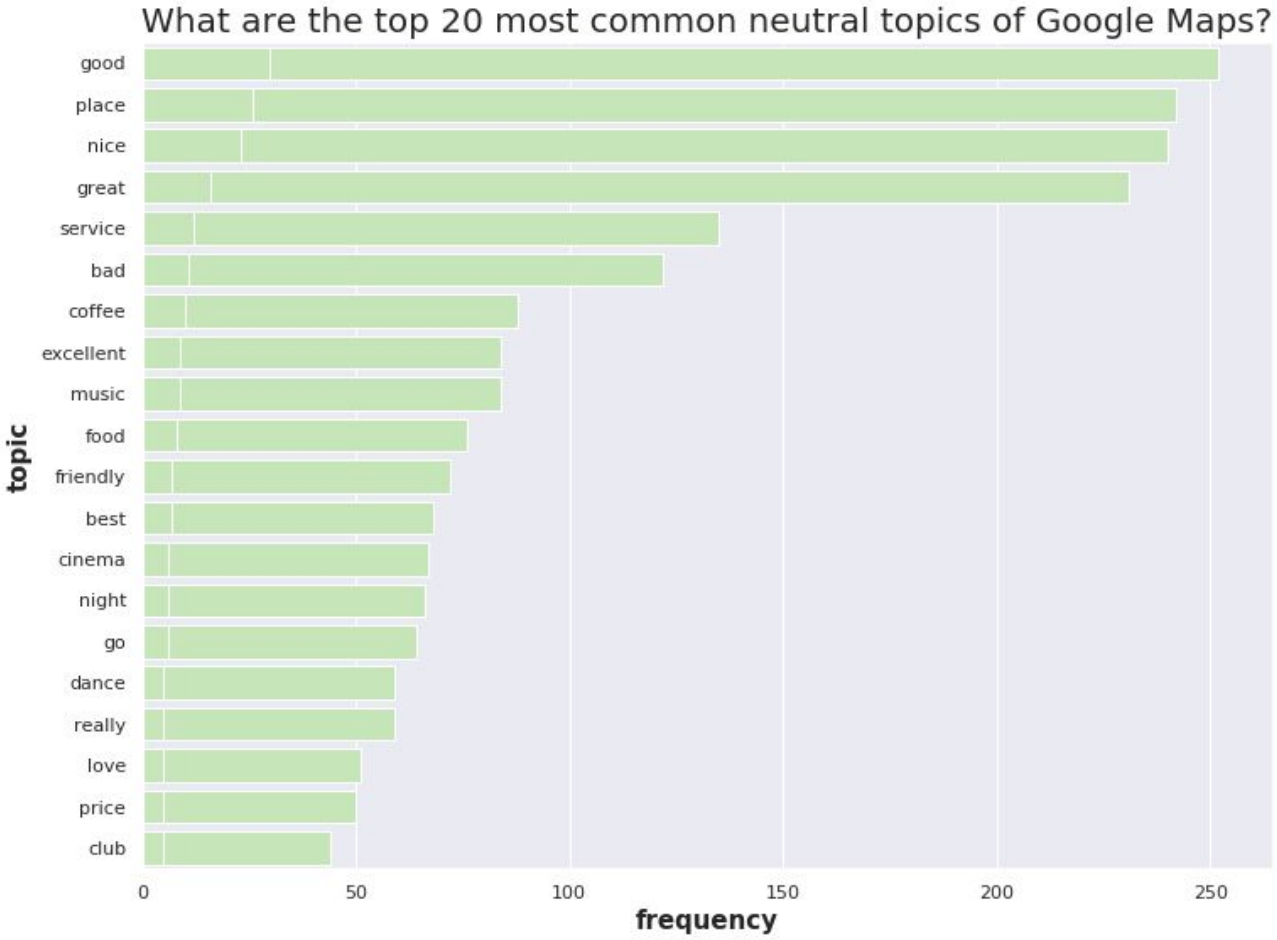


Figure 17: Top 20 most common neutral topics of Google Maps

In positive reviews, "coffee", "food", "bar" and "music are also mentioned. The fact that the words "staff", "service" and "friendly" are included means that the visitors' shaped opinion about the quality of employees is remarkable (Figure 18).

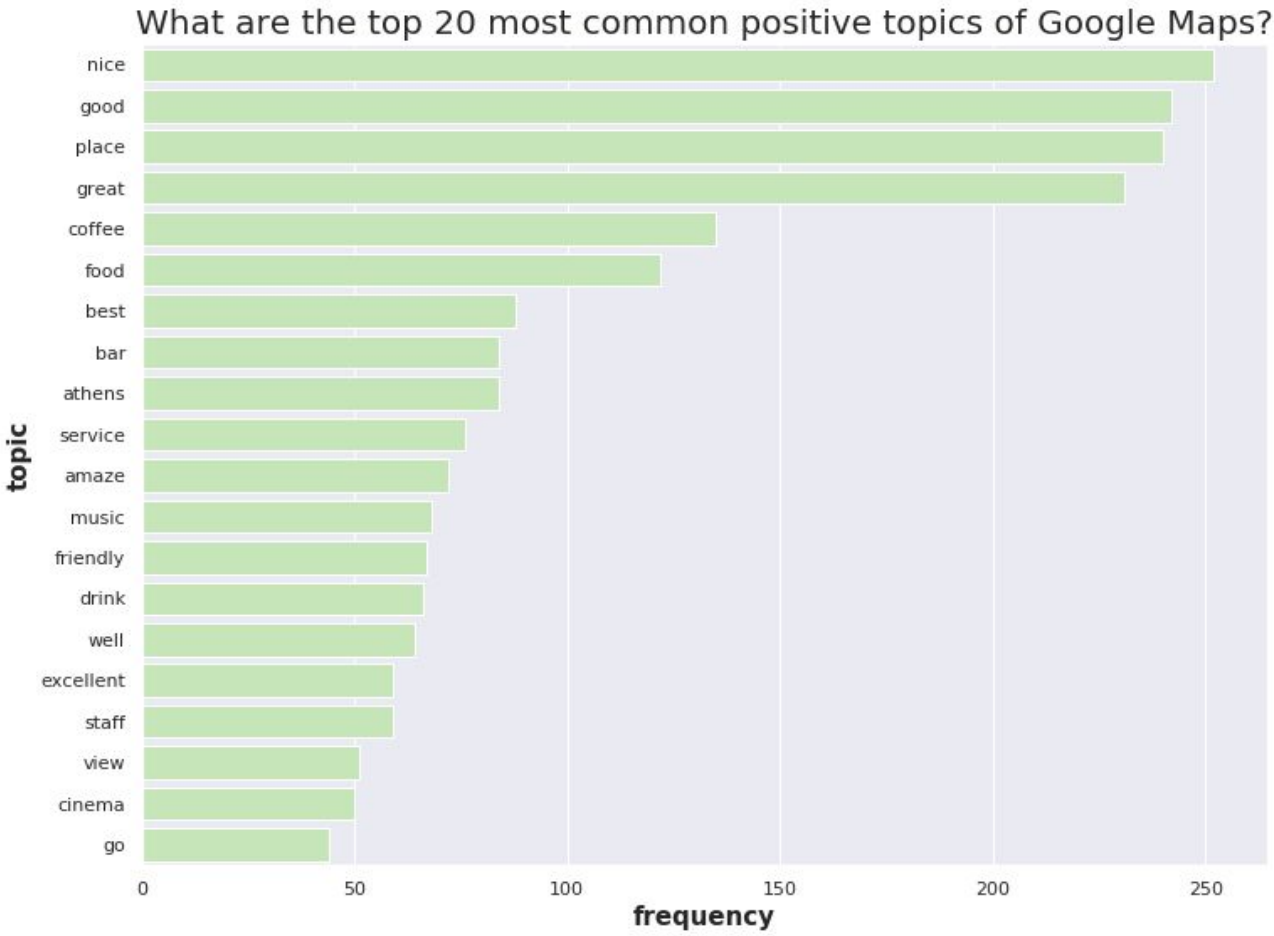


Figure 18: Top 20 most common positive topics of Google Maps

Due to the limited Google Maps dataset and the nature of the collected reviews (the two most helpful), negative reviews were almost non-existent. Thus, topics' frequency is unitary and there are no repeated words as topics. However, the words in reviews that have been labeled as negative should preoccupy the citizens. Words as "gay" and "straight" have been mentioned with a sense of aversion. Unfortunately, it appears that there are still visitors that criticize the existence of diversity (Figure 19).

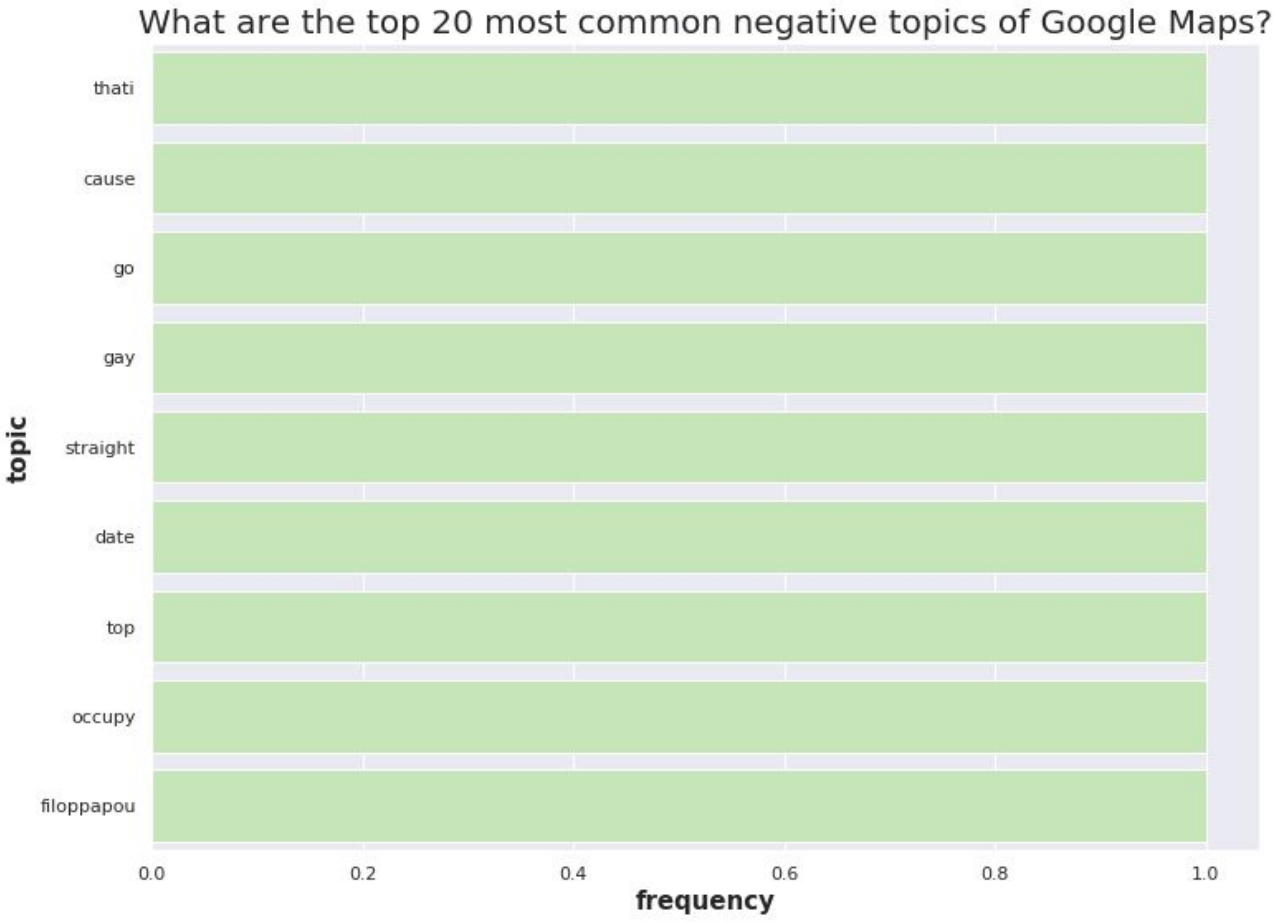


Figure 19: Top 20 most common negative topics of Google Maps

- **Twitter most common topics of discussion**

Due to the nature of this social media, topics like accommodation, food, entertainment and nightlife are not mentioned in tweets. The main discussed topics (Figure 20) pertain to locations, i.e. "athens", "acropolis", "greece", "plaka", "exarcheia" which is expected because of the existence of hashtags. Moreover, many neutral everyday words are encountered i.e. "work", "today", "city". Topics as "deport" and "attack" are unexpected and the sentiment analysis algorithm could probably label them as negative. However, migratory movement, nursing and restoration occupy the greek society, especially in its capital.

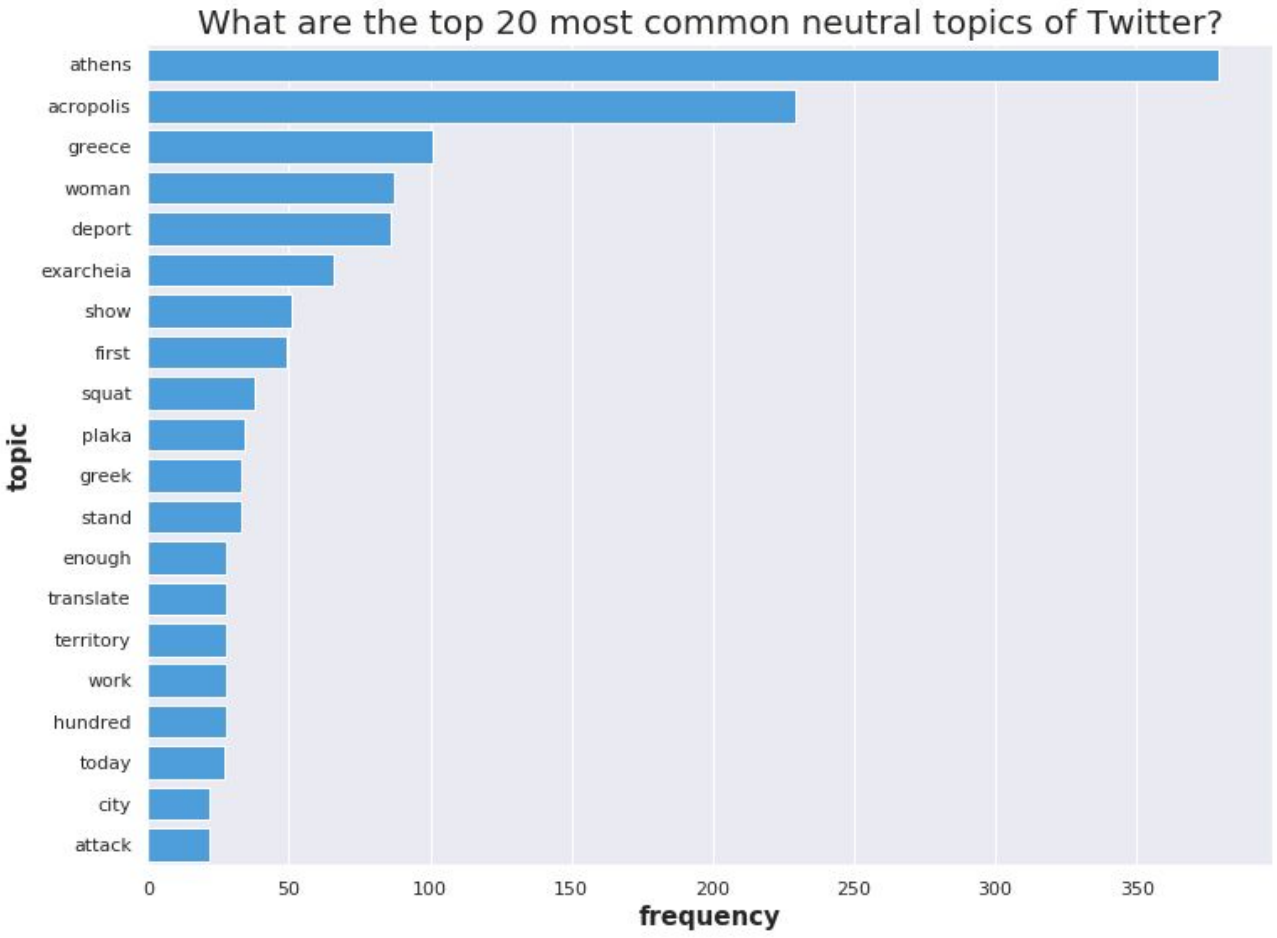


Figure 20: Top 20 most common neutral topics of Twitter

In the bar plot that present the top common positive topics (Figure 21), the dominant word is "travel". This fact indicates the positive touristic impression of Athens as a travelling destination. "Monastiraki" is highly and positively mentioned, which is encouraging as it constitutes a touristic center. In addition, many positive tweets has "restaurant" as a topic, which confirms the decent food industry in Athens.

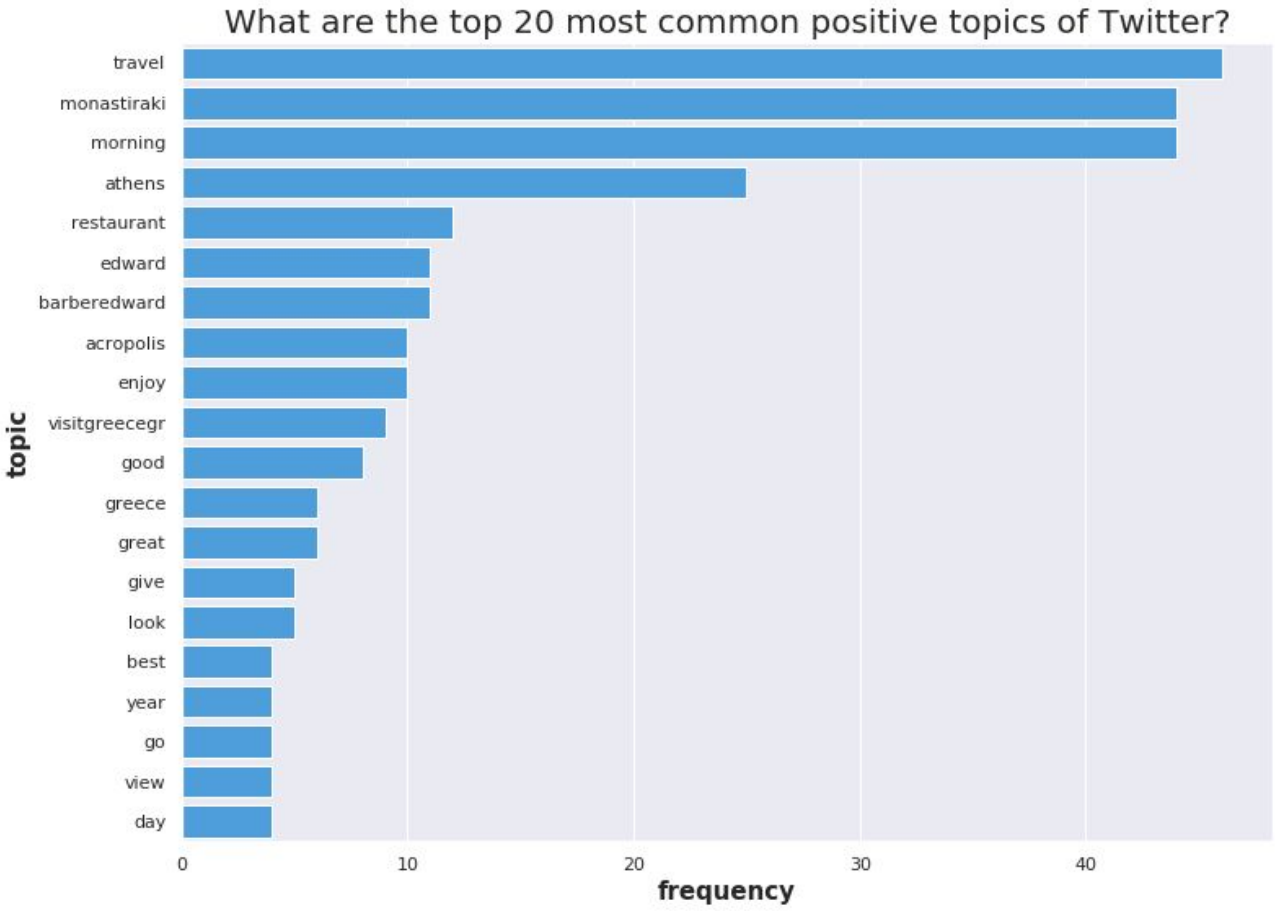


Figure 21: Top 20 most common positive topics of Twitter

Twitter comprises the only online platform which contains swearing . It is a medium where users have the opportunity to openly express themselves, therefore aggressive and offensive posts are expected. "Exarchia" is a frequent topic in negatively labeled tweets. A majority of residents and visitors portray this area as a neighbourhood that is beyond the law, a den of anarchists and criminals; hence the hesitant and negative sense (Figure 22).

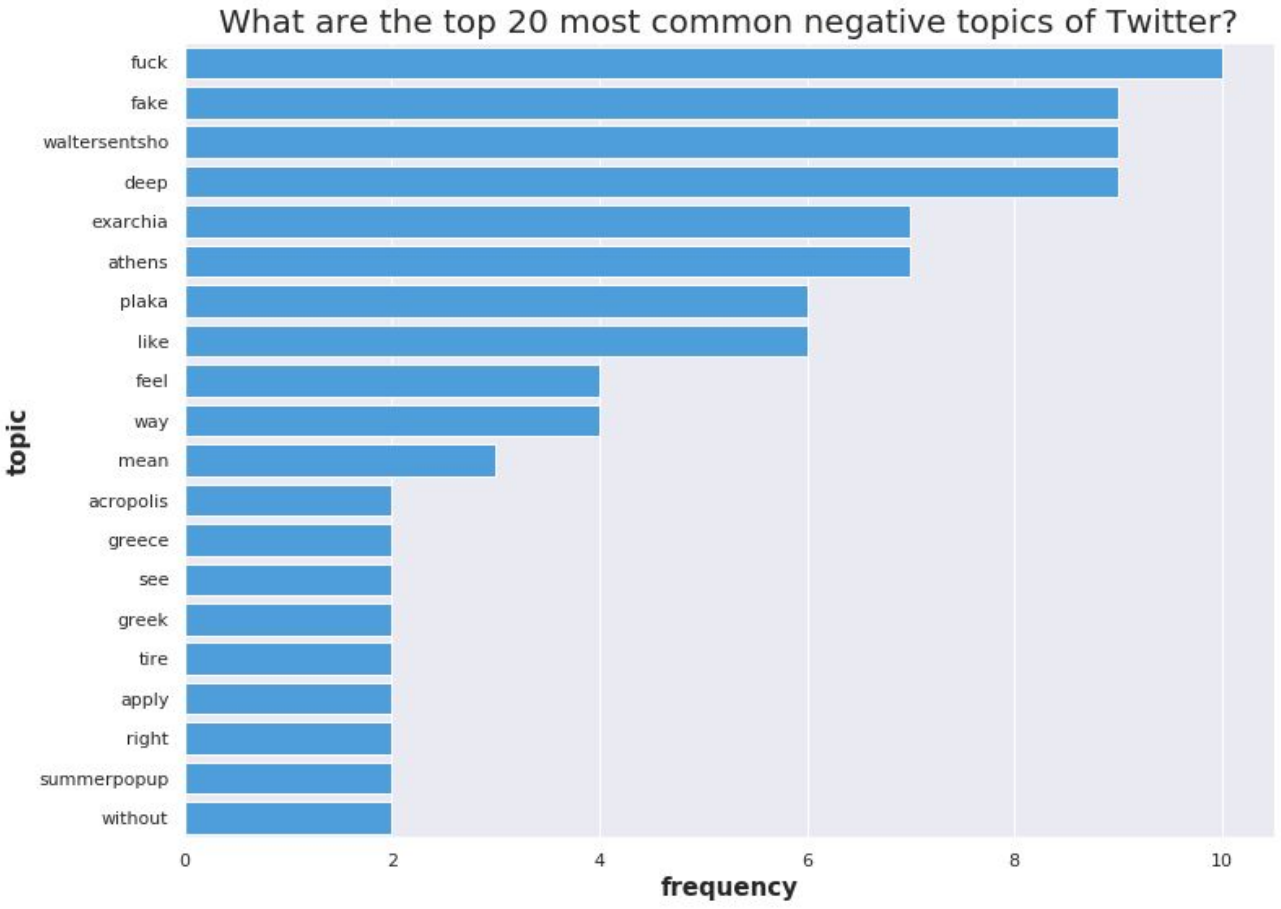


Figure 22: Top 20 most common negative topics of Twitter

4.4.3 Satisfaction rate of Athens on each platform

An interesting extracted result from the algorithms' application is the satisfaction rate for each platform based on the sentiment label (Figure 23). Airbnb reviews are mostly positive (70.8%). A small percentage (≈ 30%) is identified as neutral and the number of the negative posts are almost insignificant. Taking this into account, accommodation facilities in Athens satisfy the majority of the users on Airbnb. In the Foursquare dataset, half of the reviews have been labeled as neutral, almost 40% as positive and the negative posts gather 6.5% of the total reviews. The sentiment distribution of this platform confirms that the collected dataset constitutes a sufficient sample. On the other hand, reviews that originate from Google maps presents a satisfaction rate of approximately 92%. The rest 8% is covered by neutral posts, while negative comments barely exist. The Google maps dataset consists of the most helpful reviews, which might not be so diverse, hence suitable for data analysis. In the Twitter dataset, unlike the other platforms, the majority of the tweets is neutral. Although positive reviews prevail over negative, the numeric difference is not significant.

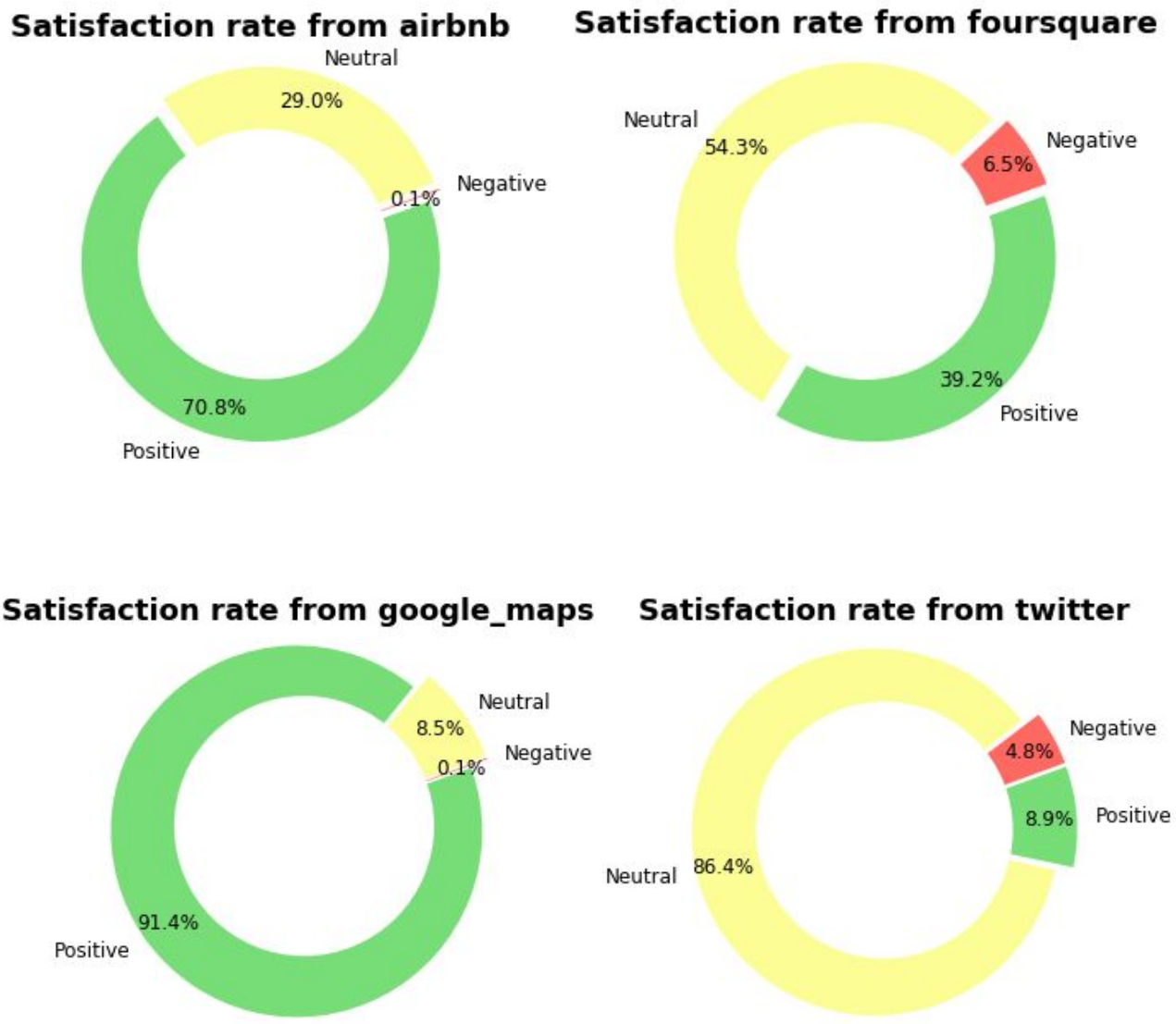


Figure 23: Satisfaction rate of Athens on each platform

4.4.4 Satisfaction rate of Athens on each category

The collected reviews/posts have been classified into five categories according to their context: accommodation, food, entertainment, nightlife and museum/archaeological sites. This particular label enables the calculation of the satisfaction rate of each type and empowers the travelling industry of Athens to feel the pulse of its infrastructure's impact on the different fields.

- **Accommodation**

  Results that are relevant with hosting activities originate from the Airbnb platform. Thus the pie chart of the accommodation category is identical with the Airbnb satisfaction diagram. As mentioned above, tourists are generally pleased by the hospitality of Athens (Figure 24).

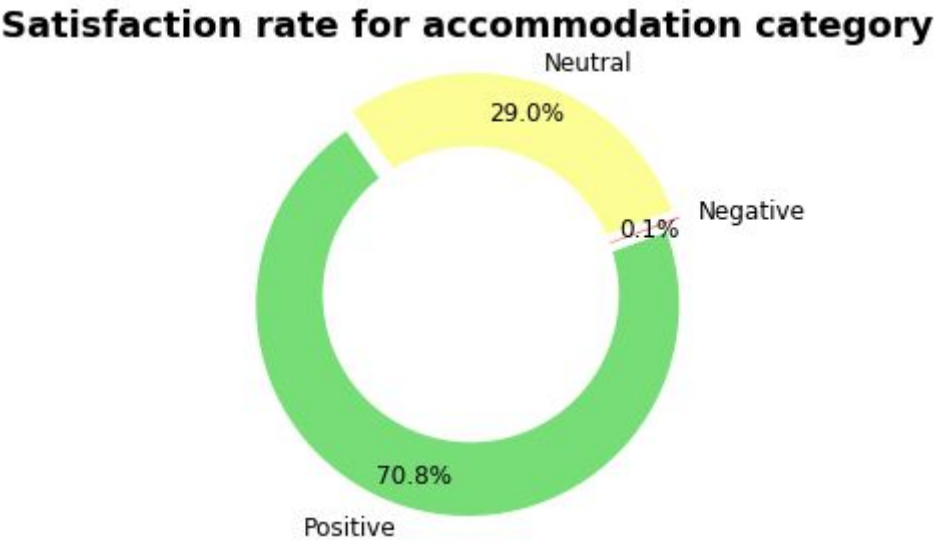


Figure 24: Satisfaction rate for Athens in the accommodation category

- **Entertainment**

  Athens is considered as one of the most entertaining metropolis of Europe. In this project, the category of entertainment mainly includes the activities that do not relate with food and drinks, i.e. theatre, cinema, galleries. Athens has a lot to offer for entertainment as numerous theatres and cinemas are spread in most areas of the city. Figure 25 indicates that a large percentage (≈72%) of visitors have a positive opinion about the recreational tourist attractions. Approximately 26% maintain a neutral position, while 2.6% are dissatisfied.

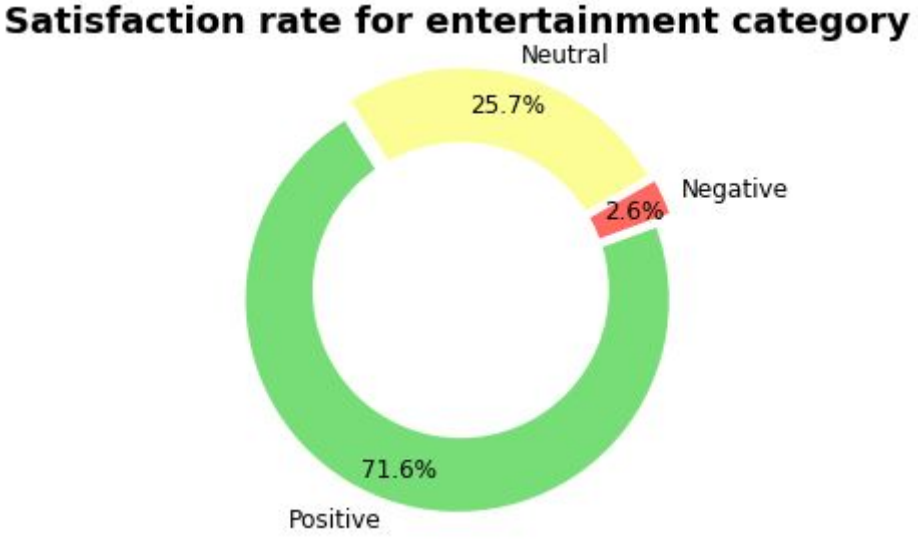


Figure 25: Satisfaction rate for Athens in the entertainment category

- **Food**

  It is a common opinion that Greek cuisine constitutes a main reason for someone to visit the country. Greek food is deeply influenced by both Eastern and Western culture, without losing its traditions and original flavours. Thus, Greece and

inevitably Athens, is one of the top worldwide gastronomic destinations. The chart verifies the expected opinion of the visitors apropos food. The vast majority is highly pleased while only a very small percentage appears disappointed. The results are encouraging for the local cuisine. Apart from the taste, the satisfaction rate includes the overall service, performance and professionalism in Athens' restaurants, which all seem to be more than adequate (Figure 26).

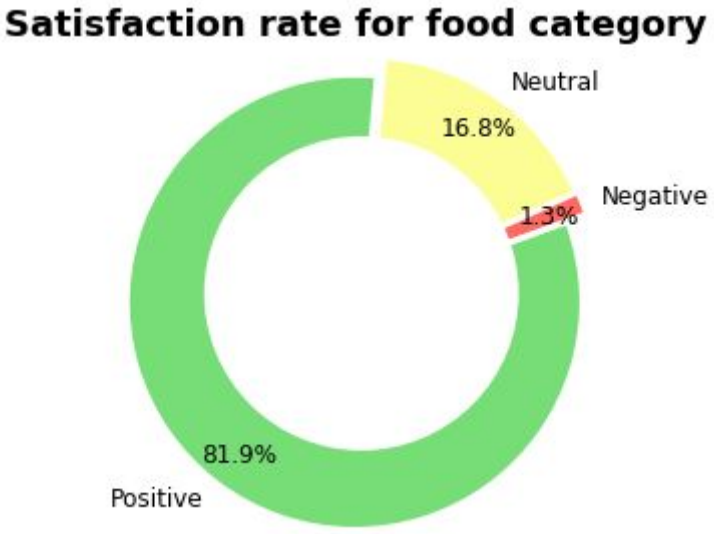


Figure 26: Satisfaction rate for Athens in the food category

- **Museum/archaeological sites**

  The cultural and social life of Athens plays out amid, around, and in landmarks that are centuries old. The remnants of Ancient Greece capture the most attention, however visitors in Athens can also admire elements from the "later" years for instance the Byzantine era and the neoclassical style architectural monuments [3]. Due to this fact, the dominance of positive reviews related to historic sites is obvious. Approximately 85% of the tourists are highly satisfied in terms of museums and archaeological places, which confirms that Athens is a destination that is suitable for history lovers (Figure 27).

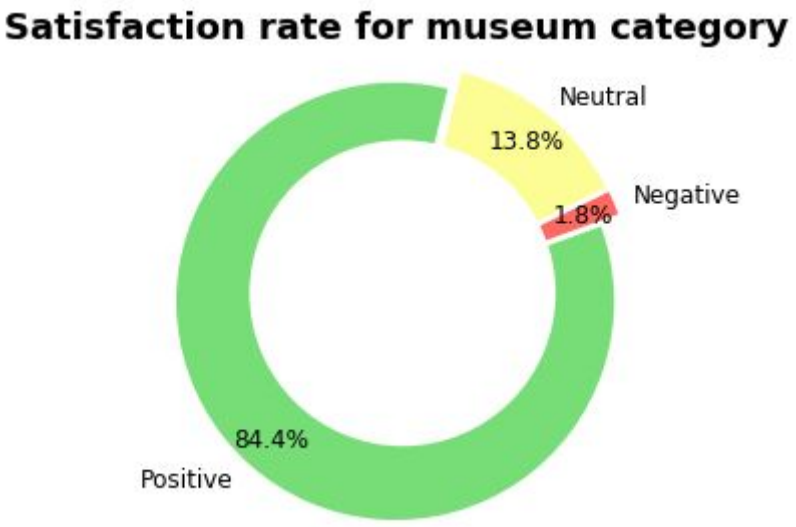


Figure 27: Satisfaction rate of Athens for museum category

- **Nightlife**

  One of the most attractive aspects of Athens is considered to be its nightlife. In general, tourists are often surprised by the fact that the Greek capital is so animated during the day as well as at night. There are numerous bars and clubs for every preference, many of which remain open until the morning. Visitors, as it emerges from the diagram (Figure 28), appreciate the active nightlife, as more than 75% of people have reviewed it positively.

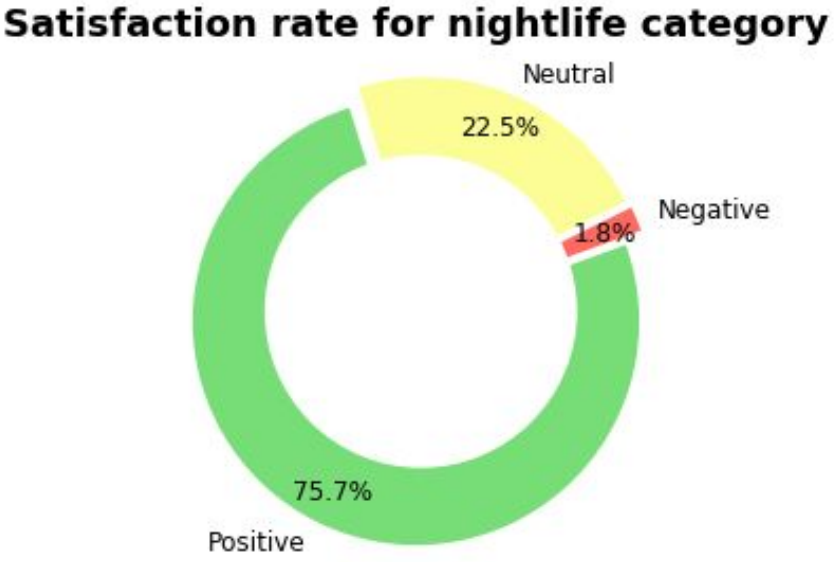


Figure 28: Satisfaction rate for Athens in the nightlife category

4.4.5 Types of venues on each category

Apart from the general category (accommodation, entertainment, food, museums, nightlife), the collected data include the specific type of the venue. The types of venues that are mentioned on each review, present the general orientation of the tourism's related operations in Athens. In case similar information could be found, it would be interesting to compare the different travelling enterprises that flourish in each city.

- **Accommodation**

  Nearly half of the Greek population lives in the Athens metro area, more than 4 million people, making it one of the most densely populated areas in Europe (1,540 inhabitants/km²) [4]. Consequently, the apartments are the most common property type of Airbnb accommodation. The bar plot (Figure 29) notices that over 200,000 reviews originate from users that have been hosted in apartments while visiting Athens. However, apartments are generally the most popular recommended property type in the Airbnb platform, hence the tremendous number. The following most common type of accommodation is the lofts, although they are not as preferred (less than 25,000 mentioned).

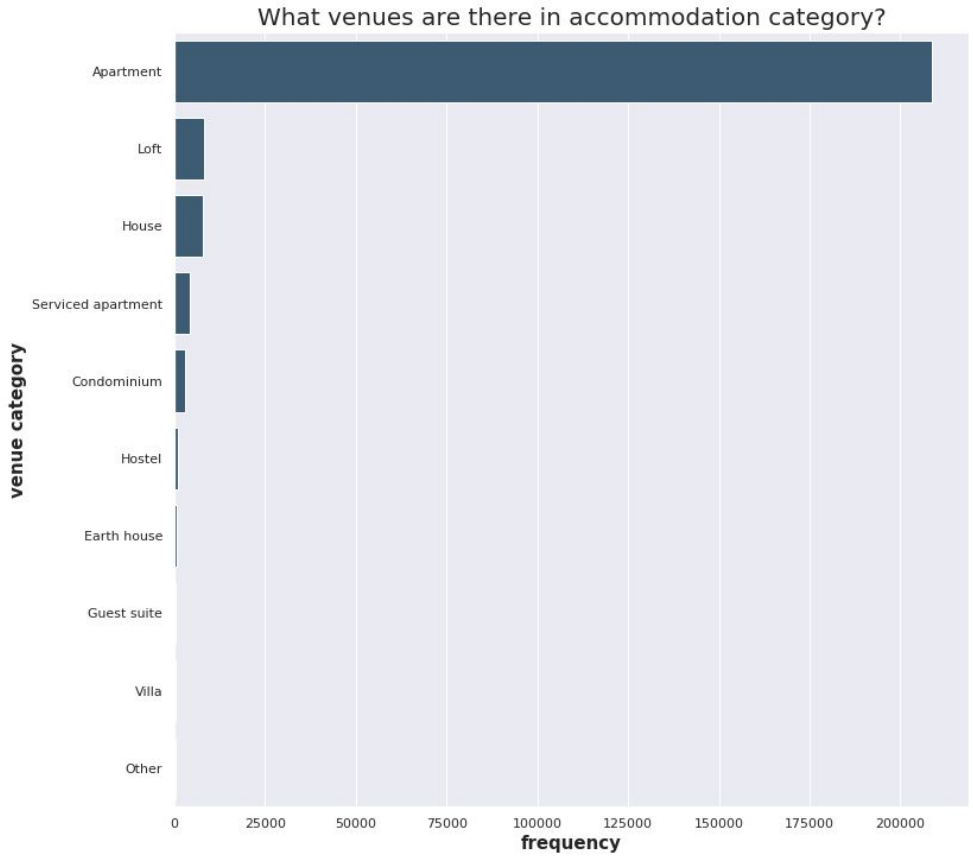


Figure 29: Common types of venues in the accommodation category

- **Entertainment**

  In the entertainment category, most venues are under the general label of "point of interest" (POI), which mostly refers to theatres and cinemas. Cinemas appear under

similar names (movie_theater, Movie Theater) in the same diagram and their sum renders them the most popular entertainment category. Athens combines many types of cinematic venues, from old-fashioned open-air cinemas to alternative movie theaters and mega cinema complexes, therefore cinemas deserve to be on top of the list. The category that follows is 'dance studios' and 'theaters'. They are considered to be a more luxury activity, which explains their position on the diagram (Figure 30).

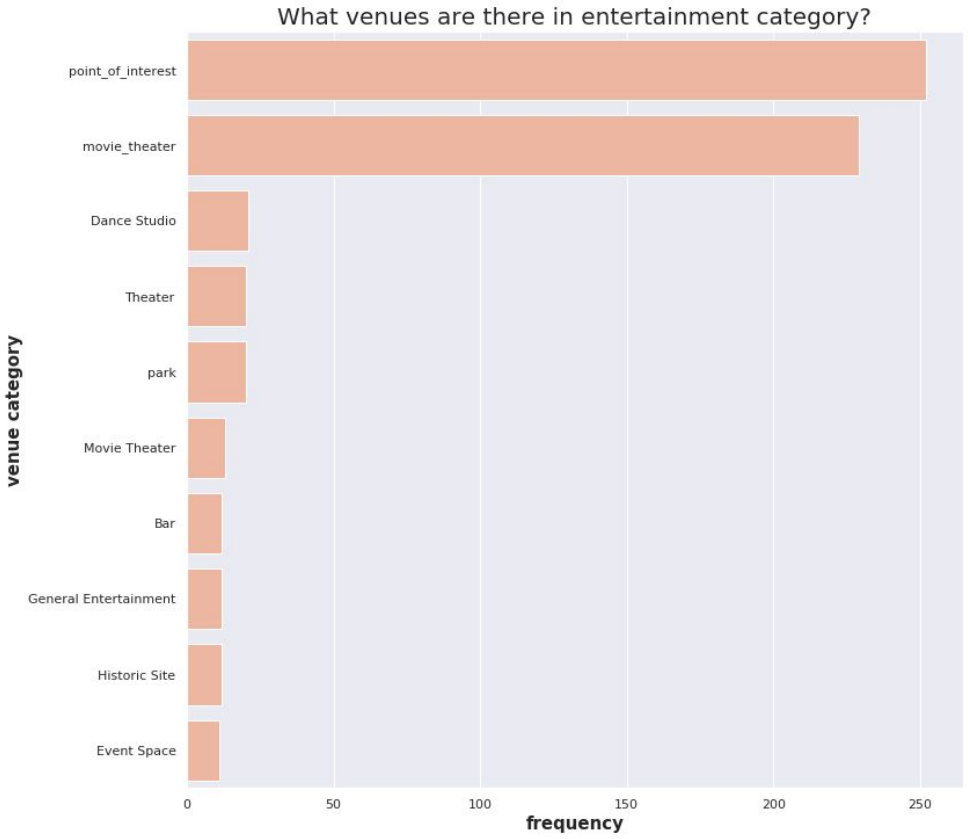


Figure 30: Common types of venues in entertainment category

- **Food**

  Greece's coffee industry is rapidly growing. In the last decade, most of the enterprises involved in the coffee industry have succeeded in developing against all the odds, due to the financial crisis. As coffee seems to be one of the last affordable daily "luxuries" for consumers, a further growth is expected. Businesses specializing in takeaway coffee flourished and new chains keep on appearing that grow in numbers and revenues [2]. Almost 500 "cafe" appear in the collected data plus the traditional "kafeneia" which are mentioned lower in the diagram. Moreover, "Greek restaurants" and "souvlaki shops" cover a great portion of the food service activities. Souvlaki has always been the most popular and cheap go-to street food in Athens. Considering the limited budget of the average Athenian, more and more such places have suddenly arose in every neighbourhood.

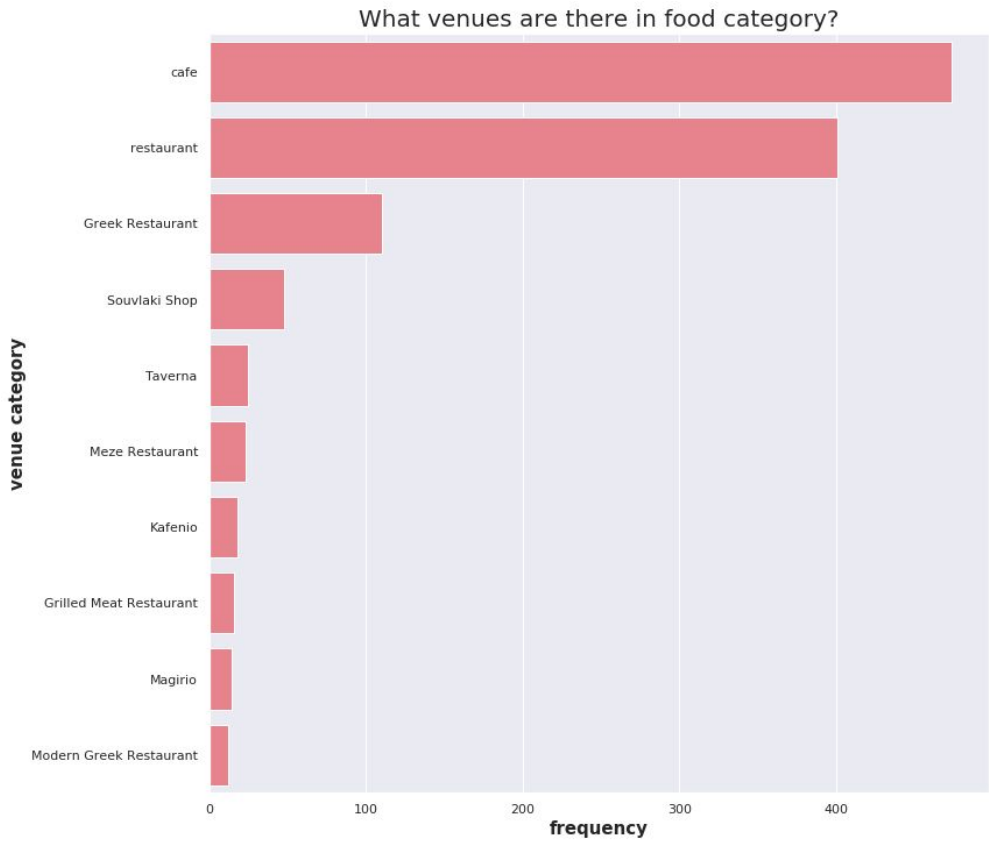


Figure 31: Common types of venues in food category

- **Museum/archaeological sites**

  As shown in the diagram (Figure 32), Athens is loaded with museums. Over 250 posts refer to museums. However, no other classification has been performed, so the only information is that there is a good number of "Art Museums" and "History Museums".

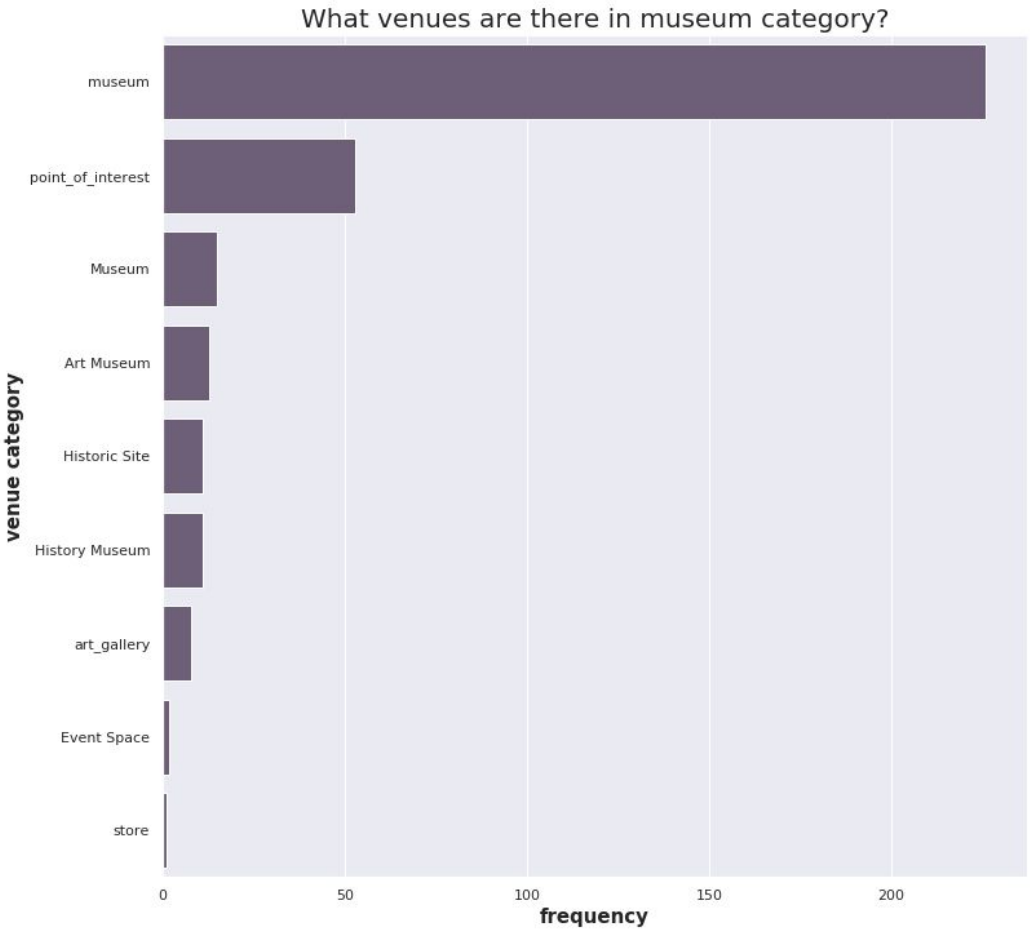


Figure 32: Common types of venues in the museum/ archaeological sites category

- **Nightlife**

  The nightlife scene in Athens has become a collection of hot spots that draw the attention of both visitors and locals. Over 700 different bars and over 100 nightclubs are mentioned in the collected data.

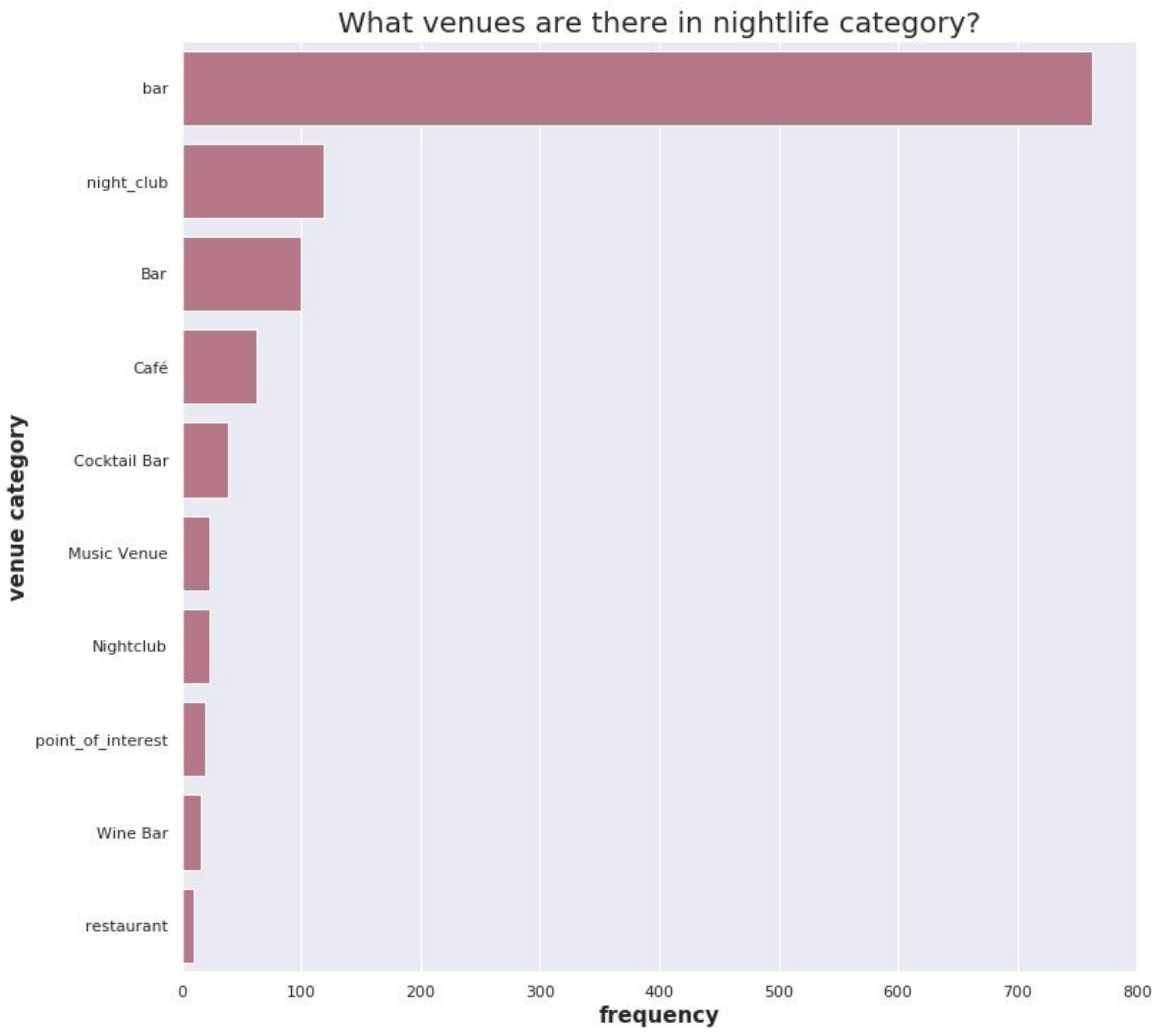


Figure 33: Common types of venues in the nightlife category

4.4.6 Most mentioned districts of Athens in reviews

The collected data offer information apropos the districts of Athens that are considered the most popular (Figure 34). The fact that the majority of the reviews originate from the Airbnb platform explains the results, which indicate non expected areas as the most reviewed. Kypseli and Dourgouti (an alternative name for the region of Neos Kosmos, Syggrou and a part of Koukaki) constitute districts under gentrification while most apartments are Airbnb rental properties. Thus, regions like the Acropolis or Monastiraki and Plaka appear lower in the current diagram. Gouva is a densely populated region and includes areas such as Pagkrati, Neos Kosmos and Mets. Although it might not be the first choice when it comes to busy nightlife, this residential yet vibrant area still offers a variety of things to do and see but mainly it is a top option with demand for Airbnb.

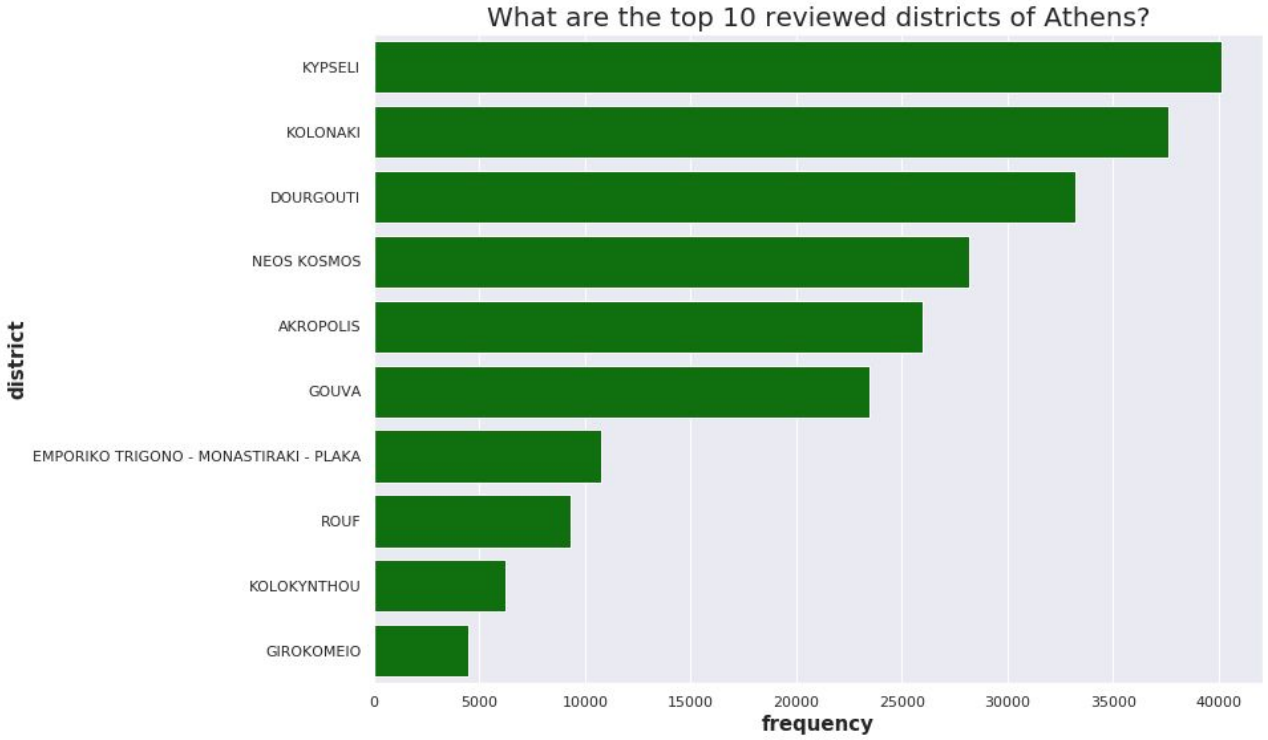


Figure 34: Top 10 reviewed districts of Athens

4.4.7 Interactive maps of sentiment

Two different styles of interactive maps constitute the final representation type of this project.

In the first map (Image 9), the collected data are able to be depicted, given the coordinates of the venues. As a result, tourists are able to get a sense of the city's districts that gather the most points of interest and municipal stakeholders are able to detect both emergent and overwhelmed areas, and to track down tourists' and locals' activities. In the long run, if similar research is conducted, it will be interesting to observe the difference with the touristic centers, hence in travelling preferences. *Folium*[27] marker clusters form a map that contains many markers. When the map is zoomed out, nearby markers are combined together into a cluster. When the map zoom level becomes closer (Image 10), the cluster separates. The marker includes information about the sentiment and the category of the represented post/review. The markers' icon represents the category (entertainment, food, accommodation and the marker's color matches the sentiment (red color for negative sentiment, blue color for neutral sentiment, green color for positive sentiment). In addition, each marker contains the equivalent topic. As it is observed, many markers may correspond to the same location because several reviews have been written for the same venue.

[27] https://python-visualization.github.io/folium/

- **Clustermap**

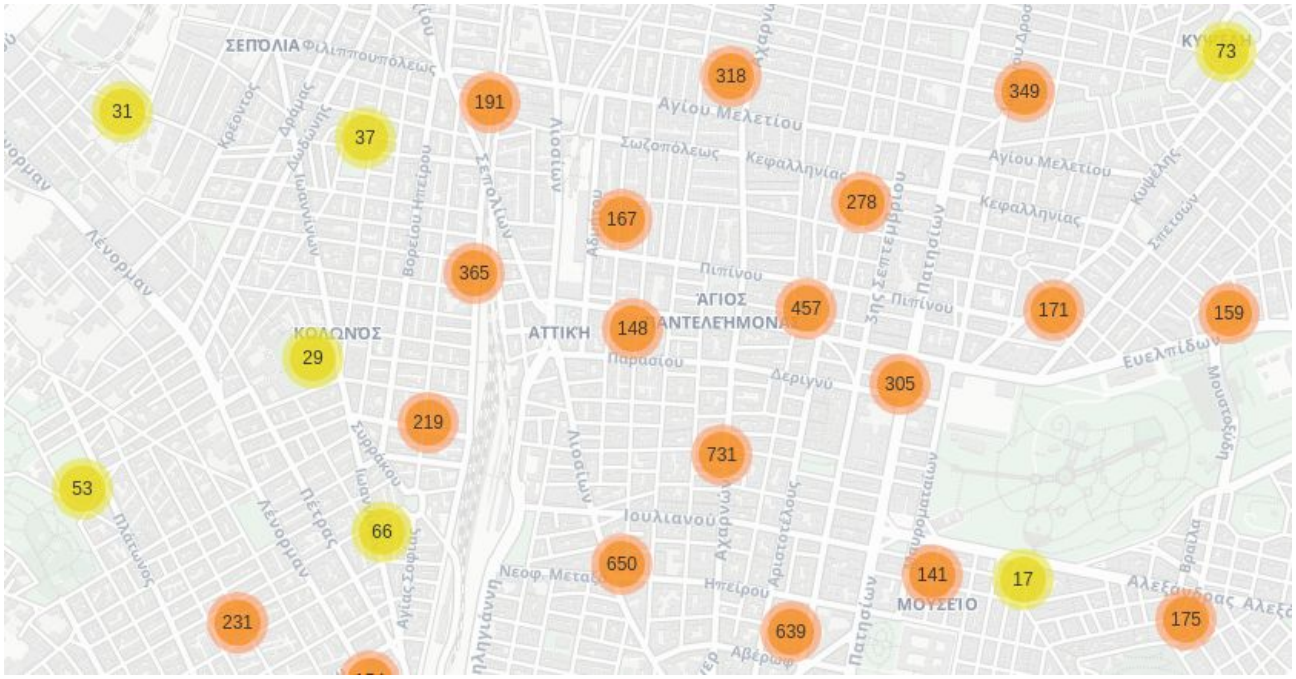


Image 9: Map of Athens after clustering the Foursquare markers (zoomed out)

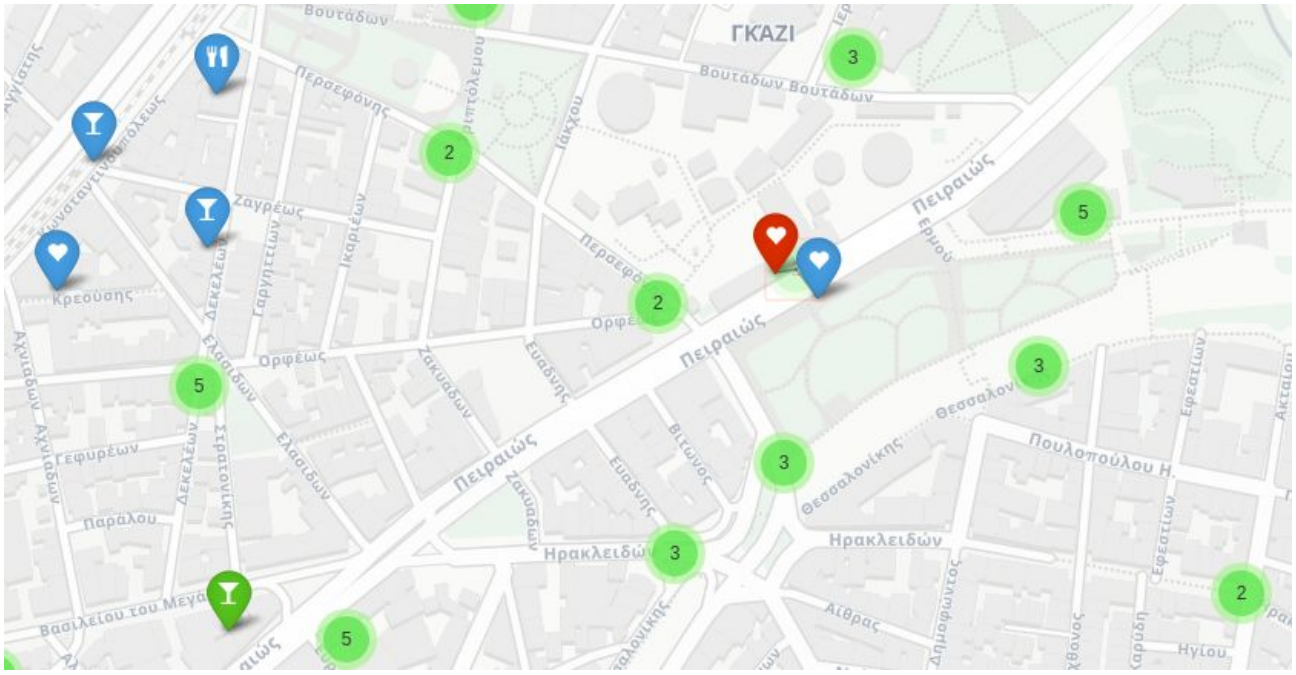


Image 10: Map of Athens after clustering the Foursquare markers (zoomed in)

In the second map (Image 11) each district of Athens is colored based on the satisfaction rate given by the collected data. The type of representation used is called choropleth map. A choropleth map is a category of thematic maps in which areas are shaded or patterned in relation to a statistical variable that visualizes how a measurement varies across a geographic area (in this case the sentiment) [1]. In this project, a choropleth map may not be so representative, because each area contains a different amount of reviews as well as neighbourhoods with zero reviews/posts. However, it constitutes the most suitable way to show the overall sense of the research's findings apropos the sentiment of tourists for Greece's capital. The color bar indicates that the darker a region, the better the satisfaction rate. The areas that are completely white are associated with either unidentified sentiment (zero reviews/posts are corresponding to this district) or neutral sentiment, hence the average satisfaction rate is zero. The choropleth map of this project includes an interactive element (Image 12). On hover, each region shows its name and satisfaction rate.

Districts' borders is a geojson file provided by the GIS department of the municipality of Athens. Geojson format is designed for representing geographical features and non spatial attributes. The geojson input contains the polygon of coordinates that demarcate each area and the name of the represented area. However, the collected data cannot exploit geojson file without undergoing further processing. Data's addresses in the data were converted to match the names mentioned in the input file. Polygons form vectors and are handled as GEOdataframe in order to be compatible with the plotting package. *Matplotlib*[28] is a library that copes with vectorial data efficiently.

- **Choropleth map**

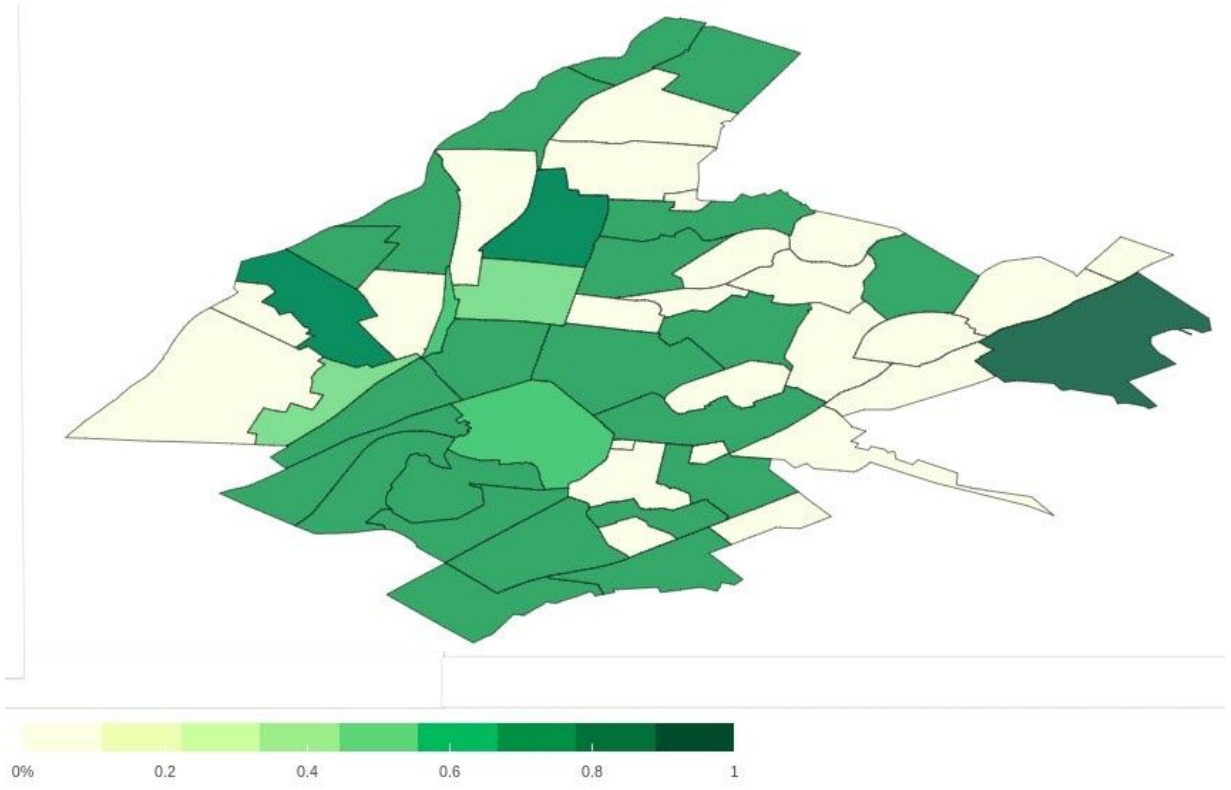


Image 11: Choropleth map of Athens with the explanatory color bar

[28] https://matplotlib.org/

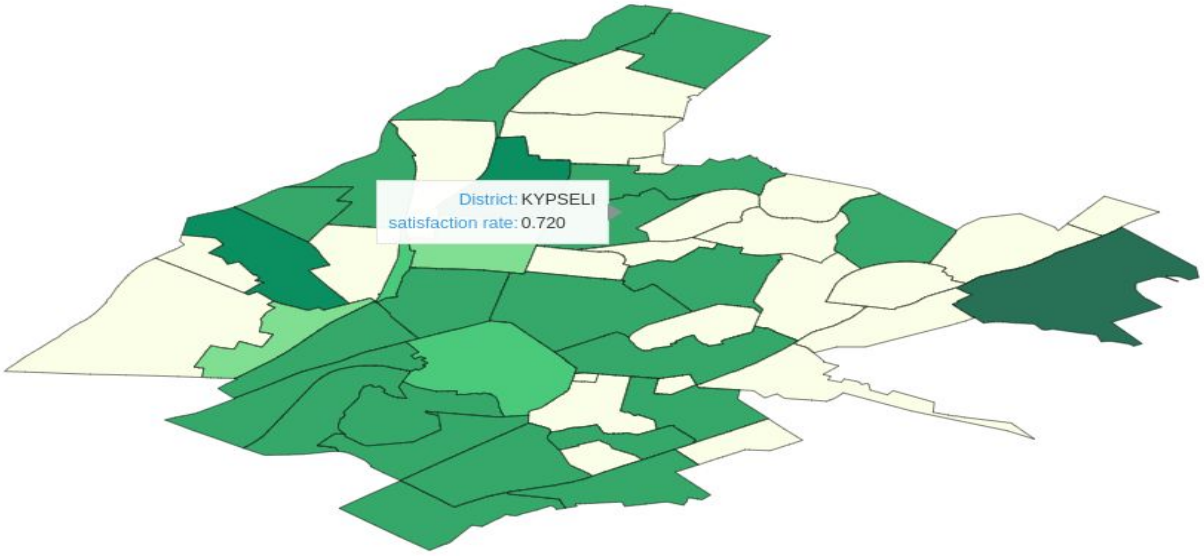


Image 12: Choropleth map of Athens and the hover feature

5. CONCLUSION

This chapter summarizes the results of this thesis and proposes future extensions that would complete the research and offer interesting contributions to the community.

5.1 Results and discussion

The initial goal of this thesis was to exploit data from online platforms and extract the general sentiment of travellers to the city of Athens, Greece's capital, as well as any other related information, in order to detect the facilities and defective points in the Athenian tourist industry. Whilst compelling in theory, in practice the task of gathering and processing a high velocity and large volumes of data has become very complex.

The sentiment analysis and data/information visualization methodologies followed are popular and widely used processes. However, each stage lurked its own challenges. During the collection of reviews , finding a decent amount of valid open source data was a slow and difficult process that required time and manual parameter entry on a daily basis; hence, it could not be fully automated. Next, raw data have undergone a series of transformations in order to be compatible input for the machine learning algorithms and the visualization libraries. The data preprocessing stage, in this project, could be labeled as an agile procedure. Questions' increase causes further data preprocessing in order to extract the necessary answers. Machine learning algorithms included both custom classifiers and ready-made tools. The implementation fulfilled the hypothesis of combining different methods to achieve higher accuracy. However there is space for improvement in case more data or a better trained dataset is collected. Moreover, the posts that have been labeled as neutral, which was the great majority, could have been dealt with more detail, taking into consideration sarcasm or opinion words. It was fascinating to see how the lexicons and the training datasets are probably the most important piece in a sentiment analysis system.

Using Big Data and deep learning approaches can help tourism research to discover relationships based on interconnected sets of data. Tourism research may further move into an approach of data driven practices in order to improve the local industry. Sentiment analysis and natural language processing is only a small step towards this direction.

A supplementary interesting element is related with the quality of research. The certainty that the collected data, hence the overall research, has offered useful information could only be shown in the last step of visualizing the results. Fortunately, this project did not conclude that the dataset was insufficient to provide any additional information. It has confirmed the general positive attitude of travellers towards the city of Athens, especially in terms of food and entertainment and it has underlined the known problems of noise pollution, cleanliness and lack of infrastructure. Despite the economic crisis of the last decade, Greece's capital seems to continue to satisfy the majority of tourists, who in return express their experiences in public online platforms thus contributing to the destination's promotion.

5.2 Future work

Machine learning projects in service of tourism activities or any social aspect are very promising and show great potential to provide truly useful indicators that can support the relevant stakeholders' decision making. Machine learning is a circular process. The

increase of the collected data will lead to better algorithm's training, thus more accurate results and higher validation scores. Consequently, research conclusions will be closer to reality and will have a higher scientific impact. Moreover, society's indicators and status alter continuously. In order to achieve social monitoring, it is essential to maintain a record of the same data in the long run.

Assuming that the research will continue to take place with up to date and similarly formatted information, the next step is to expand the data retrieved to upgrade this work from a simple sentiment analysis project to a general social monitoring research. Despite the advantages that tourism can bring to the development of an area, tourism can put pressure on natural resources when consumption is increased in areas where resources are already scarce. In the case of Athens, tourism can cause the same forms of pollution as any other industry: air emissions, noise, waste and littering. To understand the nature of the challenges and develop coping strategies, city officials and other stakeholders must learn to take advantage of change through continual monitoring and social learning in order to make a shift towards a more sustainable future. In this context, making tourism more sustainable means a continual process of making optimal use of environmental resources with respect to the host communities.

An interesting research goal for the future would be to combine the results from this research, which includes the designation of the most popular districts of Athens tourism-wise, with environmental indicators. Counting carbon footprints during high season, direct energy consumption and the amount of emissions or waste using sensors are actions that may or may not take place in research institutions. Tourism's measurements should not be narrowed to simple questions as the satisfaction rate of tourists but should also include the impact to the local community. Incorporation of those different types of data is crucial to comprehend the impacts of tourism.

With the contribution of data science, the city of Athens could better understand its tourists' needs, preserve the advantageous elements, improve the problematic issues and sustain its reputation as one of the most preferred travel destinations worldwide.

ABBREVIATIONS - ACRONYMS

2D	Two Dimensional
API	Application Programming Interface
BOW	Bag Of Words
CSV	Comma Separated Values
IoT	Internet of Things
JS	Javascript
JSON	Javascript Object Notation
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POS	Part Of Speech
ppGIS	public participation Geographic Information System
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
TSI	Tourism Sentiment Index
VADER	Valence Aware Dictionary and sEntiment Reasoner

REFERENCES

[1] "Choropleth map," *wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Choropleth_map. [Accessed: 04-Jan-2020].

[2] A. VLACHOU, "Greece's coffee industry grows despite financial crisis," *ekathimerini*, 2018. [Online]. Available: http://www.ekathimerini.com/230021/article/ekathimerini/business/greeces-coffee-industry-grows-despite-financial-crisis. [Accessed: 13-Dec-2019].

[3] "Athens is the world's ancient capital," *lonely planet*. [Online]. Available: https://www.lonelyplanet.com/greece/athens. [Accessed: 12-Dec-2019].

[4] "Athens," *urbact*. [Online]. Available: https://urbact.eu/athens. [Accessed: 13-Dec-2019].

[5] B. Stecanella, "What is TF-IDF?," *MonkeyLearn*, 2019. [Online]. Available: https://monkeylearn.com/blog/what-is-tf-idf/. [Accessed: 07-Dec-2019].

[6] Jason Brownlee, "A Gentle Introduction to the Bag-of-Words Model," *Machine Learning Mastery*, 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. [Accessed: 07-Dec-2019].

[7] S. Fan, "Understanding Word2Vec and Doc2Vec," *shuzhanfan.github.io*, 2018. [Online]. Available: https://shuzhanfan.github.io/2018/08/understanding-word2vec-and-doc2vec/. [Accessed: 07-Dec-2019].

[8] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," *towardsdatascience*, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47. [Accessed: 09-Dec-2019].

[9] J. Brownlee, "What Are Word Embeddings for Text?," *Machine Learning Mastery*, 2017. [Online]. Available: https://machinelearningmastery.com/what-are-word-embeddings/. [Accessed: 07-Dec-2019].

[10] E. L. Steven Bird, Ewan Klein, "Categorizing and Tagging Words," in *Natural Language Processing with Python*, O'Reilly Media, 2009, p. 504.

[11] "Topic model," *wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Topic_model. [Accessed: 18-Nov-2019].

[12] "Όρια Συνοικιών Δήμου Αθηναίων," *geodata.gov.gr*, 2019. [Online]. Available: http://geodata.gov.gr/el/dataset/op1a-euvo1k1wv. [Accessed: 18-Nov-2019].

[13] P. Pandey, "Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)," *medium*, 2018. [Online]. Available: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f. [Accessed: 22-Nov-2019].

[14] "Topic Modeling with Gensim (Python)," *machinelearningplus*. [Online]. Available: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/. [Accessed: 18-Nov-2019].

[15] "Lemmatisation." [Online]. Available: https://en.wikipedia.org/wiki/Lemmatisation.

[16] E. Fonseca, "State-of-the-art Multilingual Lemmatization," *towardsdatascience*. [Online]. Available:

https://towardsdatascience.com/state-of-the-art-multilingual-lemmatization-f303e8ff1a8. [Accessed: 13-Nov-2019].

[17] "Lexical analysis." [Online]. Available: https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization. [Accessed: 12-Nov-2019].

[18] P. Gil, "What Is Twitter & How Does It Work?," *lifewire*, 2019. [Online]. Available: https://www.lifewire.com/what-exactly-is-twitter-2483331. [Accessed: 28-Oct-2019].

[19] "Place Search," *Google Maps Platform*. [Online]. Available: https://developers.google.com/places/web-service/search. [Accessed: 29-Oct-2019].

[20] "Place Details," *Google Maps Platform*. [Online]. Available: https://developers.google.com/places/web-service/details. [Accessed: 29-Oct-2019].

[21] A. Ide, "A Turning Point for Tourism Informatics," *New Breeze*, vol. 29, no. 4, p. 4, 2017.

[22]. M. Roussou *et al.*, "Deliverable D4.1 - Indicators visualization and representation modelling and techniques," enventory - European eInfrastructures Observatory project report, 2011.

[23] E. Bruin, "Airbnb: The Amsterdam story with interactive maps," *kaggle*, 2019. [Online]. Available: https://www.kaggle.com/erikbruin/airbnb-the-amsterdam-story-with-interactive-maps. [Accessed: 21-Oct-2019].

[24] "Top 7 Data Science Use Cases in Travel," *ActiveWizards*. [Online]. Available: https://activewizards.com/blog/top-7-data-science-use-cases-in-travel/. [Accessed: 21-Oct-2019].

[25] M. Aggarwal, "Application Of Machine Learning and Deep Learning In the Hospitality Industry," *medium*, 2018. [Online]. Available: https://medium.com/@manuj.aggarwal/application-of-machine-learning-and-deep-learning-in-the-hospitality-industry-ca9675ce7b94. [Accessed: 21-Oct-2019].

[26] A. Bulanov, "Benefits of the Use of Machine Learning and AI in the Travel Industry," *djangostars*. [Online]. Available: https://djangostars.com/blog/benefits-of-the-use-of-machine-learning-and-ai-in-the-travel-industry/.

[27] M. Khan and S. S. Khan, "Data and information visualization methods, and interactive mechanisms: A survey," *Int. J. Comput. Appl.*, vol. 34, no. 1, pp. 1–14, 2011.

[28] GauravChhabra, "NLP - Twitter Sentiment Analysis Project," *kaggle*, 2019. [Online]. Available: https://www.kaggle.com/gauravchhabra/nlp-twitter-sentiment-analysis-project. [Accessed: 21-Oct-2019].

[29] H. M, "TripAdvisor Reviews - Data Analysis," *kaggle*, 2019. [Online]. Available: https://www.kaggle.com/harimo/tripadvisor-reviews-data-analysis. [Accessed: 21-Oct-2019].

[30] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste, "Methodology for Twitter Sentiment Analysis," *arXiv Prepr. arXiv1507.00955*, 2015.

[31] E. R. Tufte, *The visual display of quantitative information*, vol. 2. Graphics press Cheshire, CT, 2001.

[32] A. Mattaranz, "Applications of Text Analytics in the Tourism Industry," *meaning cloud*, 2017. [Online]. Available: https://www.meaningcloud.com/blog/applications-for-text-analytics-in-the-tourism. [Accessed: 21-Oct-2019].

[33] Gaurav Shankhdhar, "Sentiment Analysis Methodology," *edureka*, 2019. [Online]. Available: https://www.edureka.co/blog/sentiment-analysis-methodology/. [Accessed: 20-Oct-2019].

[34] "Sentiment Analysis The Only Guide You'll Ever Need," *MonkeyLearn*. [Online]. Available: https://monkeylearn.com/sentiment-analysis/.

[35] "Project," *Public participation city*. [Online]. Available: https://ppcity.eu/. [Accessed: 20-Oct-2019].

[36] "Tourism Sentiment Index," *destinationthink*. [Online]. Available: https://destinationthink.com/about-tsi/. [Accessed: 21-Oct-2019].

[37] J. Wu, "AI, Machine Learning, Deep Learning Explained Simply," *medium*, 2019. [Online]. Available: https://towardsdatascience.com/ai-machine-learning-deep-learning-explained-simply-7b553da5b960. [Accessed: 21-Oct-2019].

[38] B. Bettendorf, "NLP on Airbnb Data," *kaggle*, 2019. [Online]. Available: https://www.kaggle.com/brittabettendorf/nlp-on-airbnb-data. [Accessed: 21-Oct-2019].