Extraction and combination of multidimensional point-of-interest features for the classification of urban place types

Master's Thesis

Vasileios Milias

Extraction and combination of multidimensional point-of-interest features for the classification of urban place types

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by

Vasileios Milias
born in Athens, Greece

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Extraction and combination of multidimensional point-of-interest features for the classification of urban place types

Author: Vasileios Milias
Student id: 4614526
Email: milias.vasilis@gmail.com

Abstract

The digital representations of physical places, known as Points-Of-Interest (POIs), have been the core element of various studies and platforms such as online mapping services (e.g. Google Maps) and location based social networks (e.g. Foursquare). The use of POIs as proxies of the real-world-places facilitates the study of places, urban environments and, consequently, human behavior. Therefore, the extent to which the POIs manage to capture the complex multidimensional nature of physical places defines the limits of all those platforms and of humans' essential understanding of places.

Admittedly, the already existing POI data sources tend to represent differently the physical places (e.g. focus on specific aspects of places) and their data are being produced in a variety of ways (e.g. user generated data or non-user generated data). In addition, multiple sources exist that indirectly include place-related information as, for instance, Google Street View which contains images of the exterior of places without providing a direct link between the image and the corresponding place-entity. Thus, an interesting challenge arises which is how could all those diverse place-related data coming from different data sources be combined towards the creation of a better digital representation of places.

This thesis introduces an innovative approach to the extraction and combination of multidimensional POI features from various place-related data sources towards the study of urban places. It consists of two main parts: (1) the process of selecting, extracting and combining multidimensional POI features from various sources which reflect the high dimensional nature of places and (2) the use of the extracted features to discover which of those - and to what extent - better define and distinguish urban places in respect to their core characteristic, their main function.

Regarding the first part, for the combination of POI data sources a "matching" algorithm is developed whose goal is the identification of POIs which belong to different POI data sources and represent the same physical place and is based on the comparison of a set of attributes such as location, name and website. For the extraction of the POI features the need of specialized techniques according to the nature of the different data is revealed and several methods are discussed and used.

The second part concentrates in data collected from two capitals, Amsterdam and Athens. A machine learning classifier is trained on different combinations of features extracted from those data and their importance for distinguishing the urban place types is computed and compared. The results, among other, support that the functional (e.g. opening/closing times) and experiential characteristics (e.g. topics extracted from reviews) are the strongest indicators of a place's type independently of the context (e.g. city) while the exterior visual appearance of places does not provide such valuable information. The combination of the extracted features lead to an F1-score of around 60% when classifying POIs by their type among 10 classes (multiclass problem) and around 90% when predicting if a POI is of a certain type or not (binary problem).

Overall, the importance of combining multiple data sources in order to capture the complex nature of places is successfully supported by the results and the features that tend to better "describe" places in respect to their main function are discovered and further explored.

Thesis Committee:

Chair:	Prof. dr. ir. G.J.P.M. Houben, Faculty EEMCS, TU Delft
University supervisor:	Dr. ir. A. Bozzon, Faculty EEMCS, TU Delft
Daily supervisor:	Dr. A. Psyllidis, Faculty EEMCS, TU Delft
Committee Member:	Dr. Przemyslaw Pawelczak, Faculty EEMCS, TUDelft

Preface

"A place to put some remarks of a personal nature."

But... what is a place?

Vasileios Milias
Delft, the Netherlands
December 10, 2018

Acknowledgments

Before you lies my Master's thesis entitled "Extraction and combination of multidimensional point-of-interest features for the classification of urban place types". The searching of the "place" during the past nine months would have been definitely less productive without the physical or digital (long-distance) presence of some quite exceptional people to whom I feel the need to express my gratitude.

Firstly, I want to thank my supervisors for their excellent and to the point comments and suggestions. Particularly, I would like to give special thanks to Alessandro for his constructive criticism and to Achilleas for his great guidance and willingness to help me discover how research should be done.

Secondly, I want to thank the members of the sigma-team for their valuable feedback during our bi-weekly meetings. I frankly enjoyed those meetings and I feel that through them I learned not only how to provide the right answers but also how to pose the right questions.

Lastly, I want to thank my friends and family who ended up listening about "POIs" and "place" a lot more than what they would truly like. Each one of them supported me during the whole master with his/her own, unique, irreplaceable way.

Vasileios Milias
Delft, the Netherlands
December 10, 2018

Contents

List of Figures

Chapter 1

Introduction

1.1 Background and Problem Definition

Over the past years, the digital representation of real-world places as Points of Interest (POIs), which is an established term used by several web platforms, has increasingly attracted the interest of both the research community and the industry. Multiple scholars have based their studies on POI datasets [38], [42], [16] and a whole economy has been built around them which could be observed by the constantly growing amount of restrictions regarding the availability of such data [35].

A simple example of POI data use could be found in Google Maps. Google Maps basically consists of a collection of POIs pinned on a map (e.g. bars, restaurants, universities) and a collection of services which are mostly based on the POIs' features (e.g. showing nearby restaurants, opening times or giving directions to arrive at a physical place represented by a POI). Undoubtedly, POI is not only one of the core elements of online mapping services (e.g. Google Maps and OpenStreetMap) [1] but also a core element of location based social networks (e.g. Foursquare and Facebook places) and geographic information systems (e.g. QGIS [2] and ArcGIS [3]). Consequently, all those platforms' quality is inextricably linked with the POIs' quality. Moreover, POIs are used as proxies of physical places in disciplines such as urban planning and urban design [5], [61]. At the same time, people search for places' information on digital maps (e.g. Google Maps) and search engines (e.g. Google) on a daily basis. According to Wikipedia, Google Maps is one of the most used applications [4]. All those place-based queries, called "spatial queries", are being clearly also bounded by the amount of information integrated in the POIs.

Despite their proven importance and usefulness there is still a significant information gap between places and POIs since there is not a single POI data source which captures all the different dimensions of places. POIs' main components vary per platform and as a result the various POI datasets often focus on different places' dimensions. For instance, Foursquare POIs include information about how many users have checked-in to a place while Google POIs provide information about how crowded a

[1] https://www.openstreetmap.org

[2] https://www.qgis.org/en/site/

[3] https://www.arcgis.com/index.html

[4] https://en.wikipedia.org/wiki/List_of_most_popular_smartphone_apps

place is per hour[1]. Moreover, since the different POI platforms use different methods to collect their data it is not uncommon for POIs which represent the same physical place and exist in different platforms, to complement each other. An example would be two POIs, one coming from Google and another one coming from Foursquare, which both represent the same physical place and the former is missing the "address" of the place while the latter is missing the place's "opening hours". A POI which combines both POIs' information would be more complete than each one of them. In addition, other data sources, which are not explicitly considered to be "POI data sources", could indirectly include place-related information which is not being exploited. For example, a place's facade could be depicted in a Google Street View image without an explicit connection between the image and the place (e.g. through a tag) or people might be talking about their experience at a place on Twitter without including a place-tag in their tweets (figure 1.1). This variety of included information in the different POI and place-related data sources arises the need of combining sources in order to gain a more complete digital representation of places [36].

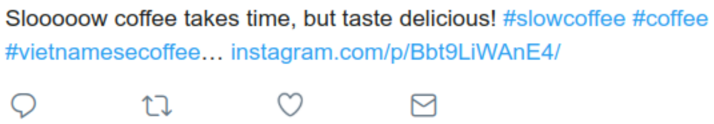

Sloooooow coffee takes time, but taste delicious! #slowcoffee #coffee #vietnamesecoffee… instagram.com/p/Bbt9LiWAnE4/

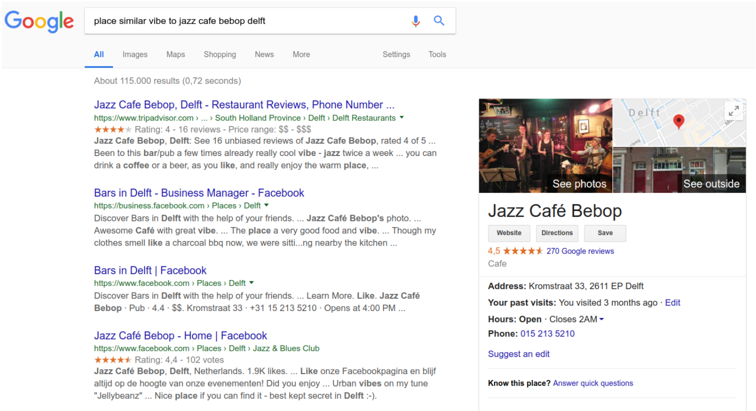
Figure 1.1: Example of tweet which contains information about a place without including a tag to link the tweet with the place (i.e. which is the coffee place the user refers to).

Indisputably, combining all those pieces of place-related information coming from different data sources is a challenging process. For the combination of the POI data sources, which directly include information about places, the challenge lies mostly in identifying the POIs which belong to different sources and represent the same physical place, a process which is called "matching". This process is not straightforward as the geolocation of the POIs deviates from one data source to another and the matching should be based on the comparison of other attributes as well (e.g. name or address) [36], [23]. For the indirect POI data, the challenge is to first identify the reference to the physical place (i.e. POI parsing) and then to extract meaningful information about it. The conversion of each different kind of data to utilizable information requires the use of various specialized techniques and algorithms. For example, extracting place-related features is completely different when dealing with images than when dealing with unstructured text.

Even though challenging, this combination could lead to a more thorough understanding of places [36] which takes the high dimensional nature of "place" into account. In addition, it could improve the way people not only study places but also "interact" with POIs in the digital world. For instance, currently POIs cannot support complex spatial queries which imitate the way people pose questions to each other and are not location-centric. A query for a place which "has a similar vibe" to another place most of the times fails. Such an example could be seen in figure 1.2 in which the returned results do not include places of similar vibe to "Jazz cafe Bebop" but results

[1] This feature is called "Popular Times"

which seem to be only based on the keywords "Jazz", "cafe" and "Bebop". This "fail" is understandable if one thinks how complex the nature of "place" is. Admittedly, to understand what are the POI features which should be combined for the POIs to reflect the multidimensional nature of a place one has to first understand what "place" truly is.



place similar vibe to jazz cafe bebop delft

About 115.000 results (0,72 seconds)

Jazz Cafe Bebop, Delft - Restaurant Reviews, Phone Number ...
https://www.tripadvisor.com › ... › South Holland Province › Delft › Delft Restaurants
★★★★☆ Rating: 4 - 16 reviews - Price range: $$ - $$$
Jazz Cafe Bebop, **Delft**: See 16 unbiased reviews of **Jazz Cafe Bebop**, rated 4 of 5 ... Been to this **bar**/pub a few times already really cool **vibe** - **jazz** twice a week ... you can drink a **coffee** or a beer, as you **like**, and really enjoy the warm **place**, ...

Bars in Delft - Business Manager - Facebook
https://business.facebook.com › Places › Delft
Discover Bars in **Delft** with the help of your friends. ... **Jazz Café Bebop's** photo. ... Awesome **Café** with great **vibe**. ... The **place** a very good food and **vibe**. ... Though my clothes smell **like** a charcoal bbq now, we were sitti...ng nearby the kitchen ...

Bars in Delft | Facebook
https://www.facebook.com › Places › Delft
Discover Bars in **Delft** with the help of your friends. ... Learn More. **Like**. **Jazz Café Bebop** · Pub · 4.4 · $$. Kromstraat 33 · +31 15 213 5210 · Opens at 4:00 PM ...

Jazz Café Bebop - Home | Facebook
https://www.facebook.com › Places › Delft › Jazz & Blues Club
★★★★★ Rating: 4,4 - 102 votes
Jazz Café Bebop, **Delft**, Netherlands. 1.9K likes. ... **Like** onze Facebookpagina en blijf altijd op de hoogte van onze evenementen! Did you enjoy ... Urban **vibes** on my tune "Jellybeanz" ... Nice **place** if you can find it - best kept secret in **Delft** :-).

Jazz Café Bebop
4,5 ★★★★★ 270 Google reviews
Cafe
Address: Kromstraat 33, 2611 EP Delft
Your past visits: You visited 3 months ago · Edit
Hours: Open · Closes 2AM
Phone: 015 213 5210
Suggest an edit
Know this place? Answer quick questions

Figure 1.2: Example of failed query imitating the way people pose questions. The used "similar vibe" doesn't truly affect the returned results.

On a theoretical level, the definition of "place" has been thoroughly discussed. To quote Yi-Fu Tuan *"Place is not only a fact to be explained in the broader frame of space, but it is also a reality to be clarified and understood from the perspective of the people who have given it meaning"* [55]. Through this sentence Yi-Fu Tuan reveals not only the complex nature of the word "place", but also the importance of human perception in the process of understanding what "place" is. So, what really makes a "place"? Various researchers from multiple disciplines such as geography, environmental psychology and urban planning have focused on this subject [11], [13], [51], [24]. The most widely used place-concept in the literature which tries to tackle this delicate question is the "Sense Of Place" (SOP) [20], [11], [50], [22], [2]. Unfortunately, SOP does not have an official, universally accepted definition, mostly due to its ambiguous nature. In this work, it is used as expressed in [50] by Richard C. Stedman: *"..sense of place is not intrinsic to the physical setting itself, but resides in human interpretations of the setting, which are constructed through experience with it. Spaces become "places" as they become imbued with meaning through lived experience* [54]." In a more simplified way, SOP describes the combination of characteristics that humans perceive or relate to a certain portion of space making it distinguishable from other places.

Thus, an important aspect to be explored is whether the combination of POI features coming from place-related data sources could facilitate an improved understanding of the places' characteristics as those are defined by the SOP.

1.2 Research Objectives

The objective of this research is twofold; first, it focuses on the combination and extraction of multidimensional POI features from various web sources. Second, by using the combined and extracted POI features as proxies of real-world places, it aims to explore which of those features contribute – and to what extent – to the classification of urban place types. "What makes a place" is a key question in this study, where "place" is defined by its core characteristic, its main function, which could be, for example "restaurant", "bar" or "university".

Obviously, not all of the different dimensions of the SOP could be extracted from the currently available data. For instance, smell and personal memories influence significantly how a certain place is perceived but gaining data which include these pieces of information is nearly impossible. This research concentrates in five categories of features which include what could be considered as the main SOP's features and could be extracted from new forms of web data. Those are the **functional features** which are relevant to the main function of a place (e.g. place type, the opening times, the website etc), the **experiential features** which concern people's experience when visiting a place (e.g. reviews and ratings), the **socio-topical features** which include social aspects of a place (e.g. popularity and topics discussed by the people around), the **visual features** which are relevant to the visual appearance of the exterior of a place and the **locational features** (e.g. nearby places) which are based on the location of a place. The main goal of this thesis is formulated into the following research question:

MRQ: How to combine and extract multidimensional POI features from various web sources to classify urban place types?

The main research question is broken down into four research sub-questions:

- **RQ1:** What are the current state of the art methods for the extraction of POI features from place-related data?
- **RQ2:** How to combine POI features from various web data sources?
- **RQ3:** How to extract multidimensional POI features from combined data sets?
- **RQ4:** Which POI features contribute the most to the classification of urban place types?

1.3 Methods

In this section, the methods used to address the research questions of this thesis are briefly presented. A more thorough description of those methods and the reasoning behind the choices being made could be found in the next chapters.

RQ1: *What are the current state of the art methods for the extraction of POI features from place-related data?*

Regarding the first research question, previously used methods and state of the art techniques concerning the extraction of POI features are reviewed and compared and this thesis' additions to the related work are discussed. Focus is given on the extraction of those features from new forms of web data and not from traditional geospatial data sources.

The collection of the previous works is divided in three groups. The first group consists of theoretical approaches and proposed methods for a more in-depth understanding of places through their digital representations. For instance, the significance of GIScience and Big Data towards this goal is discussed among those papers.

The second group includes works in which the scholars used a single data source such as Twitter, Foursquare or Yelp to extract and study POI features. The selection of those papers was also based on presenting a variety of different data sources used in the same direction.

In the third group, the selected works use the combination of multiple data sources. The methods of combining the various sources vary significantly. In some works, the researchers deal with regions, hence the geo-tagged data are combined and clustered based on their distance, while in others a more direct matching between the POIs coming from different sources is realized.

RQ2: *How to combine POI features from various web data sources?*
RQ3: *How to extract multidimensional POI features from combined data sets?*

Regarding the second and third research question three main steps are followed: the data **collection**, the data **matching** and the POI features **extraction**. Even though those three steps are being discussed consecutively, the first step is directly dependent on the other two. The design of the data pipeline which leads to the extraction of the POI features is presented in figure 1.3 including the classification step which is a core part for tackling the fourth research question of this thesis.

The **data collection** step, includes finding which are the available and suitable place-related web data sources for this thesis' objective. The selection process of the data sources is fundamental for this research, as the quality and richness of the collected data defines this work's limitations. The selected data sources include directly, or indirectly through some processing , all the aforementioned POI features. The eventually selected data sources are: **Google**, **Google Street View** (GSV), **Foursquare** and **Twitter**. For each place, its Google and Foursquare representations (POIs), the 360 panoramic Google Street View image which is taken from the outside of the place and the tweets that have been sent around that place are gathered, if existing. Those data sources are diverse in terms of avoiding overlapping information but they also contain some mutual information which is essential for the next step, the matching.

As previously stated, **matching** refers to the identification of the POIs that represent the same physical place over different data sources. Combining Google with Google Street View and Foursquare with Twitter is quite straightforward as the two sources in each pair use the same geolocation system. Thus, by having a place as a Google POI, and therefore having its latitude and longtitude, it is easy to gather the respective Google Street View images outside of that place. Commonly, by having a place as a Foursquare POI, one could collect the tweets which have been sent from around that place. The challenging part is the matching of the two pairs, Google-GSV and Foursquare-Twitter. The sources which are selected to be used for the matching of those pairs are Google and Foursquare as they contain some mutual POI features which could be compared. Thus, a matching algorithm which is based on the geographical

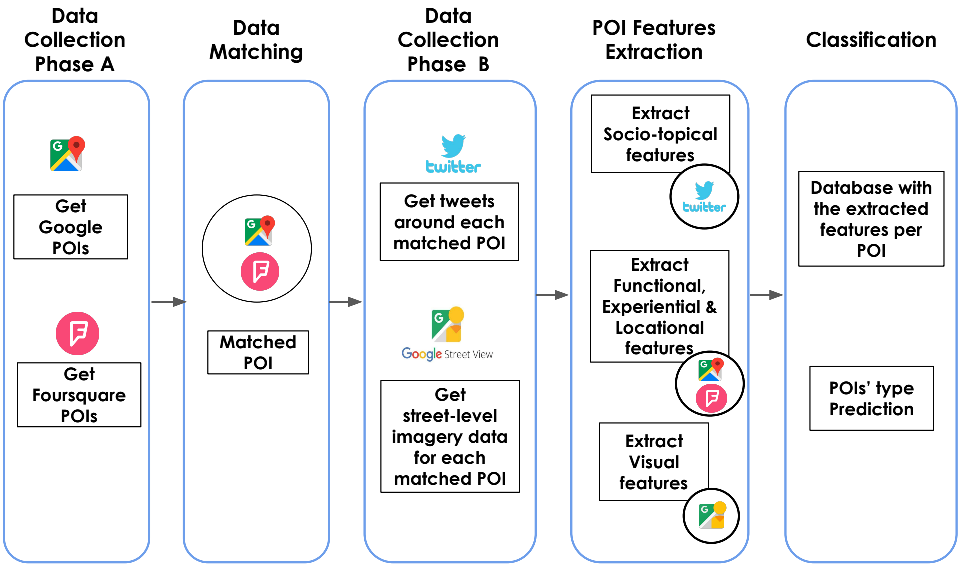


Figure 1.3: Data pipeline for the extraction of POI features and the prediction of the POI types. The pipeline could be divided into four steps: the **Data Collection** (phase A and B), **the Data Matching**, the **POI features extraction** and the **Classification** step. Firstly, the Google and Foursquare POIs are collected. Then, they are matched and the geo-coordinates of each matched POI are used to retrieve the tweets that have been sent from around this location and the street-level images from this exact location. After that, the POI features are extracted and, lastly, the classification is realized based on the extracted features.

distance, several POI features such as name, website, phone number, address and on various, previously used string similarity metrics [36], [23] is created for the matching of Google and Foursquare POIs and, consequently, the matching of all those four sources.

In the **extraction** step, the data are transformed into the functional, socio-topical, experiential, visual and locational POI features in the following ways:

- **Functional features**: extracted from Google and Foursquare POIs. The main functional features could be obtained directly from the POIs without the need of intermediate steps.

- **Socio-topical features**: extracted from Twitter. The tweets, that are gathered for each place, are analyzed and aggregated so that socio-topical features are discovered such as social media activity (how many tweets have been sent from a place) and people's (social media) sentiment. In addition, a topic model is used to detect the topics which are being discussed around each place.

- **Experiential features**: also based on Google and Foursquare POIs. Several of those attributes could come straight from the POIs such as the number of reviews, photos and likes. Moreover, a topic model is again used to extract topics from the Google Reviews.

- **Visual features**: based on Google Street View's 360 panoramic images from the outside of each place. Those images are annotated using image processing deep learning pre-trained models. The goal of the models used is to recognize scenes and detect objects. For the scene recognition part an example could be an image which is annotated as *"residential neighborhood* or *bazar*. In the object detection part, the models identify objects such as buildings and trees in each image.
- **Locational features**: based on Foursquare POIs by calculating the amount of nearby places per type (e.g. how many bars or restaurants exist in a distance of 500m or 1000m from each place).

RQ4: *Which POI features contribute the most to the classification of urban place types?*

The fourth research question is being addressed by developing the pipeline which is presented in figure 1.3. For the POI types prediction a machine learning approach, which is based on the extracted POI features, is followed. The solutions to various unexpected "problems" which were brought to light during the implementation of this pipeline are also being presented. The previously discussed data sources and techniques are used in the context of two case studies: one for Amsterdam (Netherlands) and another one for Athens (Greece). The main selected POI types which are being studied under the two case studies are: restaurant, bar, clothing store, gym, hotel, food and drink shop, cafe, college and university, coffee shop and art gallery. This selection is being carefully explained in the next chapters.

The **data collection** step, is based on each source's provided API: Google Places API [1], Foursquare Places API [2], Google Street View API [3] and Twitter's API [4]. Firstly, POI data from each city are collected from Google and Foursquare. Then, those POI data are matched and for each matched POI the Google Street View images and the tweets are gathered.

The developed **matching** algorithm works in the following way: for each Foursquare POI (FPOI), the Google POIs (GPOI) which are geolocated in a distance within 300 meters are retrieved. Then several similarity scores are computed between the FPOI and every retrieved GPOI. These scores are based not only on the POIs' geographical distance, even if this could seem enough it is not uncommon for the geographically closest POIs coming from two sources to not actually represent the same physical place [36], [28], but also on the string similarity distance of multiple features such as the name, address, phone and email. Depending on various combinations of those scores (e.g. name similarity score and address similarity score) and some specified thresholds this algorithm decides if the POIs should be matched or not. This process is thoroughly explained in the *Experiment Design* chapter in which a visualization of this algorithm is also presented (figure 3.2).

For the data **extraction**, the above-mentioned features are obtained in the following ways:

[1] https://developers.google.com/places/web-service/intro

[2] https://developer.foursquare.com/places-api

[3] https://developers.google.com/maps/documentation/streetview/intro

[4] https://developer.twitter.com/en/docs/tweets/search/overview

- **Functional features**: extracted from the POIs collected from Google and Foursquare APIs respectively. The main extracted functional features are the places' function, website, phone number, social media, price and opening times.

- **Socio-topical features**: extracted from Twitter by gathering tweets which were tweeted in a radius of 50m from each place, in the past two years. Various tweets' statistics such as the total amount of tweets and the average number of words used in the tweets are computed. Additionally, the Latent Dirichlet Allocation (LDA) topic model is used for the extraction of topics from the tweets [6]. Due to the fact that the tweets' statistics and LDA are language specific the tweets had to be filtered by language. Thus, the english and dutch tweets and the english and greek tweets were used for Amsterdam and Athens respectively. The information loss from not including the rest of the languages does not seem significant as the vast majority of the tweets belongs in the selected languages.

- **Experiential features**: extracted from Google and Foursquare POIs. They consist of features such as likes, ratings and similar features to the ones extracted from the tweets which in this case are extracted from the Google reviews.

- **Visual Features**: extracted from Google Street View's 360 panoramic images which were taken from the outside of each place. The used image processing algorithms are: Places-CNNs for scene detection [62], [63], (classifies an image in scene categories like "residential neighborhood" or "industrial area") and two models for object detection (e.g. detecting trees, buildings) using Google's Tensorflow Object Detection API [19] from which one was trained on the Open Images dataset [25] and the other one on the COCO dataset [29].

- **Locational features**: based on Foursquare POIs, the amount of nearby POI types is calculated in a distance of 100m, 1km and 3km. This calculation is realized with the use of a PostGIS special function [1].

Finally, two approaches are followed for the **classification** of the POI types: (1) distinguishing POI types by predicting what is the type of a POI among a specified set [2] (multiclass problem) and (2) analyzing specific POI types by predicting if a POI is of a certain type [3] or not (binary problem). Both cases concentrate in the comparison of the contribution of the extracted features to the classification of urban place types on a group (feature sets) and an individual (features) level. The results of this part support qualitatively the correctness of the previous steps and provide insights about the features that "make" a place.

[1] ST_Within: https://postgis.net/docs/PostGIS_Special_Functions_Index.html

[2] The most used set of types consists of: art gallery, bar, cafe, clothing store, coffee shop, college and university, food and drink shop, gym, hotel, restaurant.

[3] Focus is given on three types: clothing store, hotel and restaurant. The reasoning behind this selection is explained in the following sections.

1.4 Contributions

Supporting open knowledge, this thesis' research, code and datasets are openly shared for everyone to use [1]. The main contributions of this thesis are:

- A generated dataset of the matched Google, Foursquare, Google Street View and Twitter data.

  This dataset includes 4536 Google and Foursquare POIs from Amsterdam which are matched with a precision equal to 0.97% and 2501 Google and Foursquare POIs from Athens which are matched with a precision equal to 98%. In addition, this dataset includes tweets which were tweeted around those places (50m radius) in a time period from 01/01/2017 to 10/10/2018 and the most recent Google Street View's 360 panoramic images which were taken from the outside of each place. All those data have not been processed and could be used in numerous ways. The database used for storing this information is a postGIS database [2] and each place has been stored as a spatial object.

- A generated dataset which contains all the extracted features per POI.

  This dataset includes the previously mentioned features which are extracted from each source and stored separately for all the matched POIs. Thus, it could be used either per source, for instance only the features extracted from Google Street View or from Twitter, or combined. Again, the database used for the storing of the extracted features is a postGIS database.

- A data-driven approach for the extraction of urban place attributes from the combination of place-based web data sources.

  This approach includes the data collection methods, the POIs features extraction methods to transform the data into meaningful features and the matching methods for the combination of the different data sources.

- A methodology on how to quantify which attributes function as significant indicators of the POIs' type based on data coming from Amsterdam and Athens.

  The discussed methodology is implemented and the various urban place features are extracted from a set of specified POI types. Then, those features are used for the prediction of the POI types using machine learning techniques. The results, suggest which attributes are good indicators for the POI type classification and depict the particularities of each studied region.

1.5 Outline

The rest of this thesis is organized as follows: in the next chapter the related work is presented and in chapter 3, the *Experiment Design*, the selection of the data sources, the matching algorithm and the methods for the extraction of the POI features are being thoroughly explained. Chapter 4, the *Implementation*, contains the implementation

[1] https://github.com/MiliasV

[2] https://postgis.net/

details of the pipeline discussed in chapter 3 and the results of the algorithms used for the POI type prediction. The two final chapters are the *Discussion* and the *Conclusions*. In the former the results of the case studies and the threats to validity are discussed while the latter contains the conclusions and the future work.

Chapter 2

Related Work

In this chapter, previously used state of the art techniques for the extraction of urban place attributes are being discussed. The goal of this chapter is to tackle the first research question of this thesis:

RQ1: *What are the current state of the art methods for the study and extraction of place attributes from place-related data?*

Various researchers have tried to better understand and study places by extracting their attributes using either traditional or new forms of data. In the following sections, an overview of some of the most interesting approaches is being presented. For the selection of those papers special focus was given on the novelty of the followed approaches, on having a diversity in the used data sources and on the works which were inspirational for the realization of this thesis. The previous studies have been grouped into the following three sections and in tables 2.1, 2.2 and 2.3 the main papers discussed in each section are presented:

I. Proposed methods for the study and extraction of urban place attributes
II. Extraction of urban place attributes based on a single data source
III. Extraction of urban place attributes based on multiple data sources

2.1 Proposed methods for the study of urban environment based on POIs

The understanding of urban environment has been the subject of numerous researches from various disciplines such as geography, psychology, architecture, planning and computer science. Due to the vagueness of the "urban experience" concept and the difficulty on quantifying the importance of the urban places' attributes, multiple approaches towards this goal have been followed. In the past, those approaches followed traditional methods such as Lynch's urban cognitive maps [30] for which he used small samples of residents from Boston, Los Angeles and New Jersey (US) and Stedman's work on the contribution of the physical environment to the "sense of place" for which he used a mail survey of 1000 property owners [50]. In recent years, the exponential growth of geotagged user generated content and, in a more generic way, the awareness of the existence of large amount of information in the new forms of place-based

Title	Authors	Year	Ref.
Operationalising "Sense of Place" as a Cognitive Operator for Semantics in Place-Base Ontologies	P. Agarwal	2005	[4]
The convergence of GIS and social media: Challenges for GIScience	D.Suia, M.Goodchild	2011	[53]
Towards Platial Joins and Buffers in Place-Based GIS	S.Gao, K.Janowicz, G.McKenzie, L.Li	2013	[15]
POI Pulse	G.McKenzie, S.Gao, K.Janowicz, J.Yang	2015	[38]
Leveraging Big (Geo) Data with (Geo) Visual Analytics: Place as the Next Frontier	A.M. MacEachren	2017	[31]
Using Semantic Signatures for Social Sensing in Urban Environments	K.Janowicz, S.Gao, Y.Hu, R.Zhu, G.McKenzie	2018	[23]
OpenPOI: An Open Place of Interest Platform	G.McKenzie, K.Janowicz	2018	[35]

Table 2.1: Main discussed papers of Section I

data lead to numerous studies which propose several methods on how to use these new forms of data to study the urban environments.

Several attempts follow a quite theoretical approach. For instance, in [4], P.Argawal suggests that the operationalisation of the "sense of place" could lead to the better understanding of the cognitive dimensions of place. The "sense of place" is *"considered as a mechanism through which individual conceptualisations of place are grounded in a collective notion, defining the meanings of place and its links in the real world."* To identify the key aspects for defining a "cognitive sense of place" the author conducted two small size experiments (sample size of 50) from which she tried to show how human based experiments could help in the formalization of the vague "sense of place" concept and in the creation of a more strict definition of "place". The results of this study show the following three significant factors which better define the conceptualization of "place": distance in space, degree of familiarity and if the "place" is "in the neighborhood".

In a more technical proposed approach , D. Sui *et al.* discuss the relationship between GIS and social media and support that a convergence of the two could lead to the connection of "*the world of space (traditional GIS) and the world of place (social media)*" [53]. They characterize the future of GIS as "unpredictable" but they embrace the continuous transformation of GIS because of the new ways social media are used.

Towards the improvement of place-based GIS, Song Gao *et al.* introduce two place-based GIS operations: platial join and platial buffer [15]. "Platial join" is an operation whose purpose is to merge place entities from different sources using semantics while "platial buffer" operation is similar to the "spatial buffer" which *"involves the creation of new polygons from points, polylines, and polygons according to a specified to identify nearby features"* but based not merely on spatial features but also on the semantic relationships of the places. Those operations are of great interest as they

Title	Data	Authors	Year	Ref.
Empirical study of topic modeling in Twitter	Twitter	L.Hong, B.Davison	2010	[17]
Automatic Improvement of Point-of-Interest Tags For OpenStreetMap Data	OSM	S.Storandt, S.Funke	2015	[52]
How where is when ? On the regional variability and resolution of geosocial temporal signatures for points of interest	Foursquare	G.Mckenzie, K.Janowicz, S.Gao, L.Gong	2015	[37]
From ITDL to Place2Vec - Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts	Yelp	B.Yan, K.Janowicz, G.Mai, S.Gao	2017	[59]
A data-driven approach to exploring similarities of tourist attractions through online reviews	TripAdvisor	G.McKenzie, B.Adams	2018	[33]
Identifying spatiotemporal urban activities through linguistic signatures	Twitter	C.Fu, G.McKenzie, V. Frias-Martinez, K.Stewart	2018	[14]

Table 2.2: Main discussed papers of Section II

would be very helpful for combining POI data coming from different data sources.

A. Maceachren in [31] works also for the evolution of GIScience focusing on the significance of Big Data. In his paper, he depicts the power both of the (geo)visual analytics and the place-relevant data which are hidden in unstructured text. In his conclusions, he also shows some consideration on the ethical part of formalizing places and he proposes for future studies to focus on two main targets: making "place" as important as "space" in the GIScience and use huge amount of interconnected data to better understand places and relations among them.

Similarly, G. McKenie *et al.* emphasized the importance of Big Data in [38], in which they suggest a technical and theoretical framework for the interactive study of a city's pulse based on social media. Their inspiration came by the Foursquare's pulse videos [1] and their goal is to create interactive visualizations of the pulse of a city and discover the Big Data limitations of those visualizations (e.g. what are the limitations when rendering a vast amount of POIs in an interactive digital map). This work is a valuable example on how user generated content (UGC) could be used for the interactive study of urban environments.

A different approach for an improved digital representation of places is based on a technique called "semantic signatures" which is proposed by K. Janowicz *et al.* in [21]. The "semantic signatures" are being used for the extraction of places' attributes from human data traces and they are divided in three categories: spatial, temporal and

[1] https://foursquare.com/infographics/pulse

Title	Data	Authors	Year	Ref.
Weighted Multi-Attribute Matching of User-Generated Points of Interests	Yelp, Foursquare	G.McKenzie, B.Adams	2013	[36]
Mining point-of-interest data from social networks for urban land use classification and disaggregation	Yahoo, Dun and Bradstreet, infoUSA	S.Jiang, A.Alves, F.Rodrigues, J.Ferreira, F.C.Pereira	2015	[23]
Thematic signatures for cleansing and enriching place-related linked datas	Wikipedia, DBpedia	B.Adams, K.Janowicz	2015	[3]
Juxtaposing thematic regions derived from spatial and platial user-generated content	Foursquare, Twitter,Yik-Yak, Instagram	G.McKenzie, B.Adams	2017	[34]
The Language of Place : Semantic Value from Geospatial Context	OSM, Twitter, Google	A,Cocos, C.Callison-Burch	2017	[10]
Crowdsourcing a collective sense of place	Twitter, Wikipedia	A.Jenkins, A.Croitoru, A.Crooks, A.Stefanidis	2016	[22]
A data-synthesis-driven method for detecting and extracting vague cognitive regions	Flickr, Instagram, Twitter, Travel-Blogs.org, Wikipedia	S.Gao, K.Janowicz, D.Montello, Y.Hu, J.Yang, G.McKenzie, Y.Ju, L.Gong, B.Adams, B.Yan	2017	[16]
xNet+SC: Classifying Places Based on Images by Incorporating Spatial Contexts	Yelp, Google, Google-Street-View	B.Yan, K.Janowicz, G.Mai, R.Zhu	2018	[60]

Table 2.3: Main discussed papers of Section III

thematic signatures. The spatial signatures study attributes such as the places' geospatial distribution, the temporal signatures study attributes such as at which time-slots is a place mostly visited and the thematic signatures study semantic attributes that have to do with human experiences. For the thematic signatures, this work suggests using text (e.g. reviews or social media posts) and techniques such as the term frequency and inverse document frequency (TF-IDF) [32] or the, *"more advanced approach"*, Latent Dirichlet Allocation (LDA) [6] to extract topics and relate them to the various place types. The authors also propose new ways to compare place types based on the semantic signatures. Particularly, they represent each place type as a vector based on numeric values which were extracted from the above-mentioned attributes and they use the cosine similarity and the Jensen-Shannon divergence (JSD) to quantify the similarity of the different place types. Their work, depicts the significance of combining multidimensional data sources and extracting place-based attributes by using specialized techniques and methods.

Finally, an other interesting approach which focuses on the advancement and usability of the POIs proposes the implementation of the *"OpenPOI Platform: a dataset* *and service for storing, sharing, and interacting with a common set of places of interest"* [35]. The authors, consider this platform as a "*research enabler*" for scholars from a variety of disciplines. This study embraces the idea of "open knowledge" and emerges the need of an open, multidimensional POI platform which would be valuable for the research community to ensure the access to geospatioal information.

All the works included in this section, emphasize the importance of "place" and contribute to its understanding by following different and innovative approaches. In the next section, studies on which a single new form of place-based data source is being used for the extraction of specific place attributes are presented.

2.2 Extraction of urban place attributes from a single data source

Multiple studies have been focused on the extraction of several attributes from specific data sources such as Foursquare, Twitter or Wikipedia. Some of the most advanced and interesting researches, in which the extracted attributes originate from a single data source, are being presented in this section.

In [17], the authors investigate the potential of topic modelling in short text environments by using LDA, author-topic model [48] and TF-IDF on Twitter. Their experiments involved the prediction of popular messages and the classification of users and messages into topical categories. One of their worth mentioning insights, is that the length of a document influences significantly the quality of the topic model. Consequently, they suggest that aggregating short texts (or tweets in that case) leads to better models. Additionally, they concluded that TF-IDF scores could be proved more efficient when content information is large. The scholars in [52], deal with even shorter texts as they attempt to infer place type tags for POIs in OpenStreetMap just by using the POIs' names. The core idea of this paper is that in multiple cases there are words and phrases in the POIs' names which could work as indicators for the places' function and attributes. For their experiments they used POIs which had been tagged with *Amenity and Cuisine* tags, *Other Amenity and Shop* tags or *Tourism and Leisure* tags. Their results, imply that significant amount of information could be extracted from the POIs' names.

A rather different and interesting approach is followed by B. Yan *et al.* in [59], where they developed information theoretic and distance-lagged augmented spatial contexts which outperformed the state-of-the-art word embeddings. In addition, in their evaluation process which is based on crowdsourcing, they discovered that similarity patterns derived from their method correlated significantly with human similarity judgments. They also, proposed three evaluation systems to test POI embeddings and they published the results of the human intelligence tasks they used as well as the embeddings to facilitate the reusability of their research.

The extraction of cities' attributes has also been attempted by using TripAdvisor. G.McKenzie *et al.* in [33] used TripAdvisor's online reviews to deeper understand the similarities between cities and attractions through user generated content. Text is again the core source of this research and several textual analysis methods are used for the features' extraction such as cleaning techniques (e.g. Porter stemming algorithm [45], removing non-alphanumeric characters etc), creation of bag-of-words per specific categories (e.g. by attraction or by city), LDA and Word2Vec [40]. An interesting insight gained from this work is that reviewers have a tendency to *"...write in a manner that is linguistically similar to an attraction in their home country... At a minimum, these results confirm the notion that there is a degree of either conscious or unconscious nationalism or ethnocentrism present in travel review platforms."* Once more, the place-based user generated content proved to be a rich source of latent information.

As part of a series of papers, in 2015 McKenzie *et al.* studied the regional variability from temporal signatures using Foursquare [37]. The "temporal signatures" in this study refer to the user check-ins in Foursquare. The results support that there are regional differences in the temporal signatures. As the authors state, this paper is *part of a long term project to publish an openly available library of semantic signatures with the hope that it will be equally as transformative as spectral signature libraries have been to the field of remote sensing.* Later, in 2018, in a paper which is also co-written by McKenzie "linguistic signatures" are being used for the identification of spatio-temporal urban activities [14]. Those signatures are extracted from Twitter using, instead of the more common used LDA technique, the ST-LDA [18]. This selection is based on the fact that the length of the text of each tweet is short and since the authors did not want to aggregate the tweets, ST-LDA was a more reasonable choice as it assigns only one topic per tweet. In their results, they suggest that by analyzing users' online posts of activities it is possible to understand peoples' actual activities, in the real wold. In their conclusions, they also state clearly that to remove the bias introduced when using only one platform (e.g. Twitter) an option is to integrate different data sources which have different biases. This idea exists in the core of this thesis as well. In the next section, two more papers which follow the same direction (extraction of place signatures) are presented.

2.3 Extraction of place attributes from multiple data sources

The latest years, the popularity of collecting data from multiple sources and combining them in order to gain a better understanding of places and regions is increasing. In this section, as before, some of the most interesting and different approaches are being presented for the extraction of place-based attributes from the combination of multiple data sources.

Once again, in 2015 B.Adams and K.Janowicz dealt with the extraction of place signatures [3]. In this work, a text based approach is followed using DBpedia and Wikipedia's information for the creation of places' thematic signatures. The combination of DBpedia and Wikipedia is realized in the following way: after collecting entities from the DBpedia ontology, natural language processing techniques (e.g. topic modeling) are being used to process text coming from Wikipedia and associated with those entities. Two years later, another similar research in which thematic regions are being derived from the combination of Foursquare, Twitter, Yik-Yak[1] and Instagram was conducted by G.McKenzie *et al.* [34]. For the data collection part, 37,302 POIs belonging in one of twenty specific categories and geolocated in Los Angeles were

[1] Mobile application allowing users to post anonymous content to other users within a 5 mile radius of their location.

collected from Foursquare. In a similar way, geo-tagged social media posts were collected for the same region from Twitter, Yik-Yak and Instagram. Interestingly enough, in order to classify the thematic regions the authors focus on places' popularity which, as they argue, plays a crucial role "*...a popular bar, should contribute more to defining a bar region than a venue that has had little to no visitors in the past year*". This work concludes with a very interesting question which is being examined in this thesis as well: "*Does having access to user-contributed geographic content enhance our understanding of the relationship between space and place?.*"

A.Cocos *et al.* in [10], researched at which extend the semantically similar words occur over the same geospatial context. After collecting tweets written in English and coming from 20 specified metro areas they derived from their text, word embeddings. Then they used OSM and Google Places' entities which were located within a distance of 50m from each tweet, to geospatially enrich their context. Their results suggest that geographic context is a valuable source for studying semantic relatedness.

A.Jenkins *et al.* in [22], introduce the term "*collective sense of place*" to describe that they study sense of place on a collective level and not on an individual level. Their goal, is to better understand "place" through crowd-generated content. The used data sources in this work are Twitter and Wikipedia. The authors focused on four major cities [1] from which they collected tweets and Wikipedia entries. Then, they made a comparison between the thematic-spatial clusters extracted from Twitter and the thematic characteristics of the respective physical places which were extracted from Wikipedia. Their results depict the significance of the chosen scale of place as "*...at the neighbourhood scale the particularities of place emerge, whereas zooming out to the city scale reveals more of the medium such as Twitter and Wikipedia instead of a particular location.*".

Towards a similar goal, S.Gao *et al.*, studied how Flickr, Instagram, Twitter, TrabelBlogs.org and Wikipedia could be used to extract vague cognitive regions. Their inspiration came from a human participants study [41] in cognitive and behavioural geography. The goal of this work is to reproduce this study by combining those high dimensional sources. The data collection part was targeted on data coming from the above-mentioned data sources and geolocated in regions of California, as the original study. In addition, it was based on two sets of keywords meaning that all the collected posts included a subset of those keywords. The similarity of this work's results to the original study's results validated to a certain extent the used approach. One of the most interesting aspects of this work is the comparison between the traditional human participants study and the data-synthesis approach in which the authors discuss both approaches' pros, cons and limitations.

The combination of multidimensional data is once again suggested in [60], for the better classification of place types. Their goal is to improve the pre-trained scene-detection model [62] by adding spatial contextual information which could be broken in three parts: the *spatial relatedness* which quantifies how much the different place types relate to one another, the spatial co-location which explores the distribution of the places over space and the spatial sequence pattern which is a combination of both. The data sources they use is Yelp, Google and Google-Street-View. Their results, support that the use of the spatial context improved significantly both Mean Reciprocal Rank

[1] New York City, Los Angeles, Singapore and London

and accuracy of the initial model.

In most of the previously mentioned studies the combination of the data sources is realized quite indirectly meaning that the "matching" of the different sources is based solely on the geolocation. The "places" are basically "regions", and the data collected from those regions are grouped and analyzed. The last two studies discussed in this section, focus on how to identify POIs coming from different sources and representing the same physical place a process which, as stated before, is called matching. In [36], McKenzie *et al.* suggest a weighted multi-attribute matching based on multiple attributes of Yelp and Foursquare's POIs. The similarity score between two POIs is based on the places' name, the geographic location and the places' "topic" which is extracted using the LDA method on the POIs' reviews. For their experiments, they used 140 POIs which existed in both Yelp and Foursquare and included the needed information for their model. This is an important simplification, since it is highly doubtful that both sources would in reality include the same POIs. Then, for each Yelp POI they calculated a similarity score with every Foursquare POI and they matched the ones with the highest score. Their method resulted in an accuracy of 97%.

Later, S.jiang *et al.* combined Yahoo, Dun and Bradstreet(D&B)[1] and infoUSA[2] to identify land use estimations at the disaggregated level [23]. In this work, the POIs matching is realized for Yahoo and D&B, and the attributes on which it is based are distance, name similarity and website similarity. Their matching algorithm used a tree-structure approach which is presented in B.2 and has 98% accuracy (manual evaluation). The attributes and similarity metrics used in [36] and the tree-structure logic used in [23] are parts of the matching algorithm used in this thesis as well.

2.4 Conclusions

This chapter includes various works towards the study and extraction of places' attributes. The results of those studies reveal the large amount of latent information which could be extracted from the new forms of place-based data sources. Most of those studies examine how POIs and geospatial data could be used to understand activities and patterns of the physical world. Each of the presented studies, is mostly focused on a specific type of features (e.g. focusing on temporal or locational features,) and to what extent place-related information could be collected and used to provide valuable insights about places and human behaviour in respect with this kind of features. Even if the significance of each different type of features is evaluated in those works, a direct comparison of the importance of the different by nature features (e.g. temporal and functional characteristics of a place) is hard to be realized as each study uses different data, collected from different regions and different time periods.

In this thesis, an effort has been made to combine various features which focus on different aspects of places (e.g. functional characteristics and visual characteristics) and reflect the high dimensional nature of places. Then, a direct comparison of the contribution of those features to the classification of urban place types is being made to discover which combination of characteristics make specific place types distinguish

[1] http://www.dnb.com/ (Last

[2] http://www.infousa.com/

from the rest or in other words: what "makes" a place. In other words, instead of focusing on a specific type of features and analyzing them in multiple ways, the focus is given on extracting several dimensions of places and comparing them by computing their importance in the classification of urban place types towards understanding what are the attributes that assist distinguishing the various urban place types.

Chapter 3

Experiment Design

In this chapter, the design of the data pipeline which is used for the extraction of multidimensional POI features from place-related web data is thoroughly explained. An overview of this pipeline could be seen in figure 3.1 including the classification step which is discussed in the next chapter. The goal of the next sections is to tackle the second and third research question of this thesis:

RQ2: *How to combine POI features from various web data sources?*
RQ3: How to extract multidimensional POI features from combined data sets?

In the scope of this thesis the multidimensional POI features are defined as following:

- **Functional Features**: features relevant to the function of a place (e.g. place type, the opening times, the website etc).
- **Experiential Features**: features concerning people's experience when visiting a place (e.g. reviews and ratings)
- **Socio-topical Features**: Socio-topical aspects of a place such as the (social media) sentiment of the people around and the topics discussed by the people around.
- **Visual Features**: features relevant to the visual appearance of a place.
- **Locational Features**: features concerning what is the main type of the nearby places.

The rest of this chapter is divided in two main sections: the **Data Collection & Matching** which includes the selected data sources, the reasoning behind this selection, the methods used to collect the data and the matching algorithm that is used for the combination of the different sources and the **Extraction of POI Features** which includes the techniques to extract the above-mentioned attributes from the collected data.

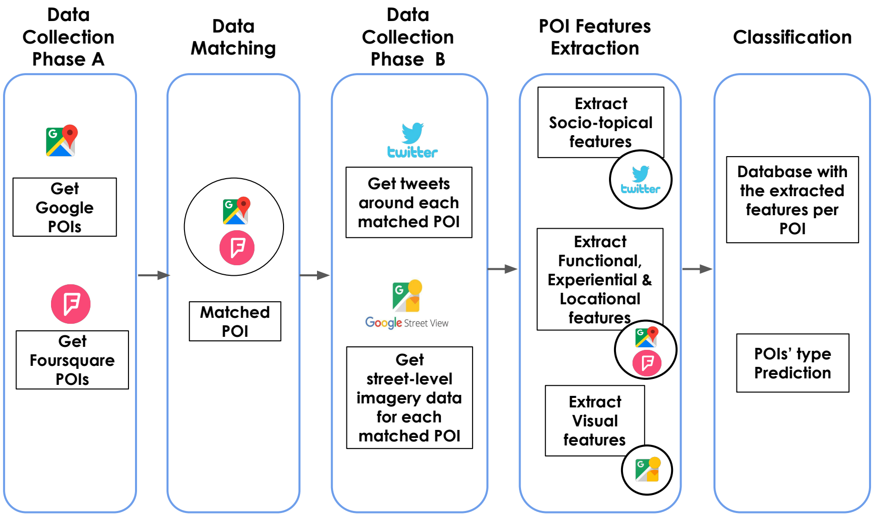


Figure 3.1: Data pipeline for the extraction of POI features and the prediction of the POI types. The pipeline could be divided into four steps: the Data Collection (phase A and B), the Data Matching, the POI features extraction and the Classification step. Firstly, the Google and Foursquare POIs are collected. Then, they are matched and the geo-coordinates of each matched POI are used to retrieve the tweets that have been sent from around this location and the street-level images from this exact location. After that, the POI features are extracted and, lastly, the classification is realized based on the extracted features.

3.1 Data Collection & Matching

3.1.1 Selection of place-based data sources

The large amount of place-related web data sources demands the understanding of each source's advantages and disadvantages. The four basic pillars that the data sources selection process is built upon are:

1. The combination of the selected data sources should be able to express directly or indirectly the selected multidimensional POI features.
2. The selected data sources should be "matchable" meaning that they should share enough features so that the identification of the POIs which come from different data sources and represent the same physical space, is possible.
3. The selected data sources should be diverse to avoid unneeded overlapping information and to reduce the possible bias introduced by a combination of platforms of similar nature [1]. For instance, there could be a "user generated content" bias if all the platforms contain information generated by users.

The starting point of the selection process consists of the exploration of the most known and widely used in the literature web data sources. Those sources, as well as

[1] As the bias introduced by a singe platform which is discussed in [14]

the various ways each one them has been previously used for similar purposes, have been discussed in the second chapter, the *Related Work*, and an overview of them is presented in table 3.1. Each row of the table corresponds to a single web data source and each column contains the sources' details which are relevant to our work. Particularly, for each source: the **API** column presents the official API which could be used to collect the data, the **Extracted Features** column contains the above-mentioned features which could be extracted (even partially), the **Matching Features** column contains the features which could be used for the matching, the **UGC/Other** column shows if the data is user generated and the **Data Kind** lists the nature of the data (e.g. text or image).

Data Source	API	Extracted Features	Matching Features	UGC/Other	Data Kind
Google	Google Places	Functional, Experiential, Visual, Locational	Name, Geolocation, Address, etc.	Both	Place info, Reviews, Images
Facebook	Facebook Places	Functional, Experiential, Visual, Locational, Social,	Name, Geolocation, Address etc.	Both	Place info, Reviews, Images
Foursquare	Foursquare Places	Functional, Experiential, Visual, Locational	Name, Geolocation, Address etc.	UGC	Place info, Reviews, Images
Twitter	Twitter	Socio-topical	Geolocation	UGC	Text
Google Street View	Street View	Visual	Geolocation	Other	Images
Instagram	Instagram	Social, Visual	Tags, Geolocation	UGC	Text, Images
Yelp	Yelp Fusion	Functional, Experiential, Visual, Locational	Name, Geolocation, Address, etc.	Both	Place info, Reviews, Images
OSM	Open Data	Functional, Locational	Name, Geolocation, Address etc.	UGC	Place info
TripAdvisor	TriAdvisor	Functional, Experiential, Visual, Locational	Name, Geolocation, Address, etc.	Both	Place info, Reviews, Images

Table 3.1: Main potential data sources to be used.

The selected data sources which are used in this work are Google, Foursquare, Google Street View and Twitter. Those four data sources act in accordance with all four above-mentioned pillars of the selection process,. Firstly, (1) Google and Foursquare contain functional, experiential and locational POI features, Google Street View contains visual POI features and Twitter contains Socio-topical POI features. Next, (2) Google and Google Street View could be matched based solely on geolocation as they use the exact same location system. The same applies for Foursquare and Twitter. The matching of all four sources could be accomplished by matching the overlapping information between Google and Foursquare POIs (e.g. name, geolocation, address). Furthermore, (3) all those sources are significantly diverse . Google Street View and Twitter differ from every other source of this set, as the former contains outdoor images and the latter user generated posts, information which is not even included in the rest of the platforms. Google and Foursquare could be considered more similar in the way they are used in this work, as they both provide basic information about POIs. However, not only their content has been produced differently, Foursquare is a user generated platform while Google is not, but their conflation proved also to be necessary as in many cases Google Places API does not always return all the POIs' information which could be found by using Google on the web. For instance, a "cafe" POI might not include "cafe" as its type but "point of interest" or "establishment". Lastly, all the selected sources have been used in a lot of previous studies and they are considered to be among the most popular and used platforms [57].

3.1.2 Google-Foursquare POI data collection

The POIs are being collected through Google and Foursquare. In table 3.2 the main, meaningful for this research, data fields of the POI data are presented for each source. As could be seen in table 3.2, the desired information could be sometimes missing (e.g the phone and the price from the Foursquare data).

Google POIs' data

Name	Types	Addr.	Phone	Website	Reviews	Photos Ref.	Rating	Opening Times
Lidl	Supermarket, Store	Zuideinde 278-282	+31 207095039	www.lidl.nl	5 reviews included	10 photos ref.	4	included

Foursquare POIs' data

Name	Types	Addr.	Phone	Website	Tips/Likes Count	Photos Count	Rating	Price
Lidl	Supermarket	Zuideinde 278-282	-	www.lidl.nl	7/0	3	6	-

Table 3.2: Example of the main information included in the POI data provided by Google and Foursquare respectively. Apart from the self-explanatory columns, *Photos Ref.* column refers to the reference ids of the Google POIs' photos.

The main drawback, which should be taken into account in the experiment design, is the strict data access limitation of both sources' APIs. The time needed for the collection of the POIs directly depends on the limitation of the amount of requests per day and per month by each API. Unfortunately, the free versions of both APIs are quite limited and there is no guarantee that their policy will not change at any point of time. Particularly, Google Places currently, since the 16th of July [1], allows 5.000 requests per month and Foursquare allows 99,500 regular API calls and 500 premium API calls per day (the requests which return the POIs details mentioned in table 3.2 are considered to be premium API calls). Thus, it is important to ensure that there is enough time to collect the data for the regions to be studied.

3.1.3 Google-Foursquare POI data matching

Matching is the process of identifying the POIs which belong to different data sources and represent the same physical space. In this work, the matching is realized between the Google and Foursquare POIs. As stated before, street-level images are collected for each Google POI and tweets are collected for each Foursquare POI thus, by matching Google and Foursquare POIs all four sources are, in a sense, matched. In that way, the collected digital representations for each place include the Google and Foursquare POIs' information , street-level images from outside of it and tweets which have been sent within 50m of it.

The matching process of Google and Foursquare POIs is not straightforward and several studies have been focused on designing POI matching algorithms [23], [36], [28]. One could wrongly assume that since POIs include geographic coordinates a matching algorithm which is just based on the distance between the POIs of the two sources would suffice. However, as mentioned in [36], ... *it is not uncommon to find a significant discrepancy in the geographic coordinates of the same location sourced from two applications.* Consequently, other POI features are also used for the matching. Naturally, the other features of the POIs from the two sources also exhibit discrepancies [28]. For instance, the Google POI name of a restaurant in Amsterdam is "Betty's Restaurant" while the Foursquare POI name of the same restaurant is "Betty's". Thus, those features should be compared based on string similarity metrics and matched if the overall similarity score is above a specified threshold. This study does not focus on the design of a novel matching algorithm. The used matching algorithm is created for practical purposes, meaning the conflation of the various data sources which are selected to be used for the extraction of POI features.

The inspiration for this works' matching algorithm is based on the tree structure approach of [23] which is presented in figure B.2 and on the following string similarity metrics, some of which are also used in [36]:

- **Levenshtein distance**: The Levenshtein distance, also called as "edit distance", between two strings is basically the minimum required number insertions, deletions or substitutions of a single character (edits) to convert the one string into the other [27].

- **Damerau-Levenshtein distance**: The Damerau-Levensthtein distance differs from the Levenshtein distance by adding on the allowed edits the transposition of two adjacent characters [12].

[1] Before the 16th of July the limitations were less strict for Google.

- **Phonetic similarity**: Using the double metaphone algorithm [44] the strings are converted to phonetic codes. The phonetic codes are then compared using the Levenshtein distance.
- **Ratcliff and Obershelp's algorithm**: This algorithm gets the longest common subsequence of the two stings and then it works recursively for the remain characters from the left and right of the this subsequence. The similarity score is then computed as the number of matching subsequences divided by the total number of characters [47].
- **Longest subsequence metric**: Measures how similar is one string A to another string B as the length longest common subsequence of the two strings divided by the total number of characters in A.

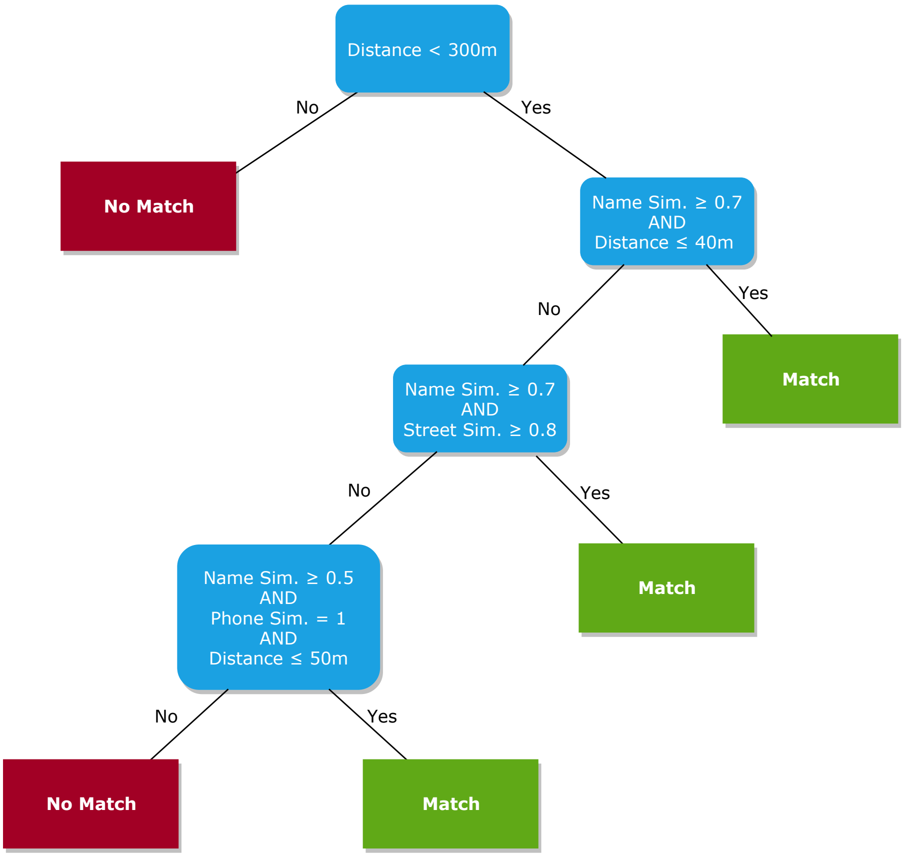


Figure 3.2: Structure of the used matching algorithm. "Sim." stands for similarity and each rounded node of this tree refers to the comparison between Google and Foursquare POIs.

The overall string similarity score for each feature is calculated by using the mean of all the above string similarity scores. The POI features which are compared by using the above-mentioned string similarity metrics are: name, address and phone.

Obviously, some of these fields might be missing for some POIs. In this work, an effort has been made to exclude as less POIs as possible from the matching process. Thus, in the worst case scenario the selected POIs for the matching process include only the name and the geo-coordinates and as a result there are cases in which Foursquare and Google POIs are compared based only on those two features.

The tree-structure of the matching algorithm is presented in figure 3.2. For defining the rules and thresholds of this algorithm a heuristic approach is followed. Basically, for each Foursquare POI all the Google POIs within a distance of 300m are retrieved. Then, the Foursquare POI is compared with each one of the Google POIs and the POIs are matched if one of the following rules is true: their name similarity Particularly, the thresholds are tuned according to the trade off between precision and recall. The precision is important so that the quality of the matches is high and the recall is tuned in a way so that the size of the matched data is sufficient for the rest of this research. The algorithm's evaluation is presented in the next chapter, the *Implementation*.

3.1.4 Google Street View data collection

Google Street View contains street-level geolocated 360 panoramic images which are not linked in any way with the Google POIs. Their collection is based on the Street View API. Specifically, the geographic coordinates (latitude and longtitude) of each previously collected Google POI are used as parameters so that the closest panorama to the POI's location, within a radius of 50m, is retrieved. Each panorama is being stored in the form of four images and each image's direction is defined according to where the compass of the camera is headed: north, east, south, and west respectively. If there are Google Street View images of the same location taken on different dates, only the most recent ones are kept.

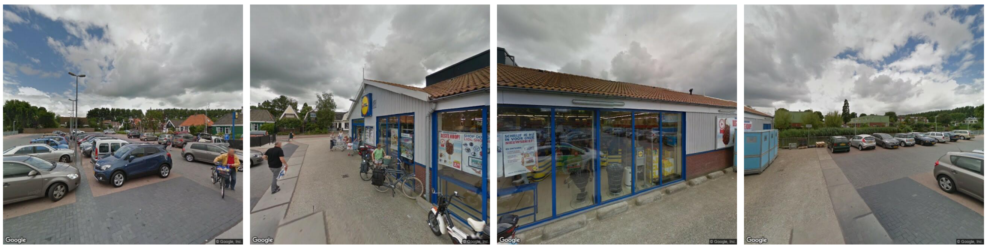


Figure 3.3: Example of collected street-level images outside a "Lidl" store

Therefore, for each collected Google POI, four street-level images which depict the outside of the place are retrieved. The fact that Google and Google Street View use the exact same geolocation system ensures that the images will indeed be taken from as close to the physical place as possible, assuming of course that the Google POI's geographic coordinates are correct.

3.1.5 Twitter data collection

Twitter also offers an API which offers the functionality to search for tweets by geographic coordinates, keywords or places using a Twitter place ID. However, searching by place returns results for a very limited amount of places.

Thus, in this work the tweets are collected based on their location. Particularly, for each Foursquare POI, its geographic coordinates are used to search for tweets which are geolocated within a radius of 50m and have been posted between 01/01/2017 and 20/10/2018. The 50m radius has also been used in [10] vice versa: each tweet is enriched with geospatial features through the tags of Google and OSM places which are located within 50m of the tweet's coordinates. For each tweet, the text, language, creation date, amount of times retweeted and favored and geolocation are collected.

The fact that Foursquare and Twitter use the exact same geolocation system ensures that the tweets will have indeed been posted within 50m radius of the POI, assuming of course that the Foursquare POI's geographic coordinates are correct.

3.2 Extraction of POI features

In this section, the extracted POI features from the collected and matched data and the methods and techniques used regarding this goal are presented. Those features are selected in the belief that combined they provide a more complete digital representation of places and their use could help in the exploration of the question "what makes a place?". Is a place considered to be a restaurant because people experience it as a restaurant, because it looks like one, because it is self-labeled like one or because all of those three reasons? The rest of this section is divided in three parts based on the nature of the extracted features which could be functional, social, experiential, visual or locational.

3.2.1 Extraction of functional features

The places' functional features are extracted from **Google** and **Foursquare** POIs' data. Some features could be used as included in the POIs while others need some processing first. The extracted functional features are the following:

Type: The main function of a place according to the POI data. This is a core feature for this study as, regarding the fourth research question, machine learning models are trained on the various extracted features for the prediction of this feature, the POI's type. Type is extracted from Foursquare POIs since the type fields are more consistent and complete than the Google POIs' type fields. The selection process of the used POI types in this work is presented in the *Implementation* chapter together with the implementation details of the extraction procedure.

Opening Times: Opening times are directly extracted from Google POIs without any further processing. This attribute has been also suggested in other works [37], [52] as a promising POI type indicator.

Website included: Whether there is a website included in the POI data. This information is collected from both Google and Foursquare POIs. The initial intuition is that this feature will not make a significant difference to the results of this study.

Phone number included: Whether there is a phone number included in the POI data. Again, this feature does not seem like a significant one.

Facebook/Twitter included: Whether there is a Facebook or a Twitter account included in the POI data. This information is extracted from Foursquare POIs.

Most Popular Timeslots: Most popular timeslots are extracted by processing the popular times from Google POIs. In Google POIs, popular times information is included as depicted in fig 3.4.

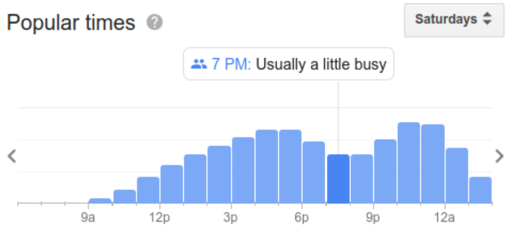


Figure 3.4: Popular times information as included in Google POIs.

Price: Information about how expensive a place is (e.g. cheap, moderate, expensive). This information is included in Foursquare POIs.

All the above mentioned features aim to express the functional attributes of a place. Their extraction obviously depends on the existence of the required POI information, which could be missing for several POIs. Some of them seem to be more valuable for distinguishing one place type from another (e.g. opening times) while others seem less important (e.g. phone number included).

3.2.2 Extraction of Socio-topical features

The Socio-topical features aim to capture some of the social aspects of the places. For their extraction, Twitter is used. The tweets collected for each place are aggregated for the extraction of the following features:

- **Amount of posted tweets**: The amount of the tweets sent around each place per language and in total.

  Average number of words: The average number of words used in the posted tweets per language and in total.

- **Average Tweets' Sentiment**: There are several studies which have focused on quantifying the sentiment of a text or even more specifically of a tweet [43]. Depending also on the language on which the text is written there are multiple open tools and techniques provided for this exact purpose. The used techniques are presented in the next chapter due to the fact that their selection is based on the specific use cases studied in this work. Different use cases could demand the use of other techniques.

- **Time Differences between tweets**: The time difference between consecutive tweets per place shows how distributed are the tweets in time, for those two years for which they have been collected.

- **Topics**: The topics discussed in the tweets (grouped by place) which are derived using one of the most used in the literature topic model, the Latent Dirichlet Allocation (LDA) [33], [16], [38]. LDA is a "bag of words" model which instead of using the words frequency it uses the distribution of words across topics. In LDA, each topic is expressed as a list of words with probabilities for them to belong to that topic and each document is expressed as probability distributions over words. In this work, each "document" consists of the collected tweets for each place. Thus, places are "*modeled as a mixture of topics which themselves are multinomial distributions over terms* "[3]. In a more simplified way, each place is described by a unique topic probability pattern derived by the tweets which have been sent from around it. LDA is an unsupervised method meaning that there is no need of labeled data and no predefined topics exist. Two worth mentioning facts for the LDA are that, since the topics are presented as a list of words (with their respective topic probabilities), the interpretation of what each topic actually represents is subjective and that the number of topics to be extracted is a parameter which should be tuned carefully as its value significantly changes the quality of the topics.

3.2.3 Extraction of experiential features

The **experiential** features aim to describe how people experience a place. The data sources from which those attributes are extracted are Google and Foursquare. The extracted experiential features are the following:

- **Ratings**: The ratings are extracted from both Google and Foursquare POIs as they are included in both platforms, and no further processing is needed. The ratings, in a sense, quantify the quality of people's experience in a place.

- **Likes**: The likes are extracted from the Foursquare POIs without the need of any further processing and they aim to show the places' (online) popularity.

- **Photos count**: The number of photos which exist in Foursquare. Again, this feature tries to show how people experience a place. A more visited or visually appealing place could include more photos taken from Foursquare users than a less popular or not so "visually interesting" place.

- **Amount of posted reviews**: The amount of Google reviews written for each place are counted and stored per language and in total. This attribute could include latent features of the places' visitors. Depending on the reviews' language a place could be more local (native language mostly used) or more international. The total amount of reviews could include information about the age or mentality of the visitors. For instance, one would expect less online reviews for a cafe which gathers elderly people than a cafe that youngsters prefer. The main drawback of this attribute is that Google returns a maximum of five reviews per place. Thus, while by using this data there is a distinction between a place with none or 1 review and a place with 4 or 5 reviews, there is not distinction between a place with 5 reviews and a place with 10.

- **Reviews' average number of words**: The average number of words used in the reviews. This feature might not be that significant but it could, for example, provide information about how connected the visitors of a place are with a place. This statement is based on the, not so strong assumption, that a person would write larger reviews for a place that she goes often and feels connected with than for a place she visited once and she liked/disliked.

- **Average Reviews' Sentiment**: As previously described for the tweets, the average sentiment of the reviews is also quantified and the selection of the techniques and tools used for the sentiment analysis are described in the next chapter.

- **Google Reviews Topics**: Similarly to the tweets, the reviews are aggregated per place and topics are extracted from them using the LDA method. The information that this attribute includes is quite interesting to be explored, as the reviews basically describe how people experience a place. Thus, a place could be labeled as "bar" but in the reviews people might reference it as a "cafe". This contradiction leads once again in the question "what makes a place?". In this context, is this place defined by its label or by how people experience it?

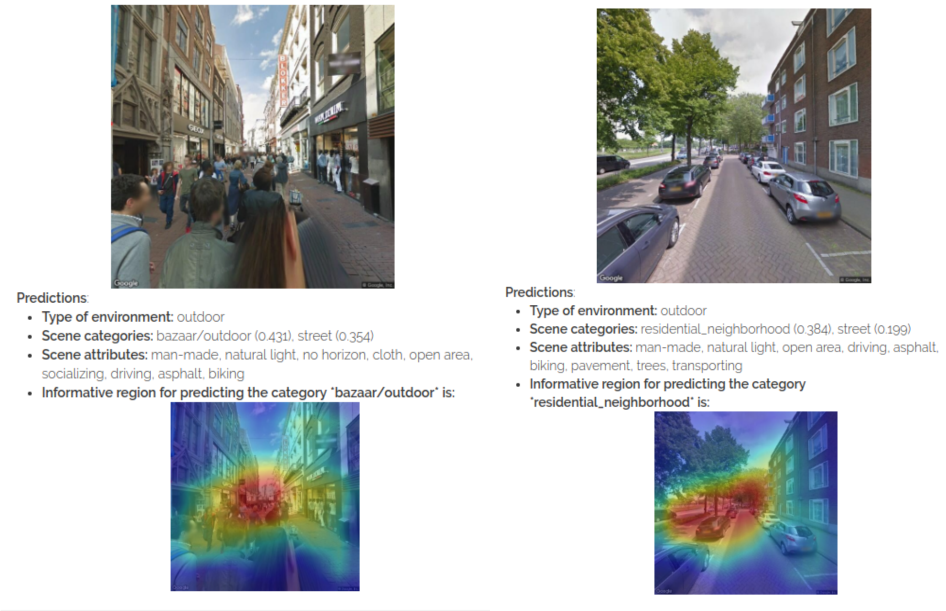
3.2.4 Extraction of visual Features

The visual features are extracted from the Google Street View images and this extraction is realized using image recognition techniques. In this work, the aim of the extracted visual features is twofold: first to describe how people perceive places appearance-wise, for instance does a place look like a junkyard, a bar or a jewelry shop, and secondly to discover objects such as buildings and trees outside of the places, which could indicate some of the places' characteristics. For example, if there are a lot of trees outside a place, this place could be considered less "urban". In addition, a large amount of trees could indicate that there is a higher probability for this place to be a university campus than a nightclub. Thus, the visual features extraction process could be further broken into two parts: the scene recognition part and the object recognition part.

Several approaches have been followed in the past in the image recognition field but during the past few years deep learning models have been quite dominated it. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [49] is being widely used for the evaluation of object detection and image classification algorithms and as could be seen by the results, since 2012 the winners of this challenge have all developed deep neural networks architectures [56]. Thus, deep learning approaches are also followed in this work for both parts of the visual features extraction.

Admittedly, the goal of this thesis is not to develop a novel image recognition model neither to improve the existing algorithms. Thus, pre-trained, state-of-the-art deep learning models which have been openly shared by their contributors are used as tools towards the extraction of the places' visual attributes.

There are various datasets which could be used for the **scene recognition** part, such as the SUN database [58], the Scene15 database [26], the MIT Indoor67 database [46] or the Places Database [62], [7]. In this work, Places Database is used because of its size, which is extremely important when a deep learning model is used, and the number of supported classes. To be more specific, Places Database, includes 1,803,460 images labeled with 365 scene categories. Each category is represented by a number of images which varies from 3,068 to 5,000. The creation process of this database involves four main steps: *...querying and downloading images, labeling images with ground truth category, to scaling up the dataset using a classifier, and further improving the separation of similar classes.* Each step is thoroughly explained in [62]. It is worth mentioning that the images ground truth annotation task was crowdsourced to Amazon Mechanical Turk [8] and, therefore, the labels of those images are based on people's perception. Several pre-trained, on the above-mentioned dataset, Convolutional Neural Networks (CNN) are openly shared. In the context of this thesis, these models are used towards the extraction of semantic scene categories from each one of the collected street-level images. An example, of such a model's outcome could be seen in figure 3.5. The implementation details concerning the used models are discussed in the next chapter.



Predictions:
- **Type of environment:** outdoor
- **Scene categories:** bazaar/outdoor (0.431), street (0.354)
- **Scene attributes:** man-made, natural light, no horizon, cloth, open area, socializing, driving, asphalt, biking
- **Informative region for predicting the category 'bazaar/outdoor' is:**

Predictions:
- **Type of environment:** outdoor
- **Scene categories:** residential_neighborhood (0.384), street (0.199)
- **Scene attributes:** man-made, natural light, open area, driving, asphalt, biking, pavement, trees, transporting
- **Informative region for predicting the category 'residential_neighborhood' is:**

Figure 3.5: Scenery detection example. As could be seen the "scene categories" successfully describes what the scenery looks like. The class activation maps of the two pictures are also presented.

For the **object detection** part, pre-trained, deep learning state-of-the-art models trained on the Google's Open Images dataset [1](around 9 million images with 600 box classes) and on the COCO dataset [2] (2.5 million labeled instances in 328k images) [29] are used. Both datasets are significantly large (the Open Images dataset might be the largest dataset with object annotations) and widely used. From all the possible detected objects using those two datasets, we focus on the "outdoor objects" presented in table 3.3.

[1] https://storage.googleapis.com/openimages/web/index.html
[2] http://cocodataset.org

Open Images Dataset	COCO Dataset
tree, houseplant, flower, building, skyscraper, house, convenience store, office, streetlight, traffic light, traffic sign	fire hydrant, stop sign, bench, potted plant

Table 3.3: Main potential data sources to be used.

3.2.5 Extraction of locational features

The locational features depend purely on the location of a place relatively to the location of other places. For the extraction of the locational features Foursquare POIs are used. For each place the amount of the (studied) POI types which exist in a distance of 100m, 1km and 5km is retrieved (e.g. within 100m of a place 4 bars and 2 restaurants exist).

3.3 Chapter Summary

In this chapter, the data sources selection process, the collected data, the matching among the sources and the methods for extracting functional, semantic and visual attributes of urban places are presented. Table 3.4 shows an overview of the extracted features. In the next chapter, the implementation details of the above-mentioned process are discussed.

Extracted F.	Data	Sources	Extraction Methods
Functional	POI Data	Google, Foursquare	Straight from POIs
Socio-topical	Tweets	Twitter	Text Processing & LDA
Experiential	Google Reviews, POI Data	Google, Foursquare	Text processing, LDA & Straight from POIs
Visual	GSV images	GSV	Scene recognition and object detection deep learning algorithms
Locational	POI Data	Foursquare	Postgis distance query

Table 3.4: Overview of the extracted attributes

Chapter 4

Implementation

In this chapter, the techniques, models and software used for implementing the data pipeline shown in 3.1 and described in the Experiment Design chapter and the results of the POI types classification are presented. The core of the implemented system has been built on Python [1]. The selection of Python is based on its multiple key-packages which are used in the different sub-parts of this system and presented in the next sections (e.g. for the topic modeling or the object detection). The used database for storing and retrieving the data is PostGis, *"a spatial database extender for PostgreSQL object-relational database which adds support for geographic objects allowing location queries to be run in SQL"* [2]. PostGIS proved to be ideal in the context of this thesis as the functionality of querying in a location-wise manner was needed in several parts of the implementation which are analyzed in the next sections. The main goal of this chapter is to tackle the fourth research question of this thesis:

RQ4: *Which POI features contribute the most to the classification of urban place types?*

In the next sections the implementation steps which lead to the study of the POI features contribution to the classification of urban place types are described. The main steps are the *Data Collection & Matching*, the *Extraction of POI features* and the *POI Type Classification*.

4.1 Data Collection & Matching

4.1.1 Use cases selection: Amsterdam & Athens

The study is focused on two capital cities: Amsterdam (Netherlands) and Athens (Greece). The selection of those cities is mostly based on the fact that the quality of the derived conclusions in a research is directly connected with the knowledge of the researcher about the studied object. Therefore, when studying urban environments, one could gain more insights if the studied urban environment is familiar to her. Moreover, in both of those two cities all the four selected place based platforms are used and

[1] https://www.python.org/

[2] https://postgis.net/

consequently the needed data for the features extraction exist. Furthermore, Amsterdam and Athens are diverse in various terms such as their economy, visual appearance and language, thus, the comparison of the two of them is valuable in understanding to what extent the results of this study could be taken into consideration for other regions as well.

In figures 4.1 the studied regions coming from Amsterdam and Athens are presented, respectively. For both cities a same size rectangle area which is located in the centre of each city is selected to be studied. The selected city-centre of Amsterdam is based on the region depicted in figure 4.2 (left), and the selected city centre of Athens is based on the inner ring of Athens depicted in figure 4.2 (right). Firstly, POIs were collected from larger regions of Amsterdam and Athens and then the areas to be studied were adjusted so that they contain the largest part of each center and at the same time include as many of the retrieved POIs as possible. The presented rectangles in figure 4.1 are the final selected regions which are analysed in the next sections.

Focusing on the centre of both cities aims to reduce to a certain extent other bias which could be introduced due to the different structure or components of the city. For example, if a city which includes multiple beaches is compared with a city which includes only industrial sites, their comparison could include differences which are not based on place-centric differences between the two cities such as how people experience places but other kind of differences which are not relevant to this research. In the scope of this thesis, the goal is to compare the two cities while keeping, as much as possible, other parameters the same. Towards this goal, the selected regions are the same size and they both include the "heart" of each city.

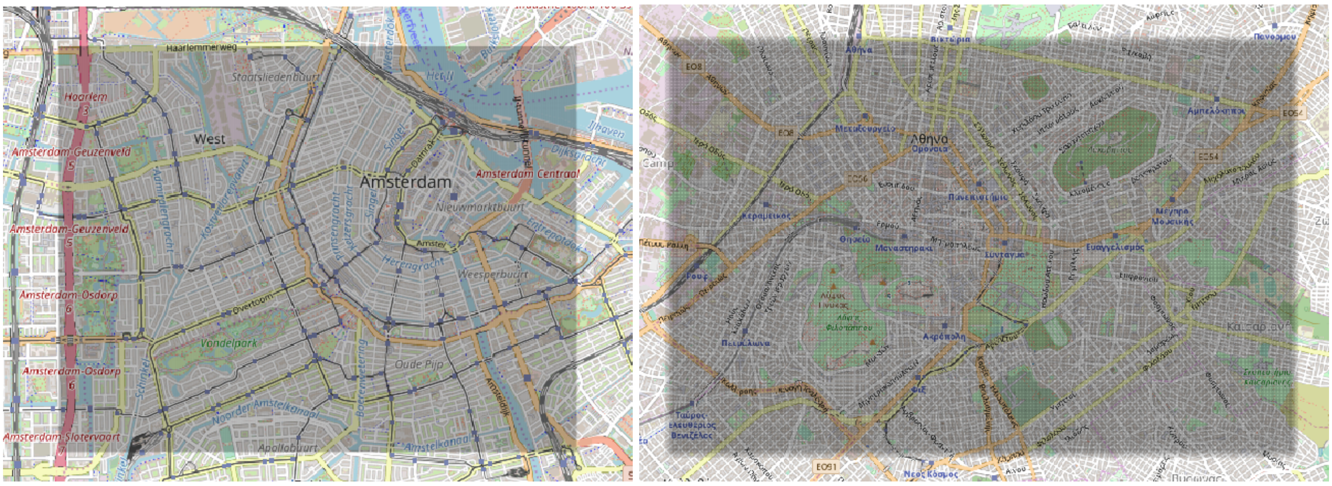


Figure 4.1: Amsterdam (left) and Athens (right) regions used in this study.

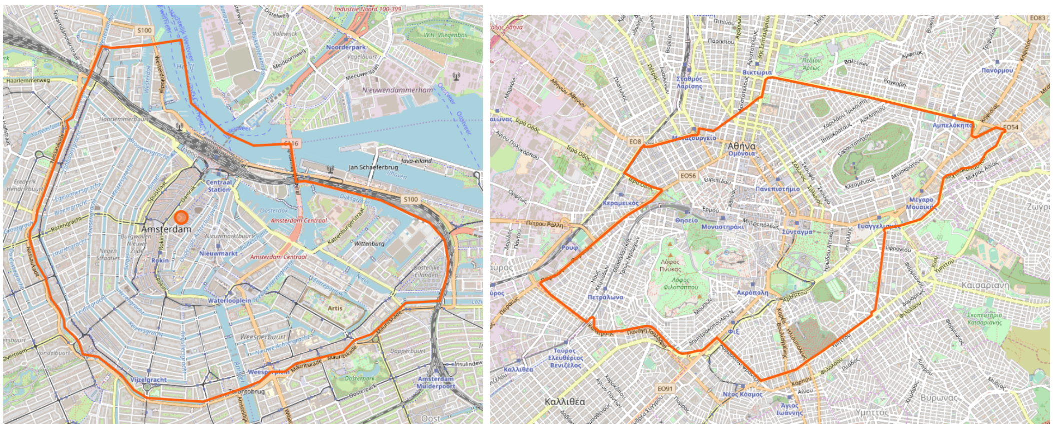


Figure 4.2: Amsterdam Centre (left) and the inner ring of Athens (right) as defined by OpenStreetMap.

4.1.2 POI data collection

As previously stated, the POI data are being collected through Google and Foursquare. The access to the Google POIs is realized using the Google Places API. The Google Places API provides the functionality of requesting Google POIs which are geolocated in a specified location through what is called a "nearby search request" [1]. The access to the Foursquare POIs is realized using the Foursquare Places API which also provides the functionality of searching for POIs near specified geographic coordinates, through the "search for venue" request [2].

A core difference of those two types of requests, is that by using the Google's "nearby search request" it is not possible to retrieve only POIs of specified types (e.g. bars and restaurants) [3] something which is possible when using Foursquare's "search for venue" request. For that reason, the queries for retrieving Google POIs return any possible POI type. On the contrary, due to the below-mentioned reasons, considering Foursquare only POIs which belong to one of the following categories are retrieved [4]: **food, arts and entertainment, college and university, nightlife spot, outdoor and recreation, bank, clothing store, drugstore, hotel**. The excluded categories are:

- **Event** (e.g. Christmas market and festival): excluded because of the places' seasonal nature.

- **Professional and other places** (e.g factory, animal shelter) excluded due to the vagueness of this category's definition.

- **Residence** (e.g. home): excluded for privacy reasons

[1] https://developers.google.com/places/web-service/search

[2] https://developer.foursquare.com/docs/api/venues/search

[3] It is actually possible only if searching for a single type of POIs. The discussed problem refers to the situation where more than one type is needed.

[4] The rest of the categories could be found here https://developer.foursquare.com/docs/resources/categories

- **Travel and transport** (e.g. airport, bus station): excluded either because the number of such places within a city would be very small (e.g. for airport) or because they are not geographically static (e.g. cable car) or because they are not places for which people share their experiences (e.g. bus station). "Hotel", is the only sub-category which is kept from the main Travel and transport category.

- **Shop and service** (e.g. ATM, Betting shop): Several sub-categories of the "Shop and Service" category were excluded either because they are too specialized (e.g. betting shop) or because they are not places for which people share their experiences (e.g. ATM). "Bank", "clothing store", "drugstore" and "food and drink shop", are the only sub-categories which are kept from the main shop and service category.

Furthermore, a main drawback of both APIs (apart from the data access limitation discussed in *Experiment Design*), is that when searching for POIs within a customized radius from a specified location, Google returns up to 60 and Foursquare up to 50 POIs. Thus, to collect the POIs of a city, multiple calls must be made using a relatively small radius over a large amount of locations.

The selection of the locations to be used for the POIs retrieval is based on the assumption that the amount of existing POIs within a radius of 30m from any given location is less than 50. For selecting those locations QGIS [1] is used. Particularly, a 40mx40m grid (set of tiles) is first created on top of the selected regions. Then, the geo-coordinates of the centroids of each created tile are calculated (figure 4.3). Lastly, those geo-coordinates are used as parameters in the "search requests" for both Google and Foursquare with the additional radius parameter equal to $20\sqrt{2}$ (the distance between each centroid and the corners of the tile the centroid belongs in). In that way, the whole selected regions is being traversed and the POIs from every part of each city-centre are being collected (example in figure 4.4). In this part, the goal is not to retrieve a complete list of the POIs of each region but to collect enough POIs distributed in the two centres, so that a large as less biased as possible POI dataset is created and studied. Particularly, for Amsterdam **64,906** Google POIs and **15,624** Foursquare POIs are retrieved and for Athens **44,633** Google POIs and **11,294** Foursquare POIs.

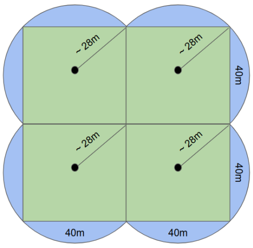


Figure 4.4: Example of the searching area of each request. For simplicity four tiles are presented. Using four different geo-coordinates and four requests with a radius equal to $20\sqrt{2}$m all the POIs existing in the green and blue area are returned.

[1] https://www.qgis.org/en/site/

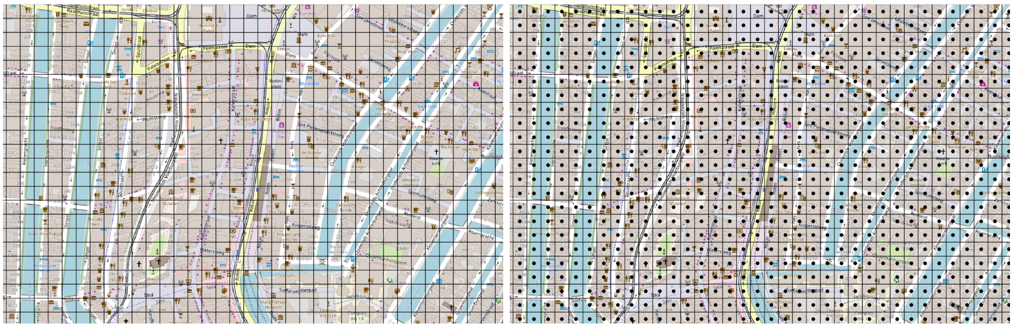


Figure 4.3: Example of the locations used as parameters for retrieving the POI data. On the left the created grid is shown and on the right the centroids of each of the polygons created by the grid are presented. These locations correspond to the queried locations.

4.1.3 POI data matching

The algorithm used for the matching of Google and Fousquare POIs has been explained in the Experiment Design chapter. The structure of the algorithm is, once again, presented in figure 4.5. It is essentially based on four elements: (1) the geographical distance among the two POIs, the string similarity of their (2) name, (3) street and (4) number.

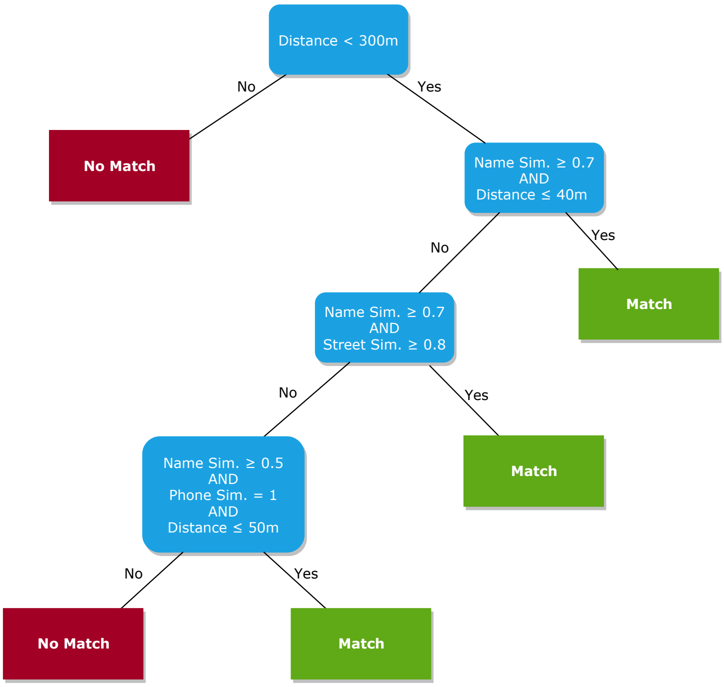


Figure 4.5: Structure of the used matching algorithm. "Sim." stands for similarity and each rounded node of this tree refers to the comparison between Google and Foursquare POIs

The matching algorithm is implemented in Python and the following packages are used for the computation of the string similarity metrics whose function is also thoroughly described in the *Experiment Design* chapter:

- **Levenshtein distance**: python-Levenshtein [1]
- **Damerau-Levenshtein distance**: pyxDamerauLevenshtein [2]
- **Phonetic similarity**: fuzzy [3] (DMetaphone)
- **Ratcliff and Obershelp's algorithm**: difflib [4] (SequenceMatcher)
- **Longest subsequence metric**: difflib (SequenceMatcher)

The rules and thresholds presented in figure 4.5 were chosen based on manual evaluations of the algorithm. For a human, judging if two POIs represent the same physical place is quite objective and straightforward, thus, crowdsourcing this task to multiple people was not considered necessary. The rules and thresholds are quite strict in order to obtain matches with high confidence, as in [23]. The core idea of the tuning process used, is that focus is given on the precision, meaning that the matched data should have indeed be matched, while recall is taken into account just so that the amount of the retrieved matched POIs is sufficient for the further analysis. The results of the matching algorithm could be seen in table 4.1.

City	% of correct Matches	GPOIs Amount	FPOIs Amount	Matched POIs Amount
Amsterdam	0.97	64,906	15,624	4,532
Athens	0.98	44,633	11,294	3,275

Table 4.1: Matching algorithm results. The "% of correct matches" is based on manually evaluating 200 matched POI pairs for each city. The initial amount of Google POIs is unsurprisingly larger than the amount of Foursquare POIs, as from Google all POIs were collected while from Foursquare only specific POI types were collected. In addition, both cities similar percentage of POIs are matched relatively to the amount of Foursquare POIs, around 29% .

4.1.4 POI Data Selection

The amount of different POI types in the "matched POI dataset" is quite large with several POI types existing only once or twice. In order to gain insights from the analysis of the POI features, multiple examples per type should be considered, thus, a selection of POI types to be studied should be made. As stated before, the POI types are extracted by Foursquare POIs as the "type" fields in the Google POIs include in various cases vague information (e.g. "point of interest" or "establishment"). Foursquare

[1] https://pypi.org/project/python-Levenshtein/
[2] https://pypi.org/project/pyxDamerauLevenshtein/
[3] https://pypi.org/project/Fuzzy/
[4] https://docs.python.org/2/library/difflib.html

POIs' type fields seem to be more consistent and the structure of the Foursquare POIs' categories is clear and openly shared [1].

The selection of the POIs to be studied is based on the amount of instances per POI type. Particularly, The ten most frequent POI types are selected to be included in the rest of the analysis. As an example of this process, in figure 4.6 the amount of Amsterdam POI instances grouped by their main type is presented (top 25). The labels are color-coded with each color representing the types which are merged into a single type. This merging is realized because some types are very specialized, for instance "Restaurant" and "Italian Restaurant" could be both labeled as "Restaurant", and is based on the secondary type of the POIs (e.g. "Italian Restaurant" has "Restaurant" as its secondary type). Moreover, some POI types are excluded due to the vagueness of their type (e.g."Pizza place" could be considered either as "Restaurant" or "food and drink shop" and it has "Food" as its secondary type which does not lead in a better understanding).

The resulting categories, after merging the similar types, are the following: **Hotel**, **Bar**, **Coffee Shop**, **Restaurant**, **Cafe**, **Clothing Store**, **Art Gallery**, **Food and Drink Shop** and **Gym**. In addition, a category in which a large amount of POIs belongs but could not be seen in figure 4.6 is the **College and University**. The reason that those POIs are not presented in this figure is that their main type field might vary (e.g. "Academic Building" or "University"). Thus, for the rest of the study those POIs are also included and merged under the "College and University" category. The amount of selected POIs is **3331** and **2501** for Amsterdam and Athens, respectively, and figures 4.7 and 4.8 show the amount of selected POI instances per type.

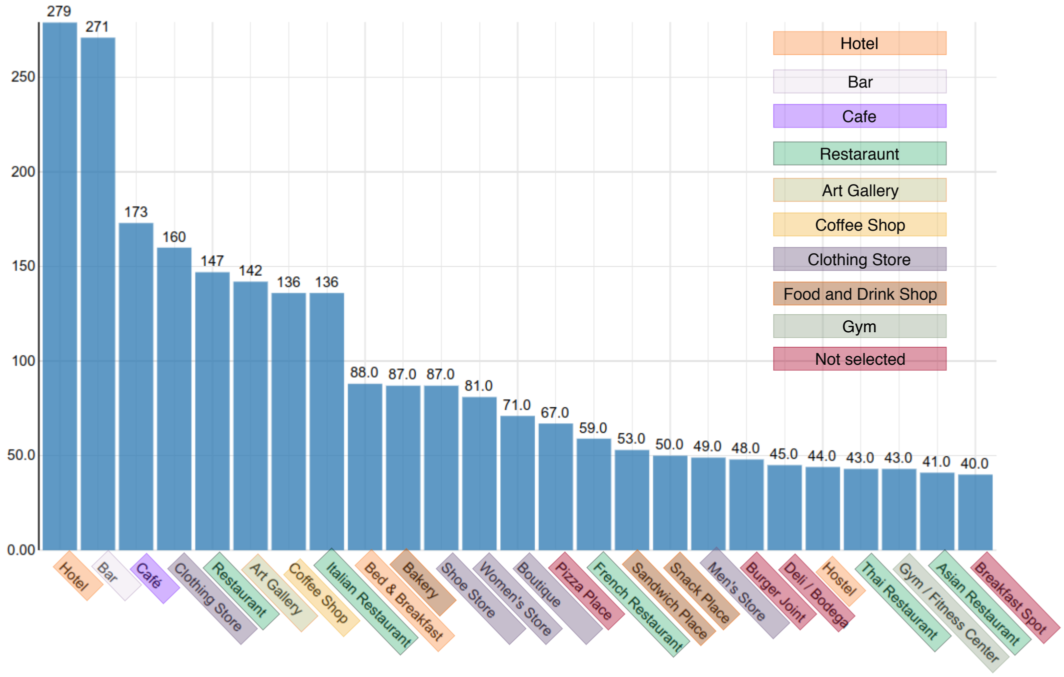


Figure 4.6: Amount of Amsterdam POI instances per type.

[1] https://developer.foursquare.com/docs/resources/categories

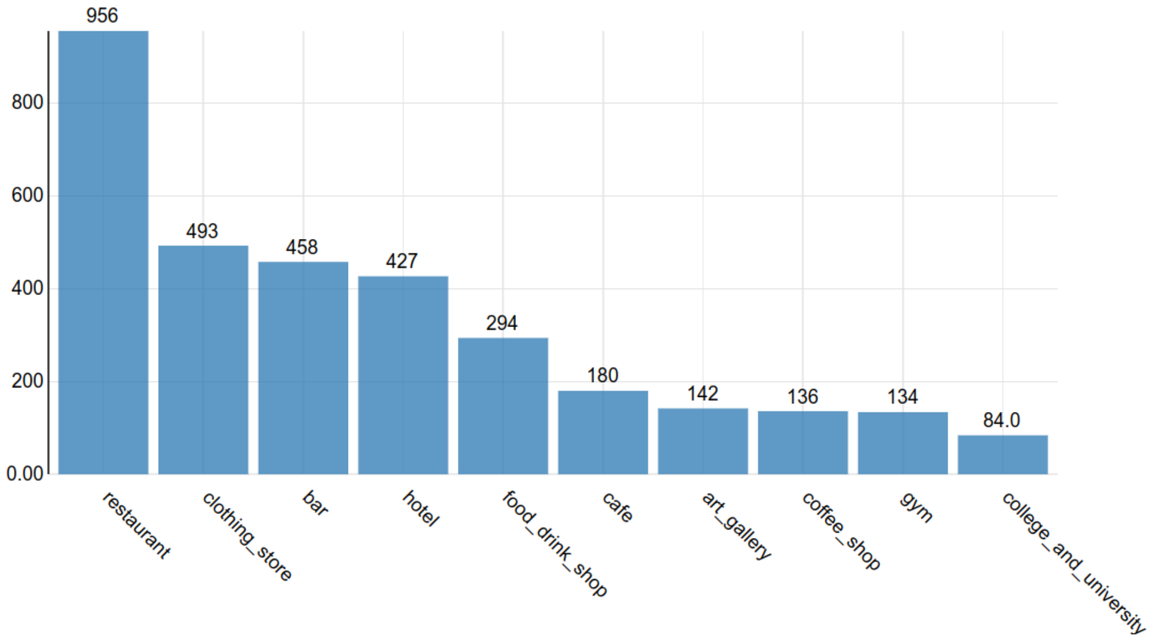


Figure 4.7: Amount of **selected** Amsterdam POI instances per type.

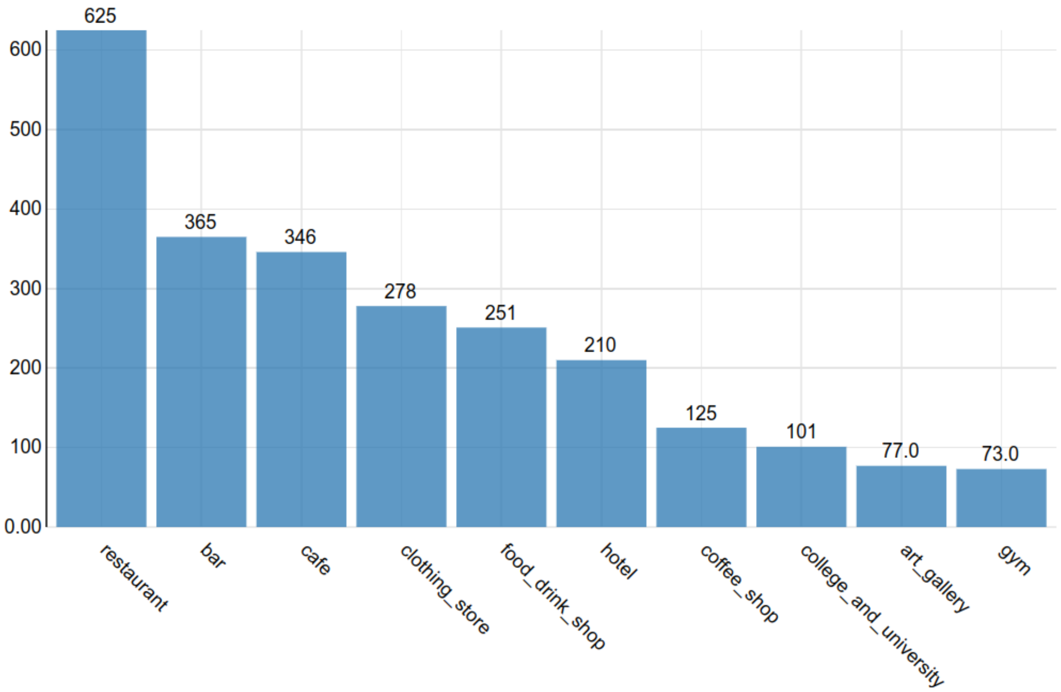
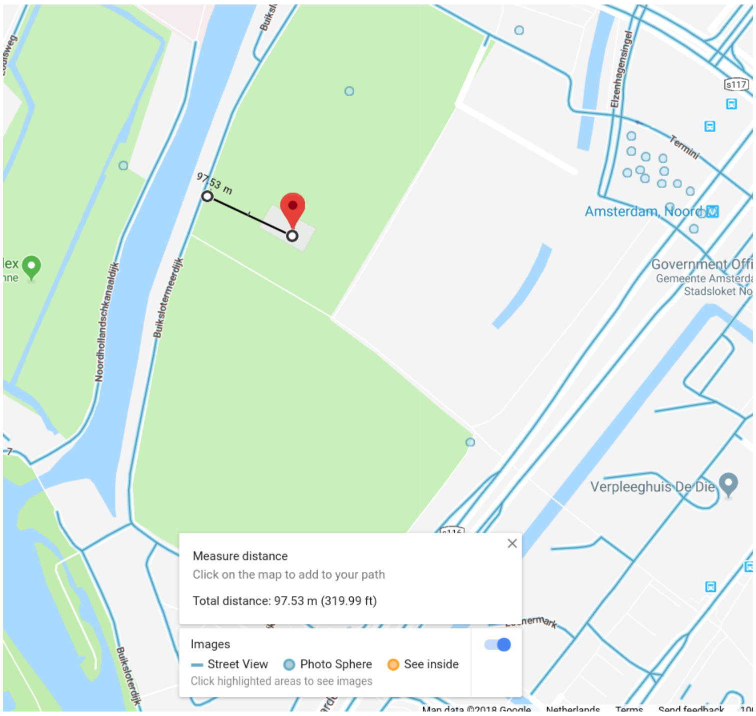


Figure 4.8: Amount of **selected** Athens POI instances per type.

4.1.5 Google Street View Data Collection

As previously stated, the Google Street View data collection is based on the Google Street View API. For each one of the selected POIs, the geo-coordinates taken from the Google POI are used as parameters for the street-view image request. The API, searches in a radius of 50m and returns the closest panorama to the specified location. Unsurprisingly, for some POIs no results are returned as no Google Street View image has been taken in a distance of 50m from the POI's location. An example is presented in figure 4.9. For that reason, for 55 and 41 POIs for Amsterdam and Athens, respectively, no street-level images are retrieved.

As explained in the *Experiment Design* chapter, each panorama is being stored in the form of four images. This is accomplished by adjusting the *heading* parameter, which indicates the compass heading of the camera [1] of the Google Street View API equal to 0, 90, 180 and 270. Thus, the final amount of the collected street-level images for the selected POI data is **13,104** for Amsterdam and **9,840** for Athens.



Measure distance

Click on the map to add to your path

Total distance: 97.53 m (319.99 ft)

Images

Street View ○ Photo Sphere ○ See inside

Click highlighted areas to see images

Figure 4.9: Example of POI for which no Google Street View Image is returned. The blue lines represent the locations for which Google Street View images exist. As could be seen the closest street-level image to the pointed location is in a distance of 97.53m which is more than the limitation of the API (50m). Thus, no image is returned for this POI.

[1] https://developers.google.com/maps/documentation/streetview/intro

4.1.6 Twitter Data Collection

The collection of the tweets is based on the scraping Twitter by searching and then automatically "scrolling" to retrieve old tweets. For each POI, the geo-coordinates taken from the Foursquare API (as Foursquare and Twitter use the exact same geolocation system) are used as parameters to retrieve the tweets which have been sent from each POI's location. Additionally, only tweets which have been posted between 01/01/2017 and 20/10/2018 are kept. The reasoning behind the selection of this timeslot is twofold: one relatively "new" tweets should be collected. This is important because the tweets are indirectly linked with places. If when the tweets were posted, the place in that location was not the same as now, then tweets have been collected around the correct location, but for a different place. Second, by gathering tweets which have been sent throughout the last two years, the seasonality bias meaning the introduced bias when tweets only from specific months/seasons are collected, is reduced.

The amount of tweets collected is, for Amsterdam, **120,250** and for Athens **93,858**. As could be observed, the amount of the tweets collected for Athens is almost equal to the 75% of the amount of tweets collected for Amsterdam. This is reasonable as the amount of POIs collected for Athens is also equal to the 75% of the amount of POIs collected for Amsterdam.

4.1.7 Data Overview

The final amount of the collected, matched and selected data is, for Amsterdam 3,331 POIs, 13,104 street-level images and 120,250 tweets and for Athens 2,501 POIs, 9,840 street-level images and 93,858 tweets. A visualization of the POIs pinned on the map could be seen in figure 4.10 for both Amsterdam and Athens.

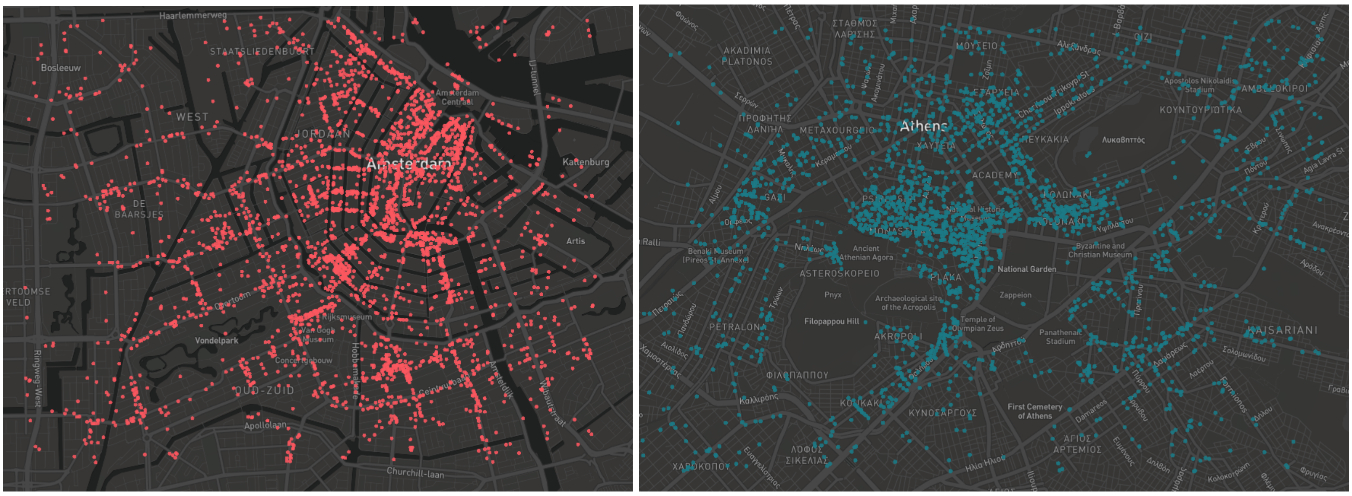


Figure 4.10: Overviews of collected POIs from the centre of Amsterdam (left) and Athens (right)

4.2 Extraction of POI features

In this section the implementation details of the system used for the POI features extraction are being discussed. The reasoning behind the selection of the POI features is presented in the *Experiment Design* chapter.

4.2.1 Extraction of functional features

The functional features are extracted from the collected Google and Foursquare POIs. Both sources' APIs provide the POI data as a json object. Thus, most of the features could be extracted quite straightforwardly, without the need of extra steps. Particularly, the extraction per feature is realized in the following ways:

Type: The Google POIs, gathered through the Google API, have in many cases their type defined only as "point of interest" or "establishment". Therefore, the main places' function is derived from the Foursquare POIs for which the "type" fields are more consistent. Foursquare POIs contain a maximum of four different "type" characterizations with the first being the most specialized (e.g. hotel) to the fourth being the most generalized (e.g. Travel & Transport). In this work, the most specialized types are the ones used to describe the places' main function. Depending on which are the place types to be studied, the type extraction process could slightly change. For instance, to gather all the "restaurants" one has to retrieve all the Foursquare POIs which include the word "restaurant" in their primary or secondary type. In that way, a restaurant labeled as "Turkish restaurant" and another labeled as "Italian restaurant" both belong under the same type "restaurant". The same holds for the rest of the selected POI types as well (e.g. "Boutique" and "Women's Store" are both labeled as "Clothing store" which is their secondary type in Foursquare). Obviously, if someone needs to study specific types of restaurants this process would differ.

Opening Times: Opening times are directly extracted from Google POIs without any further processing. For each day, the places' opening and closing times are stored in a 24-hour clock form as a single number (e.g. 0800 and 2000).

Most Popular Timeslots: Most popular timeslots are extracted by processing the popular times from Google POIs. Basically, for each hour of each day a number is assigned which quantifies how popular this place is. In this work, each day is broken in four parts : morning (05:00 - 12:00), afternoon (13:00 - 17:00), evening (18:00 - 21:00) and night (22:00 - 04:00). Only integer values of time are used since the popular times are also provided for "integer hours" (e.g. 13:00 and not 13:20). For each day, the most popular part of the day is extracted and stored (e.g. Monday: afternoon, Tuesday: evening etc).

Phone number included: This feature is directly extracted from the Google and Foursquare POIs.

Facebook/Twitter included: These features are extracted directly, from the Foursquare POIs.

Price: This information is directly extracted from the Foursquare POIs .

4.2.2 Extraction of Socio-topical features

The socio-topical features are extracted from the tweets which have been sent within a radius of 50m from each POI. They are used to express some of the places' social aspects. More details about the reasoning behind each selected feature is presented in the Experiment Design chapter. Each feature is being extracted by the aggregation of the tweets per place. The extraction methods used per feature are the following:

- **Amount of posted tweets**: Counting of the tweets, in total and per language.
- **Average number of words**: Calculation of the average number of words , per language and in total.
- **Average tweet's sentiment**: For the calculation of each tweet's sentiment two natural language processing python libraries have been used: TextBlob [1] and Polyglot [2]. TextBlob is a widely used library which produces nice quality results. Its main drawback is that it supports only the English language. Thus, for calculating the sentiment of the tweets written in dutch and greek, the tweets had to be translated first. Since this translation, might lead to a great information loss the polyglot library is also used due to the fact that it supports the sentiment analysis in multiple languages (including dutch and greek).
- **Time differences between tweets**: The average and median time difference between consecutive tweets per place. These attributes' main disadvantage is that if some tweets are missing the time difference might end up significantly different and, therefore, it might not represent the reality accurately.
- **Topics**: The collection of the tweets per city is used as the corpus from which the topics are defined. At first the tweets are pre-processed, meaning removing stopwords and punctuation, lemmatizing words and converting emojis to text. This pre-processing was different for each language (e.g. different stopwords among english, dutch and greek). Then, for each POI the respective tweets are merged into a single "document" and a probabilistic distribution of topics is assigned to each POI according to this "document". In that way each POI is expressed as a unique topic probability pattern. As previously explained, the used method for topic modelling is the Latent Dirichlet Allocation which one of the most widely used methods. The implementation of this part is based on the gensim Python library [3] which is specialised in unsupervised semantic modeling from plain text. LDA method requires to tune the amount of topics (k) to be extracted. In this work, the tuning of k is based on the coherence and perplexity of the model. Coherence is used to reassure the interpretability of the extracted topics and is used as it is important to understand what do the extracted topics represent and consequently why do they contribute significantly or not to the prediction of POI types. Perplexity is a common statistical measure which is used to compare LDA models for which a different value of k has been used, and it basically compares the distribution of words in the extracted topics with the actual distribution of words in the

[1] https://textblob.readthedocs.io/en/dev/

[2] https://pypi.org/project/polyglot/

[3] https://radimrehurek.com/gensim/index.html

used documents (in this case the tweets per each POI). Thus, according to the values of coherence and perplexity the chosen amount of extracted topics k is equal to 10 and 4 LDA models are built: one for each city and language. It is worth mentioning that the models are not that sensitive to the different amount of topics (apart for extreme values of k).

4.2.3 Extraction of experiential features

The experiential features consist of features extracted from POI data and features extracted from the processing of Google reviews. According to their extraction they could be divided into two categories: in the first category of features are the **ratings**, **likes** and **photos count** which are extracted straight from the POI data without the need of extra actions. In the second category, the features are extracted from the Google reviews, namely **amount of posted reviews**, **review's average number of words**, **average reviews' sentiment** and **reviews' discussed topics**, using the exact same methods and techniques which are above described for the extraction of the respective features from the tweets. Basically, for each POI the respective collected Google reviews are merged into a single "document" and using the LDA method 10 topics are extracted from them. As for tweets, once again 4 LDA models are used: one for each city and one for each language. It is worth mentioning that the topics extracted from the Google reviews are more interpretable than the ones extracted from the Tweets and in various cases they describe in a straightforward manner the POI types, in contrast with the topics extracted from the tweets for which the interpretation is harder. The most important topics extracted from the Google reviews are thoroughly presented in the next sections (e.g. in table 4.5).

4.2.4 Extraction of visual features

As stated before (Experiment Design chapter) the extracted visual features could be divided into two categories: scene detection features and object detection features. For the implementation of the previously described scene detection algorithm the python library places 365 [1] is used. In this library several models are included from which a deep residual network (ResNet) with 18 layers, a ResNet with 50 layers, AlexNet and DenseNet161 are used. Since the contribution of the visual features to the classification of urban place types is almost the same for all those three models, only the results when using the ResNet model are presented. An example of the result of this algorithm could be seen in figure 4.11

For the **object detection** part, the Tensorflow system is used [1] and the implementation is based on pre-trained state-of-the-art models trained on the Open Images[2] and the COCO dataset [3] [29] offered by the tensorflow object detection API [4]. An example of the result of this algorithm could be seen in figure 4.11.

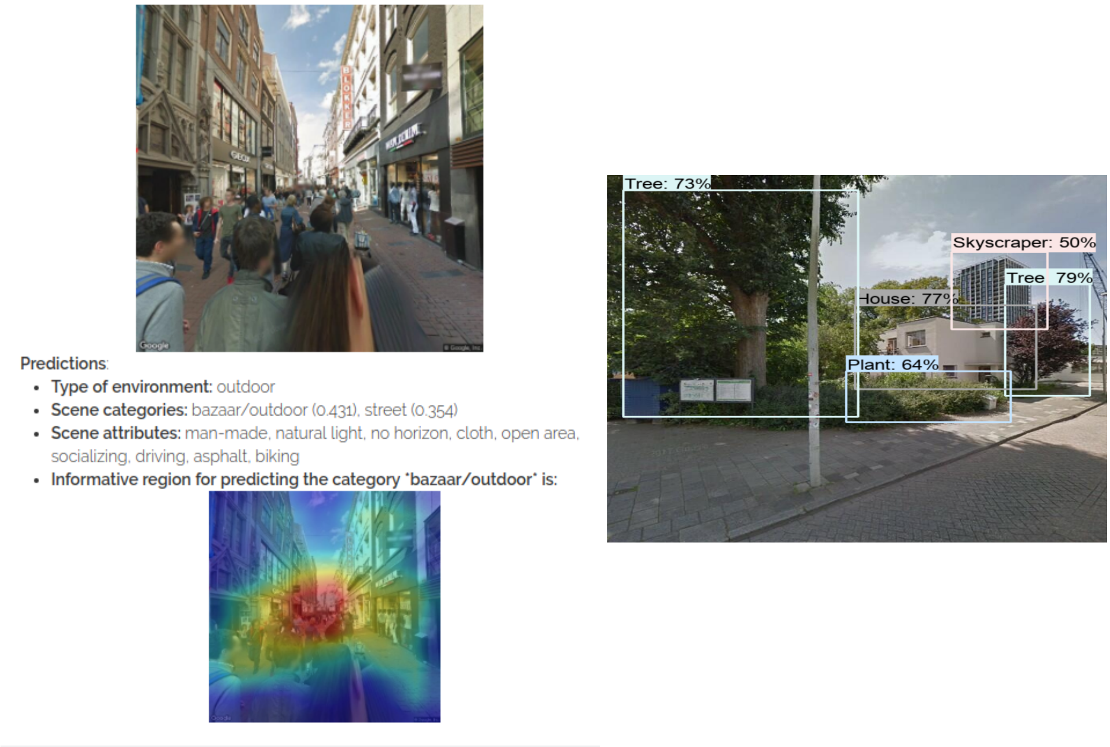
[1] https://github.com/CSAILVision/places365

[2] https://storage.googleapis.com/openimages/web/index.html

[3] http://cocodataset.org

[4] https://github.com/tensorflow/models/tree/master/research/object_detection

Since four images are used for each POI, the final step of the extraction of the visual features is the aggregation of the extracted entities per POI. In that way, for each POI a total amount of detected objects and recognized scenes is stored.


Predictions:

- **Type of environment:** outdoor
- **Scene categories:** bazaar/outdoor (0.431), street (0.354)
- **Scene attributes:** man-made, natural light, no horizon, cloth, open area, socializing, driving, asphalt, biking
- **Informative region for predicting the category 'bazaar/outdoor' is:**

Figure 4.11: Results of the scene recognition (left) and the object detection (right) algorithms.

4.2.5 Extraction of Locational features

All the data are stored in a PostGIS database so that spatial queries are feasible. The extracted locational features for each place consist of the amount of POIs per type which exist within 100m, 1km and 5km of the POI's location. The extraction of this information is based in a method of PostGIS [1] which enables retrieving all the PostGis database's elements which exist within a specified distance from a location.

4.3 POI Type Classification

This section aims to reveal which of the extracted POI features contribute – and to what extent – to the classification of urban place types. "What makes a place" is transformed in this part in questions such as "what makes a bar" or " what makes a restaurant". The two cities are first analysed separately. Then, to study the extent on which the results are region-specific a direct comparison between Athens and Amsterdam is being made.

[1] ST_DWithin method: https://postgis.net/docs/ST_DWithin.html

The initial set of POI types consists of: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym, college and university, coffee shop and art gallery. This set of classes is referred as ***Set A*** in the rest of this document. In addition, ***set B*** is defined which consists of the following types: shop and service, nightlife spot, food, travel and airport, arts and entertainment and college and university. Set B uses the less specialized type fields of Foursquare POIs. The reasoning behind using two type sets is to see if the results are affected by the level on which the POI types are specialized.

4.3.1 Prediction model selection

This study has focused so far to the extraction of valuable POI features towards the classification of POI types. The final step before the prediction which is important in order to understand the value of the extracted features is the selection of a model suitable to the nature of this problem. In other words, the selection of a classifier which is able to take advantage of the extracted features for the prediction of the POI types.

Due to the amount of the features used and the different nature of each feature it is relatively hard to make assumptions on which type of classifier would be the most beneficial to use in the specified problem. To gain a first impression of which classifier could better perform, different types of classifiers which have been used in the literature effectively on various multiclass problems and belong to different "families" are trained on the above-mentioned data. The tuning process of the classifiers' parameters is based on a validation set which includes POIs coming from the outer part of Amsterdam and Athens, and not the center. It is worth mentioning that the selected list of classifiers does not include deep learning models for two reasons: first the amount of the data is not that large to support a deep learning solution and second the main reason of the classification part is to understand the features importance, something which is much clearer using machine learning than deep learning approaches. The used classifiers are the following:

- **Linear Discriminant Analysis (*lda* [1])**: The *lda* classifier is one of the most appropriate linear classifier when dealing with multiclass classification problems. It is based on the assumption that the independent variables are normally distributed and it works well when the data are linearly separable. In addition, *lda* is not supposed to work well when the features are strongly correlated. Intuitively, it is possible that several of the extracted features are strongly correlated (e.g. the similar features gathered from Google and Foursquare POI data such as the "ratings") and, therefore, *lda* is not expected to perform very well. In addition, this correlation could make the feature importance task very hard since the importance of the correlating features would be underestimated.

- **SVM**: Support vector machines choose the decision boundaries in such a way that the space between the regions defined by the features of each class, represented as points in space, are as separated as possible. Several kernels were tried for the SVM model and the one which performed the best is radial basis function kernel [2]. Generally, svm

[1] Not to be mistaken with the Latent Dirichlet Method (LDA) topic model mentioned before.

[2] https://en.wikipedia.org/wiki/Radial_basis_function_kernel

models tend to work well with high dimensional data, are quite flexible and relatively robust to noise. Thus, this model is expected to perform relatively well.

- **K-Nearest Neighbor (KNN)**: The KNN classifier decides the label of each instance, which is represented by its numeric features, based on the majority of the k-nearest-neighbors' labels, where "nearest" is used in terms of a metric distance (e.g. Euclidean distance). This classifier often performs badly for high dimensional data as the distance between the nearest and the farthest neighbor is quite small in the high dimensional space.

- **Random Forest**: Random forest is an ensemble model which selects the most often predicted class of several randomly created different decision trees. An important aspect of random forest is that it generally works well with missing data. This is important for the specified task as there are indeed some missing values in the created dataset (e.g. for some POIs no tweets or ratings were found). In addition, random forest provides quite straightforward methods for the computation of the features importance which is an important advantage for this work.

- **eXtreme Gradient Boosting (XGB)**: The Gradient Boosting classifier is similar to random forest in the sense that it is an ensemble of multiple decision trees. However, in the case of Gradient Boosting those trees are not created at random. Each tree basically tries to correct the errors of the previously used tree. The XGB classifier follows the same principle as the gradient boosting but with a different regularization method which in many cases improves the results. A more suitable name of the XGB, quoting its author of XGB could be "*regularized gradient boosting*". XGB is admittedly a highly flexible algorithm which has been successfully used in various classification problems, thus, is a promising one for this case as well.

- **Ensemble**: An ensemble classifier is also built whose results are based on the majority voting of the results of the four classifiers which performed the best: RF, LDA, SVM and XGB. Intuitively, the combination of conceptually different classifiers could be proved beneficial for predicting the POI types for which the classifiers tend to "disagree". If, for instance, the majority of the classifiers agree on the classification of a POI type, there is a higher probability that this classification is correct even if the best performed classifier has made a different prediction.

In figures 4.12 and 4.13 the F1 (macro) score of each classifier is presented for Amsterdam and Athens, respectively. F1- score is the harmonic mean of precision and recall (formula 4.1). It is used as the evaluation metric in this situation, due to the fact that it takes into account both precision and recall of each class and it is not affected by the unbalanced dataset. Moreover, the evaluation method which is used is 10-fold cross-validation.

$$F_1 score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.1}$$

As could be observed, the tree-based algorithms seem to work quite better from the rest for the current problem, as expected. Unexpectedly, LDA is also one of the best performed classifiers. Overall, for both cities the ensemble method and the XGB classifier perform the best. As a result XGB is selected for the rest of the predictions since its results compared to the ensemble method are almost identical and XGB is much faster. In addition, it is worth noticing that for both cities the order of the classifiers according to their performance is the same (e.g. KNN classifier is the worst in both cases, SVM the second worst and so on) and the results are impressively similar.

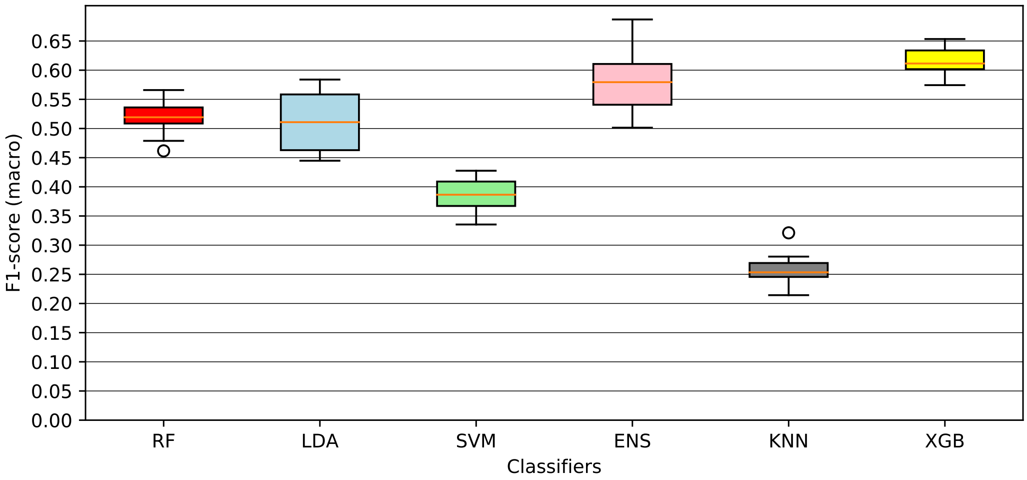


Figure 4.12: **Amsterdam - F1 (macro) Score per Classifier**. The predicted classes (Set A) are: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym, coffee shop, college and university, art gallery and nightclub. Evaluation method: 10-fold cross validation. "LDA" refers to Linear Discriminant Analysis Classifier.

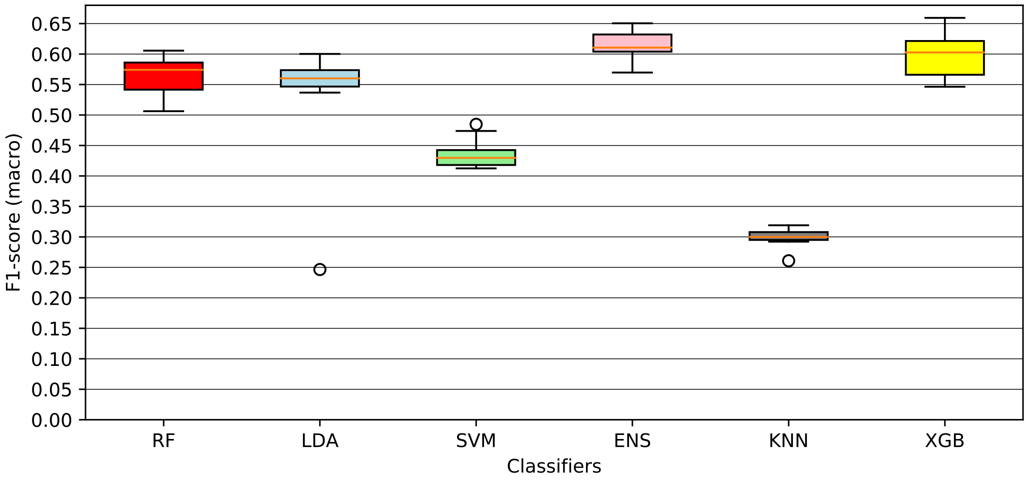


Figure 4.13: **Athens - F1 (macro) Score per Classifier**. The predicted classes (Set A) are: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym, coffee shop, college and university, art gallery and nightclub. Evaluation method: 10-fold cross validation. "LDA" refers to Linear Discriminant Analysis Classifier.

After the selection of the classifier, two approaches were followed in order to handle the **unbalanced dataset**. The first approach is data-centric and is based on the Synthetic Minority Over-sampling Technique (**SMOTE**) [9], one of the most common oversampling techniques. The second approach is classification model-centric and is based on adjusting weights to the XGB classifier. The weights for each class are computed by using the sklearn library [1] and they are inversely proportional to the class frequencies. Since the latter approach worked better than the former, the **XGB classifier** with **adjusted weights** is selected to be used for the rest of this study.

Finally, due to the fact that various features are categorical and the used libraries work only with numerical data, the categorical features had to be transformed as dummy variables [2] and the amount of the finally used features is relatively large (394). Consequently, the principal components analysis (PCA) dimensionality reduction technique was also tried. However, the use of PCA didn't lead to better results, therefore it is not used in the rest of the analysis.

4.3.2 Study of the POI features

The study of the POI features is focused into two main parts: revealing which are the features that make POI types distinguishable[3] (**Distinguishing POI Types** part) and which are the features that "make" a place (**Analysing POI Types** part).

In other words, the **first part** aims to uncover which features tend to be the best indicators towards the prediction of the POI type among a specified set of types and which are the most similar and dissimilar POI types. The important features in this case do not explicitly describe what are the features that "make" a place but what are the features which help the most in discovering what is the type of POI. Towards this goal the performance of the classifier when trained on different feature sets (using 10 POI types) is computed and compared. For instance, imagine trying distinguishing bars, restaurants and clothing stores. To do so, one could be based on various qualities of the places such as the opening and closing times of a place (part of functional features) or the visual appearance of the exterior of a place (part of visual features). Maybe it would be even more beneficial to focus on what are the qualities of this place that the people write about in their reviews (experiential features) or on what are the nearby places (part of locational features)? Finally, useful information could be indirectly provided by the tweets which have been sent from around this place (part of socio-topical features). Thus, this part concentrates on which are the most important features for discovering what is the type of a POI among a set of specified POIs.

In the **second part**, focus is given to some specific POI types and the goal is to understand which are the features which better define those POI types. Thus, a machine learning algorithm is trained again on different feature sets but this time to predict if a POI is of a specified type or not. In that way the features which are the most important when answering a binary question such as "is this a restaurant or not?" are being discovered. The rest of this section is organized as following:

[1] http://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_sample_weight.html

[2] https://en.wikipedia.org/wiki/Dummy_variable_(statistics)

[3] The initial list of studied POI types is: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym, college and university, coffee shop and art gallery.

I **Distinguishing POI Types**

a) **Feature Set Importance - Classes: Set A**. The extracted features are firstly analysed grouped in the above mentioned sets [1]. Thu, if for example the classifier performance is higher when trained on the functional than on the visual features it is indicated that the POI types are being represented more accurately by the places' functional characteristics than their visual appearance.

b) **Feature Set Importance - Classes: Set B**. This part is similar to the above with the difference that more generalized classes are used in order to discover the importance of the "level of specialization" of the classes (e.g. if the fact that "Bar" is a more specialized class than "nightlife spot" affects the results).

c) **Feature Set Importance - Regions Comparison**. According to the above results (classes set A) a comparison between Amsterdam and Athens is realized.

d) **Confusion Matrices**. Intuitively "cafe" and "bars" are more similar than "bars" and "clothing stores". The goal of this part is to reveal if this is indeed the case and, in more general terms, to discover which classes tend to be more similar. For this purpose the confusion matrices are computed and analyzed.

e) **Features Importance**. In this part the features are analyzed individually and not in sets, as before.

f) **Most Important Features Analysis**. This part includes a deeper analysis of the most important features which is needed to understand how is their importance is justified due to the complex nature of the multiclass problem.

g) **Least Important Features Analysis**. As some of the features proved to be way less significant than others, special focus is given to the least important features and feature sets in order to understand if they contain any valuable information.

II **Analysing POI types**:

a) **Feature Sets Importance** . Similarly, the importance of the feature sets is analyzed when focusing on specific POI types. The previous problem aims to classify a POI among several classes while in this part the classification problem is binary: does a POI belong to one specified type or not?

b) **Features Importance**. This part reveals what are the most important features that "make" a specified POI type.

[1] Functional, experiential, socio-topical, visual and locational.

4.3.3 Study of features for distinguishing POI types

Feature Set Importance - Classes: Set A

In Table 4.2 the performance of the XGB classifier when predicting the POI types among the 10 classes included in set A is presented based on four performance metrics: accuracy, precision, recall and (macro) F1-score. Each row of the table corresponds to the performance of the classifier when trained on a separate feature set. The F1-score (formula 4.1) is emphasized and is considered to be the most valuable metric as it is not affected from the dataset being imbalanced, it takes into account both precision and recall and it is commonly used in the literature for the evaluation of classifiers in multiclass problems.

As could be seen in table 4.2, the combination of all the feature sets improves the performance of the classifier with an F1-score of **0.606** for Amsterdam and **0.609** for Athens. Additionally, the most important feature sets are the **functional**, includes features such as the opening/closing times of a place (**0.54** for Amsterdam and **0.481** for Athens), closely followed by the **experiential**, includes features such as topics extracted from the Google reviews (**0.464** for Amsterdam and **0.365** for Athens). The least important feature sets are the **socio-topical**, includes features extracted from the tweets (**0.194** for Amsterdam and **0.143** for Athens) and the **visual**, includes features extracted from the Google Street View images (**0.145** for Amsterdam and **0.154** for Athens).

XGB Classifier - Classes set: A

Amsterdam Data

Feature Set	**Accuracy**	**Precision**	**Recall**	**F-score**	**Data Sources Used**
Functional	0.645	0.590	0.540	0.540	Google & FSQ POIs
Experiential	0.624	0.532	0.456	0.464	Google-FSQ POIs & G. Rev.
Locational	0.389	0.383	0.234	0.244	Foursquare POIs
Socio-topical	0.364	0.278	0.199	0.194	Tweets
Visual	0.326	0.245	0.156	0.145	Google Street View
All	0.728	0.631	0.599	0.606	All

Athens Data

Feature Set	**Accuracy**	**Precision**	**Recall**	**F-score**	**Data Sources**
Functional	0.528	0.556	0.458	0.481	Google & Foursquare POIs
Experiential	0.455	0.415	0.359	0.365	Google-FSQ POIs & G. Rev.
Locational	0.436	0.480	0.350	0.364	Foursquare POIs
Socio-topical	0.263	0.171	0.150	0.143	Tweets
Visual	0.251	0.186	0.163	0.154	Google Street View
All	0.641	0.651	0.588	0.609	All

Table 4.2: Prediction performance based on different features. Predicted classes: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym, coffee shop, college and university, art gallery and nightclub. "FSQ" stands for "Foursquare".

Feature set Importance - Classes: Set B

Furthermore, an important aspect to be studied is how much the results are influenced by how "specialized" the selected POI types are. To understand the "specialization", "Restaurant" and "food and drink shop" Foursquare POIs include a second label which is "Food" for both. For that reason, a second set of POI types is being taken into account, *set B* in which the studied POIs are grouped according to their secondary type to the following categories: arts and entertainment, shop and service, college and university, travel and transport, nightlife spot and food.

Similarly to the table 4.2, table 4.3 is created and it presents the performance metrics of the XGB classifier when predicting the classes of *set B*. Once again, the classifiers perform best when all features are used with an F1-score of **0.671** (+0.065 comparing to set A) for Amsterdam and **0.643** (+0.034 comparing to set A) for Athens. As expected, when using the set B labels the classifier performs better than for set A (less amount of labels and more diverse classes). Overall, the results are quite consistent with the results of table 4.2 in terms of the significance of the feature sets yet some differences exist.

XGB Classifier - Classes set B

Amsterdam Data

Feature Set	Accuracy	Precision	Recall	F-score	Data Sources Used
Functional	0.721	0.651	0.593	0.598	Google & Foursquare POIs
Experiential	0.719	0.655	0.520	0.544	Google-FSQ POIs & G. Rev.
Locational	0.549	0.525	0.380	0.305	Foursquare POIs
Socio-topical	0.564	0.461	0.254	0.256	Tweets
Visual	0.490	0.350	0.190	0.180	Google Street View
All	0.805	0.719	0.645	0.671	All

Athens Data

Feature Set	Accuracy	Precision	Recall	F-score	Data Sources
Functional	0.678	0.671	0.453	0.501	Google & Foursquare POIs
Experiential	0.648	0.547	0.378	0.415	Google-FSQ POIs & G. Rev.
Locational	0.676	0.656	0.396	0.453	Foursquare POIs
Socio-topical	0.554	0.253	0.179	0.173	Tweets
Visual	0.511	0.266	0.186	0.155	Google Street View
All	0.768	0.763	0.581	0.643	All

Table 4.3: Prediction performance based on different features. Predicted classes: shop and service, nightlife spot, food, travel and transport, arts and entertainment, college and university

If the feature sets are ordered by the corresponding F1-score, the resulted order is the same for Amsterdam (i.e. Functional, Experiential, Locational, Socio-topical, Visual). For Athens, there are two small changes: one, the experiential and locational feature sets seem to be of equal importance when set A is used while when set B is used the locational feature set is of slightly higher importance than the experiential. Second, the visual feature set is more important than the socio-topical feature set when set A is used while the opposite occurs when set B is used. In addition, for set A the overall F1-score is almost the same for both cities while for set B the F1-score for Amsterdam is higher than for Athens.

Feature Set Importance - Regions Comparison

In this sub-section, Amsterdam is compared with Athens. The reason for this comparison is to discover if the POI features contain information about the peculiarities of each city. If, for instance, the visual features are good predictors of the POI type for Amsterdam and not for Athens, it might mean that places in Amsterdam are more visually diverse than in Athens. In addition, this comparison aims to investigate to what extent those results are region-specific. By already knowing some of the peculiarities of each city a qualitative interpretation of those results is presented in the Discussion chapter.

In table 4.4, the F-score performances of the XGB classifier trained on each city's data and per feature set are presented. As could be observed, the difference between the two cities is relatively low when comparing most of the feature sets. The highest difference is found for the **locational** features (**+0.12 for Athens**). The second and third largest differences are presented for the **experiential** (**+0.099 for Amsterdam**) and the **functional** (**+0.049 for Amsterdam**) feature sets. Another worth mentioning fact depicted by the results is that, the classifier performs better for Athens when trained on two feature sets: the locational and the visual. For all the rest of the feature sets the F1-score is higher for Amsterdam. Eventually, when all features are used the difference between the two cities is very low (+0.003 for Athens). Thus, even if the various feature sets (and features) have not equal importance for the two cities, the combination of all the features lead to similar results. Thus, the obtained results suggest that the followed approach worked equally for both Amsterdam and Athens.

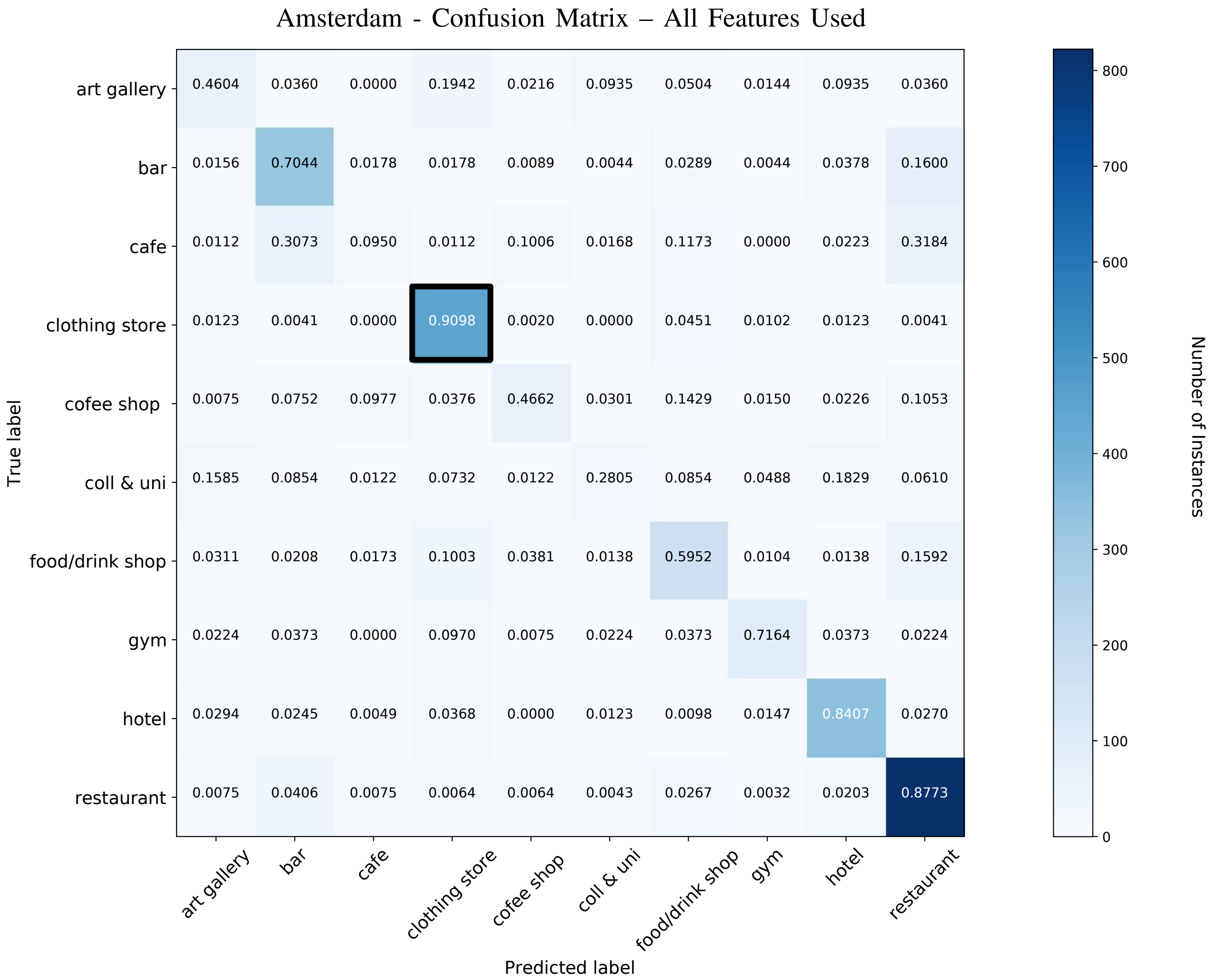
XGB Classifier - Classes set: A (Cities Comparison)				
Feature Set	**F-score - AMS**	**F-score - ATH**	**Difference**	**Best Case**
Functional	0.540	0.481	0.049	Amsterdam
Experiential	0.464	0.365	0.099	Amsterdam
Locational	0.244	0.364	0.120	Athens
Socio-topical	0.194	0.143	0.051	Amsterdam
Visual	0.141	0.154	0.013	Athens
All	0.606	0.609	0.003	Athens

Table 4.4: Prediction performance based on different features for Amsterdam and Athens. Predicted classes: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym, coffee shop, college and university, art gallery and nightclub.

Confusion Matrices

To gain a deeper understanding of which POI types seem "similar" to the classifier, the confusion matrices when using all the feature sets, for both Amsterdam and Athens, are computed and presented in figures 4.14 and 4.15. Interestingly, for both cities the three most accurately predicted classes are the same, namely "**clothing store**", "**hotel**" and "**restaurant**".

Particularly, regarding **Amsterdam**, figure 4.14, "clothing store" is the best predicted label with **91%** of instances being correctly classified while "cafe" is the worst predicted label, with a large difference from all the rest, with only the **10%** of them being correctly classified. This misclassification of the POIs labeled as "cafe" is due to the fact that most of the "cafe" are wrongly classified as "bars" (31%) and "restaurants" (32%). The second most misclassified label is the "college and university" with **28%** of correct predictions. "College and university" instances tend to be classified mostly as "art gallery" (16%) and "hotel" (19%).

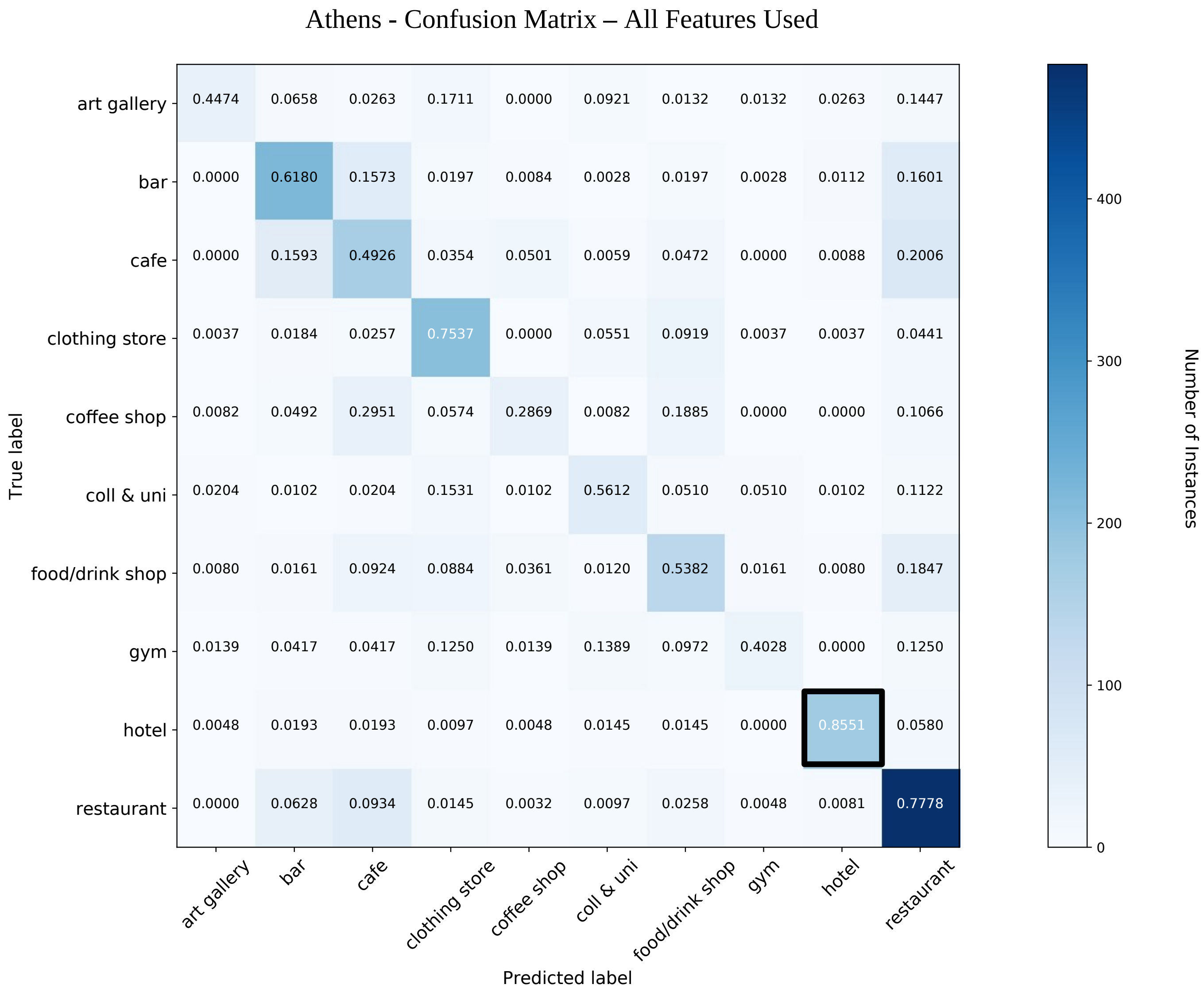


	art gallery	bar	cafe	clothing store	cofee shop	coll & uni	food/drink shop	gym	hotel	restaurant
art gallery	0.4604	0.0360	0.0000	0.1942	0.0216	0.0935	0.0504	0.0144	0.0935	0.0360
bar	0.0156	0.7044	0.0178	0.0178	0.0089	0.0044	0.0289	0.0044	0.0378	0.1600
cafe	0.0112	0.3073	0.0950	0.0112	0.1006	0.0168	0.1173	0.0000	0.0223	0.3184
clothing store	0.0123	0.0041	0.0000	0.9098	0.0020	0.0000	0.0451	0.0102	0.0123	0.0041
cofee shop	0.0075	0.0752	0.0977	0.0376	0.4662	0.0301	0.1429	0.0150	0.0226	0.1053
coll & uni	0.1585	0.0854	0.0122	0.0732	0.0122	0.2805	0.0854	0.0488	0.1829	0.0610
food/drink shop	0.0311	0.0208	0.0173	0.1003	0.0381	0.0138	0.5952	0.0104	0.0138	0.1592
gym	0.0224	0.0373	0.0000	0.0970	0.0075	0.0224	0.0373	0.7164	0.0373	0.0224
hotel	0.0294	0.0245	0.0049	0.0368	0.0000	0.0123	0.0098	0.0147	0.8407	0.0270
restaurant	0.0075	0.0406	0.0075	0.0064	0.0064	0.0043	0.0267	0.0032	0.0203	0.8773

Figure 4.14: Confusion Matrix for **Amsterdam**. It is important to notice that the different colors represent the absolute amount of predicted instances and not the accuracy which is represented by the actual numbers.

For Amsterdam, the types could be divided in three groups according to how accurate are their instances predicted. Starting from the most accurate, the **first group** includes "restaurant", "bar", "hotel", "clothing store" and "gym" (over **70%** of correct predictions per each class), the **second group** includes "food and drink shop" and "art gallery" (over 45% of correct predictions per each class) and **the third group** includes "cafe", "college and university" and "coffee shop" (10%, 28% and 47% respectively).

For **Athens** (figure 4.15) the best predicted label is "hotel" for which **86%** of the POI instances are correctly classified. The worst predicted label is "coffee shop" for which the **29%** are correctly classified. Most of the misclassified "coffee shops" are classified as "cafe" (29%) and "food and drink shop" (19%).



	art gallery	bar	cafe	clothing store	coffee shop	coll & uni	food/drink shop	gym	hotel	restaurant
art gallery	0.4474	0.0658	0.0263	0.1711	0.0000	0.0921	0.0132	0.0132	0.0263	0.1447
bar	0.0000	0.6180	0.1573	0.0197	0.0084	0.0028	0.0197	0.0028	0.0112	0.1601
cafe	0.0000	0.1593	0.4926	0.0354	0.0501	0.0059	0.0472	0.0000	0.0088	0.2006
clothing store	0.0037	0.0184	0.0257	0.7537	0.0000	0.0551	0.0919	0.0037	0.0037	0.0441
coffee shop	0.0082	0.0492	0.2951	0.0574	0.2869	0.0082	0.1885	0.0000	0.0000	0.1066
coll & uni	0.0204	0.0102	0.0204	0.1531	0.0102	0.5612	0.0510	0.0510	0.0102	0.1122
food/drink shop	0.0080	0.0161	0.0924	0.0884	0.0361	0.0120	0.5382	0.0161	0.0080	0.1847
gym	0.0139	0.0417	0.0417	0.1250	0.0139	0.1389	0.0972	0.4028	0.0000	0.1250
hotel	0.0048	0.0193	0.0193	0.0097	0.0048	0.0145	0.0145	0.0000	0.8551	0.0580
restaurant	0.0000	0.0628	0.0934	0.0145	0.0032	0.0097	0.0258	0.0048	0.0081	0.7778

Figure 4.15: Confusion Matrix for **Athens**. It is important to notice that the different colors represent the absolute amount of predicted instances and not the accuracy which is represented by the actual numbers.

An important difference between the two confusion matrices is that in the case of Athens, the accuracy among the classes vary less than in the case of Amsterdam. Particularly, the correct classifications of each class vary for Athens from **29%** to **86%** while for Amsterdam they vary from **10%** to **91%**. Thus, both extremes, meaning the highest and lowest scores, are observed for Amsterdam.

Feature Importance

Moreover, the importance of each single feature (top 15) for both Amsterdam and Athens is presented in figure 4.16 and 4.17. This computation is based on the "Gain" of the XGB classifier which basically expresses how much is the training loss decreased when each feature is used for splitting. Since the goal of the classifier in this section is to distinguish the POI types, those features are not meant to better describe every POI type, but to be able to distinguish one POI type from another.

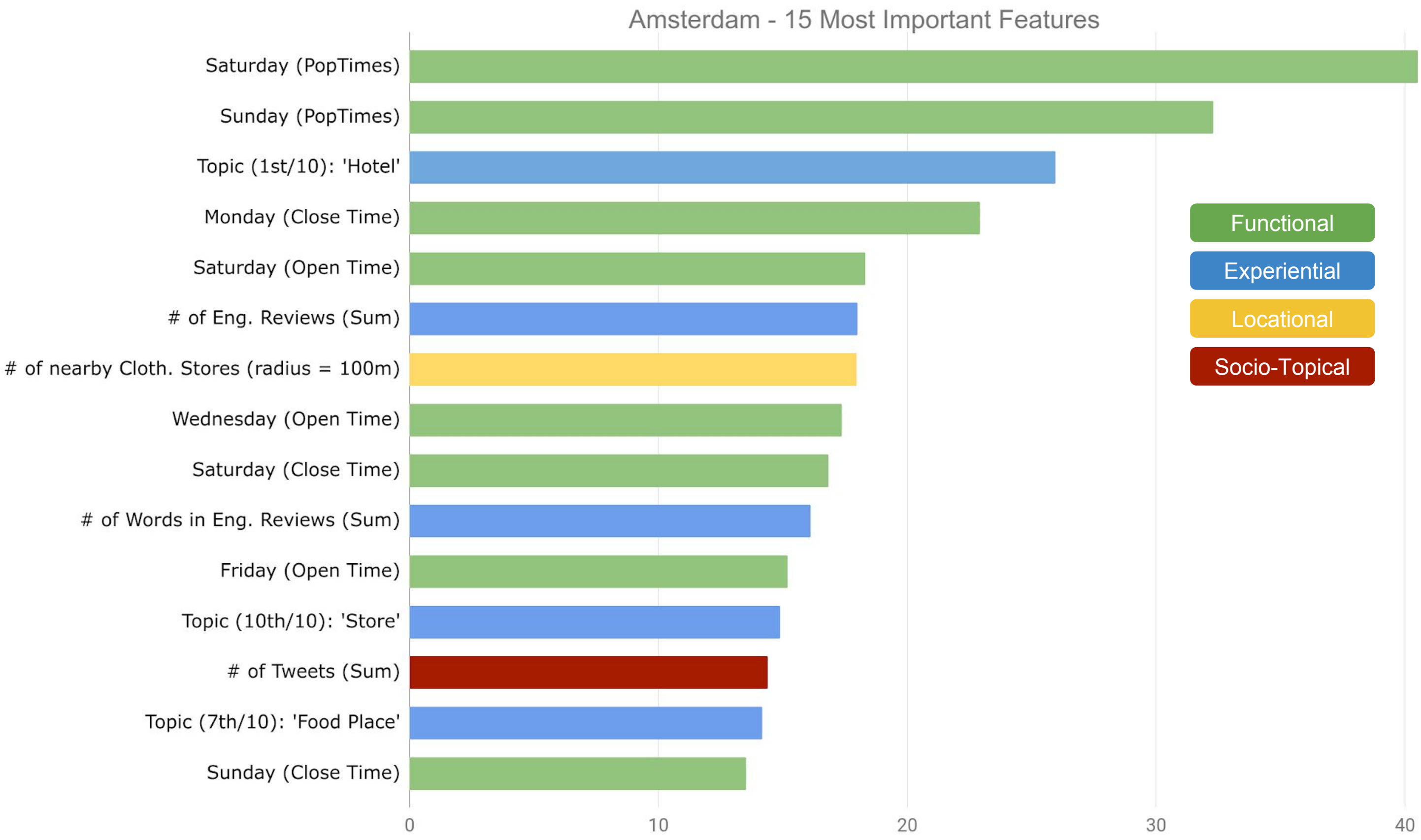


Figure 4.16: Top 15 most important features - **Amsterdam**. The "Topic" features have been extracted using the **LDA** method. Thus, Review Topic (1) expresses the first extracted topic from the **English** written **Google reviews**.

The feature importance scores support the results of table 4.2 as for both cities, the majority of the 15 most important features indeed belong to the **functional**, **experiential** and **locational** feature sets. For Athens, two locational features are included while for Amsterdam only one them is included, and this is also reasonable as the locational feature set is more important for Athens than for Amsterdam. In addition, features which are correlated with specific POI types such as the topics which are interpreted as "Hotel", seem to be of great importance for distinguishing the POI types. Thus, it is indicated that if a feature is able to represent accurately one of the classes, its importance will be high.

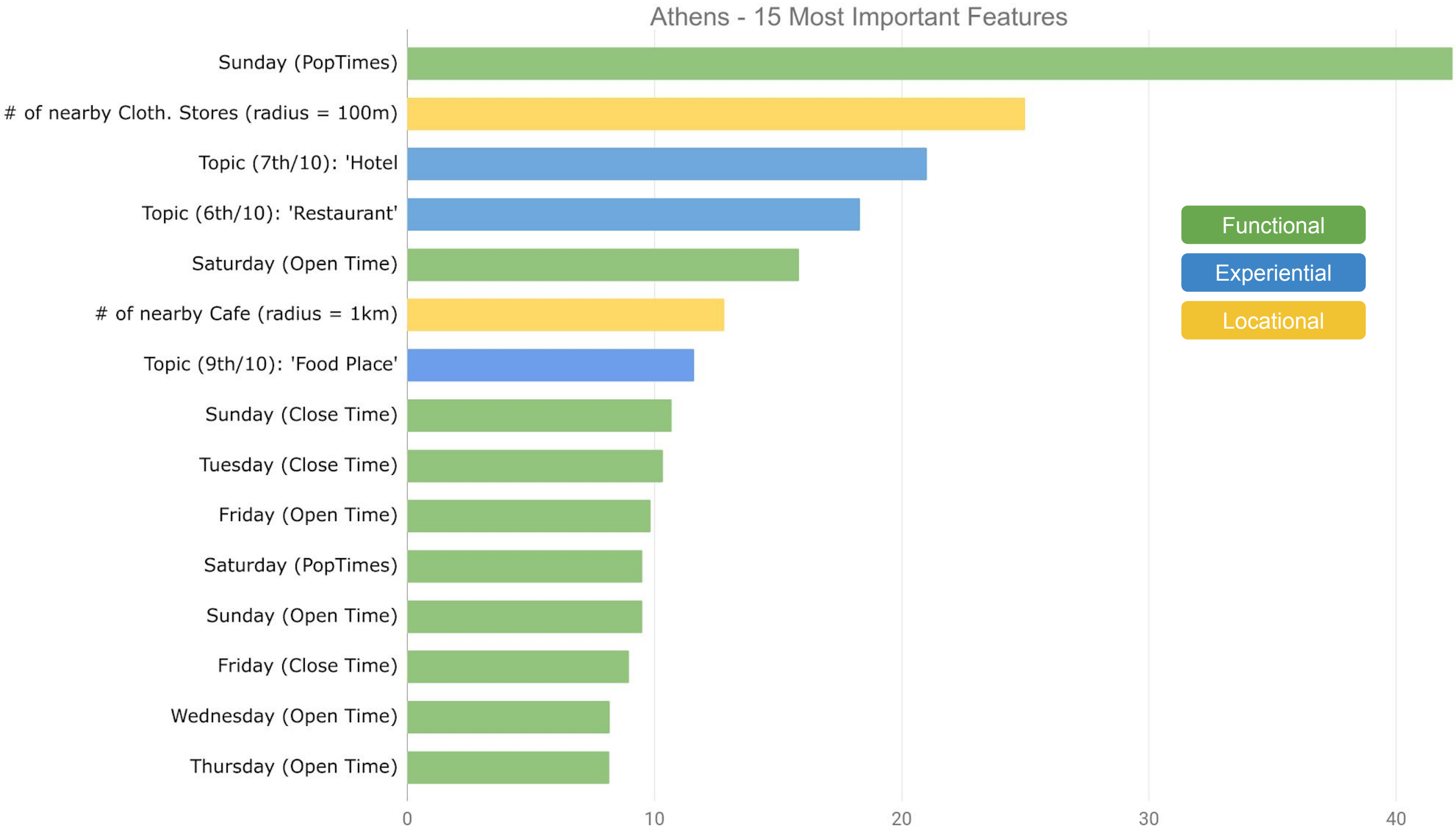


Figure 4.17: Top 15 most important features - **Athens**. The "Topic" features have been extracted using the **LDA** method. Thus, Review Topic (1) expresses the first extracted topic from the **English** written **Google reviews**.

In figures 4.16 and 4.17 a feature which, for example, is named "Review Topic (6th/10): "Restaurant", refers to the **6th** topic extracted from the **English written Google reviews** by the LDA method for which the total amount of extracted topics is equal to **10**. As could be observed the extracted topics from the reviews are among the most important features. On the contrary, the topics extracted from the tweets do not seem to be important as features. Interestingly enough, again for both cities the topics extracted from dutch or greek written reviews are also not that important. Furthermore, for **Amsterdam** the total amount of Tweets sent seem to be included as one of the 15 most important features even if the socio-topical feature set, in which this feature belongs to, is one of the least powerful according to the previous results. Thus, even if generally the socio-topical features are not of great interest according to table

4.2, some of the socio-topical features seem to be useful.

Most Important Features Analysis

As depicted in figures 4.16 and 4.17 some of the most important features for both cities include the extracted **topics** from the **Google reviews**, the **popular times on Sundays** and the **amount of clothing stores within a distance of 100m**. To better understand why those features work well in distinguishing the POI types in both cases a further exploration of those features is realized.

Starting with the features which represent the topics extracted from the **Google reviews** and are important for both cities, table 4.5 includes the explanation of how those topics are interpreted by showing their "main" words. As explained before, for each POI, after the Google reviews have been collected , they are merged into a single "document" and the probability of how relative this "document" is to each topic is computed through the LDA method. Thus, each POI is expressed through a topic probability distribution according to its Google reviews. Each topic is described as the weighted sum of a set of words. Those words and their respective weights, which define to what extent each word represents the subject of the respective topic, are included in the column "Main words". The "Interpretation" column includes a subjective interpretation of what seems to be the subject that each weighted sum describes. As could be seen, the resulted topics are quite specific to certain POI types.

City	**Topic**	**Main Words**	**Interpretation**
Amsterdam	1st/10	0.048*room + 0.035*hotel + 0.022*staff + 0.021*location + 0.018*clean	Hotel
Amsterdam	7th/10	0.045*food + 0.026*great + 0.022*service + 0.021*place + 0.019*friendly	Food Place
Amsterdam	10th/10	0.037*store + 0.028*shop + 0.011*love + 0.010*product + 0.009*brandy	Store
Athens	7th/10	0.030*room + 0.026*hotel + 0.018*view + 0.017*staff + 0.015*breakfast	Hotel
Athens	6th/10	0.035*food + 0.019*restaurant + 0.016*greek + 0.013*nice + 0.012*best	Restaurant
Athens	9th/10	0.066*nice + 0.055*place + 0.049*good + 0.045*great + 0.032*food	Food Place

Table 4.5: Main words of the most important topic-features extracted from **Google Reviews** using the LDA method for Amsterdam and Athens. The "Interpretation" column is a subjective interpretation of which subject the specified topic seem to describe.

Furthermore, to understand how the extracted topics from the Google reviews are correlated with the POI types, the average probability of four topics, two for each city and for each POI type are presented in figures 4.18, 4.19. As could be seen, for all those topics the connection between what the topic represents and the POI types (for which this topic has the highest probability to be "present" in their reviews) is quite straightforward. For instance, for Athens the topic which is interpreted as "**restaurant**" has indeed a higher probability to be matched with the Google reviews of "restaurant" POIs and the **Store** topic for Amsterdam is indeed associated mostly with the "clothing store" POIs. Therefore, the high importance of the topics extracted from the Google reviews is justified as they seem to distinguish certain POI types from all the rest.

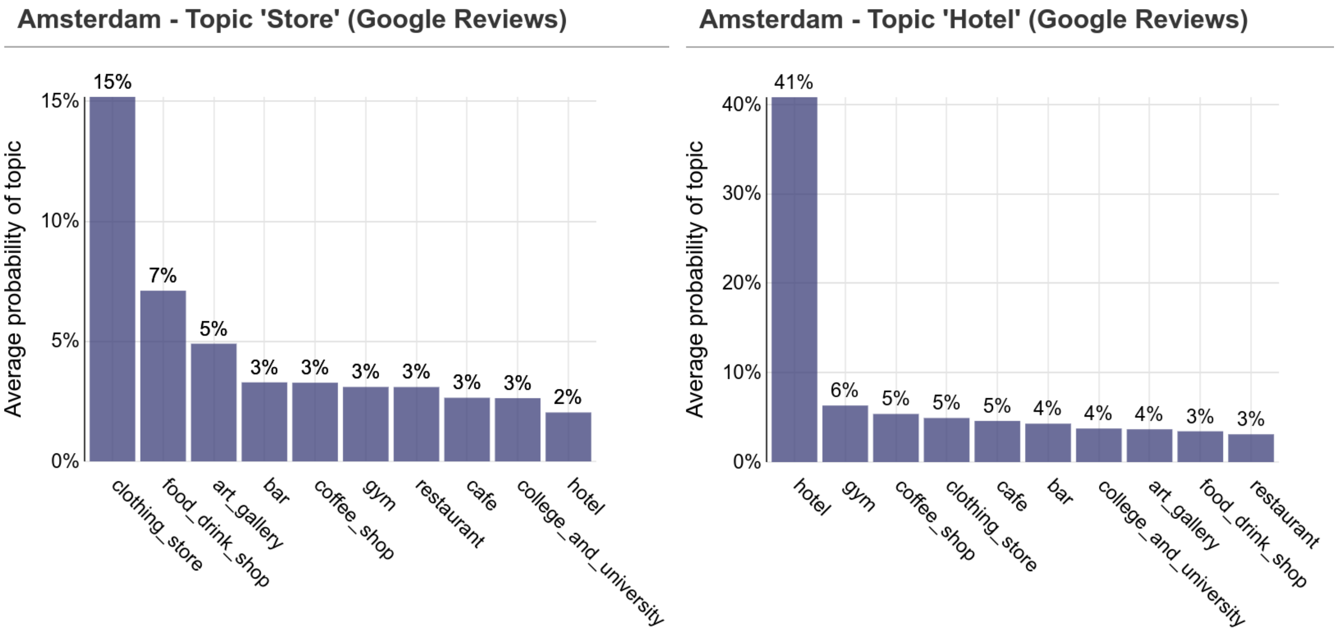


Figure 4.18: Average probability for the topics **Store** and **Hotel** to be present in the english written Google Reviews.

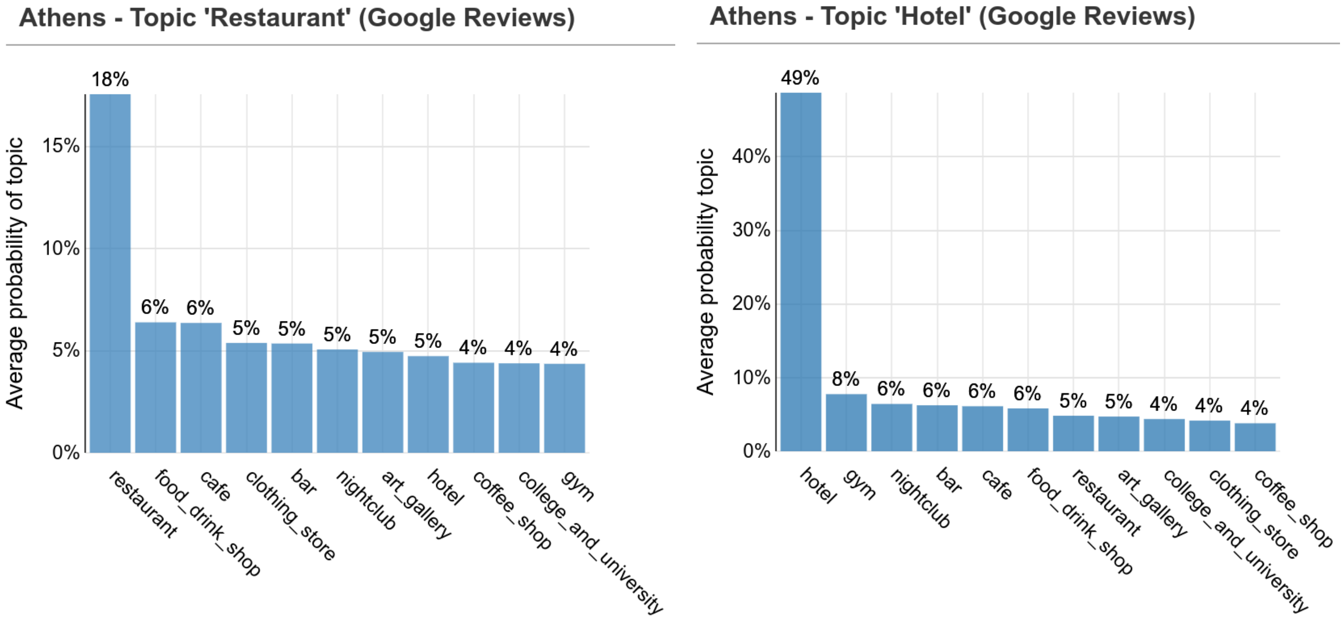


Figure 4.19: Average probability for the topics **Restaurant** and **Hotel** to be present in the english written Google Reviews.

In addition, figures 4.21 and4.20 depict the amount of Amsterdam and Athens POIs, respectively, grouped by type, according to their **most popular timeslot during Sundays**. This feature is the most important one for Athens and the second most important one for Amsterdam. There are four timeslots in total, namely "morning", "afternoon", "evening" and "night" as explained in the Extraction of functional features sub-section. Through this figure, the importance of the "Sunday popular times" features could be understood as the differences among the POI types seem to be quite high. For instance for Athens on Sundays, gyms and clothing stores are mostly visited in the morning while bars are mostly visited, unsurprisingly, at night. Hotels seem to be mostly visited in the afternoon and evening. It is important to notice that the information of "popular times" is not present for all POIs and in figure 3.4 the total amount of the instances per each POI type which include the popular times is showed. For example, as could be seen only for two "art galleries" the popular times are present and as a result all the art galleries seem to be visited in the morning, which could be false. It is possible that if this information existed for every POI the performance of the classifier would be significantly improved.

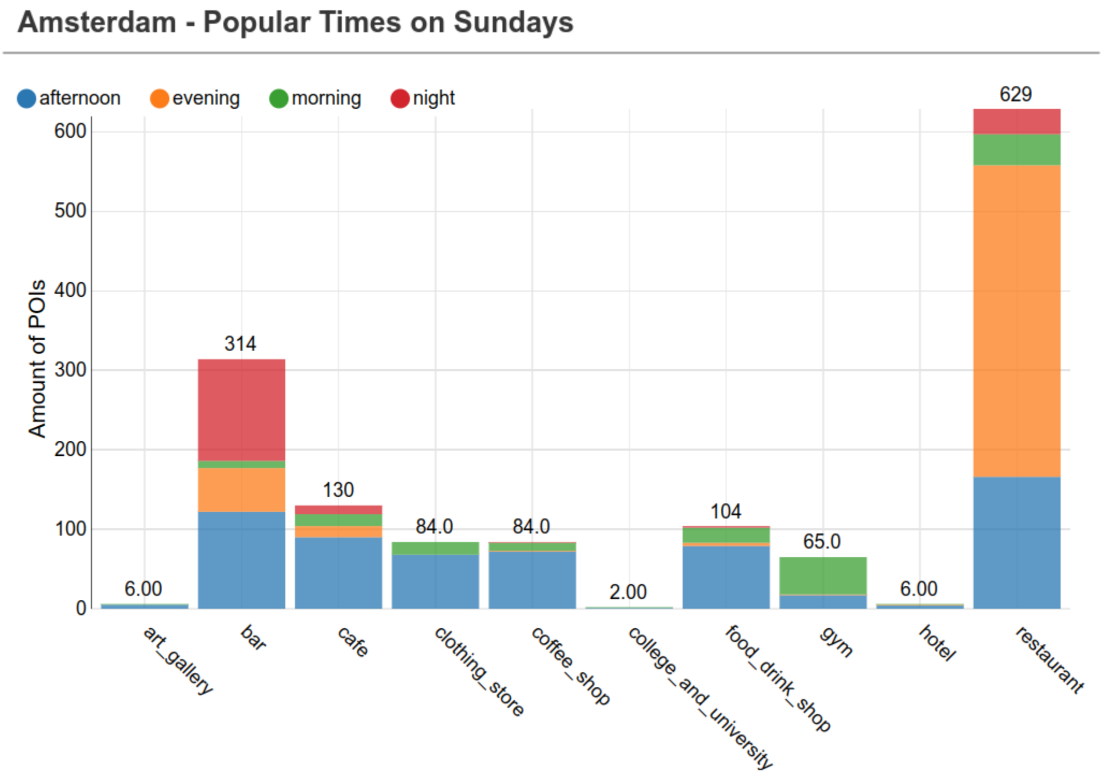


Figure 4.20: Amount of POIs grouped by their **most popular timeslot on Sundays for Amsterdam** among morning (05:00 - 12:00), afternoon (13:00 - 17:00), evening (18:00 - 21:00) and night (22:00 - 04:00) - Third most important feature for Amsterdam (figure 4.16). For example, this figure suggests that restaurants are most popular at the evening.

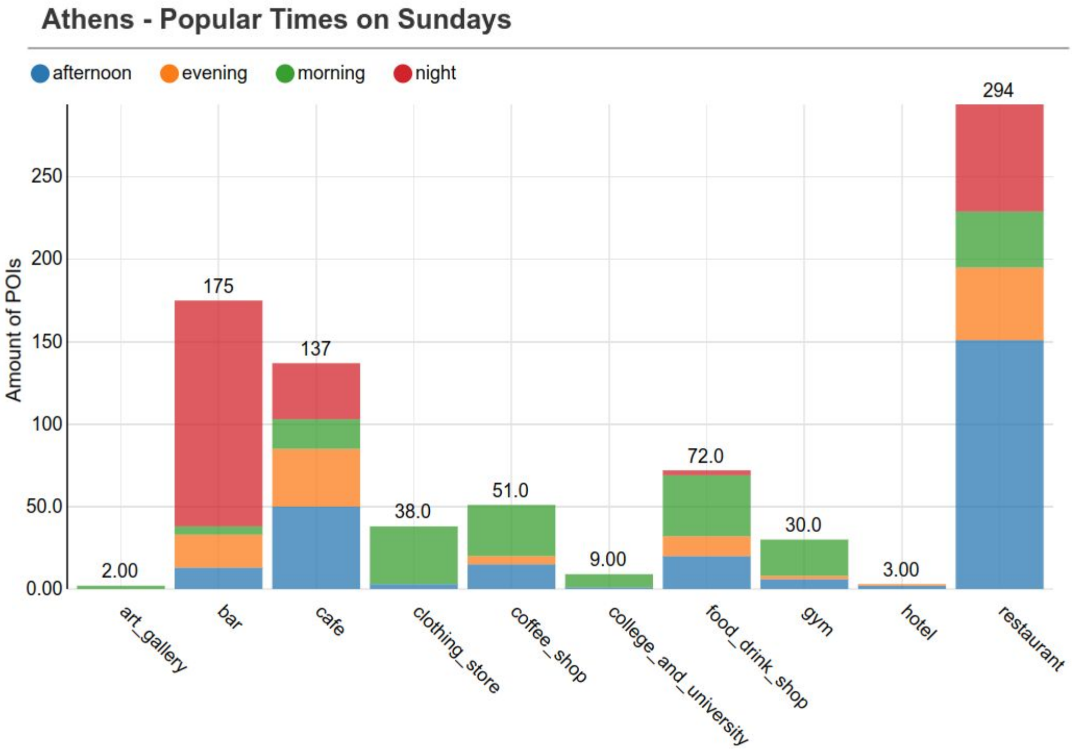


Figure 4.21: Amount of POIs grouped by their **most popular timeslot on Sundays for Athens** among morning (05:00 - 12:00), afternoon (13:00 - 17:00), evening (18:00 - 21:00) and night (22:00 - 04:00) - Most important feature for Athens (figure 4.17). For example, this figure suggests that the bars are most popular at night.

Lastly, figure 4.22 presents the **amount of nearby clothing stores**, withing a radius of 100m, and the actual clothing stores according to their location, on the map. As could be seen, this feature seems to sketch quite accurately the main parts where the clothing stores are located for both cities. Of course, this feature does not seem to be able to capture all the clothing stores as some of them are not clustered and, therefore, the amount of the nearby clothing stores is equal to zero for them. However, since most of the clothing stores seem to be indeed clustered an overview of the distribution of the clothing stores in each city is indeed depicted by the "nearby clothing stores" feature. It is also quite interesting to notice that among all the POI types [1], the clothing stores are those which tend to be the most spatially clustered, and as a result among all the spatial features the "nearby clothing stores" is the most important one, **for both cities**.

[1] All the POI types meaning: Art Gallery, Bar, Cafe, Clothing Store, Coffee Shop, College and University, Food and Drink Shop, Gym, Hotel, Restaurant.

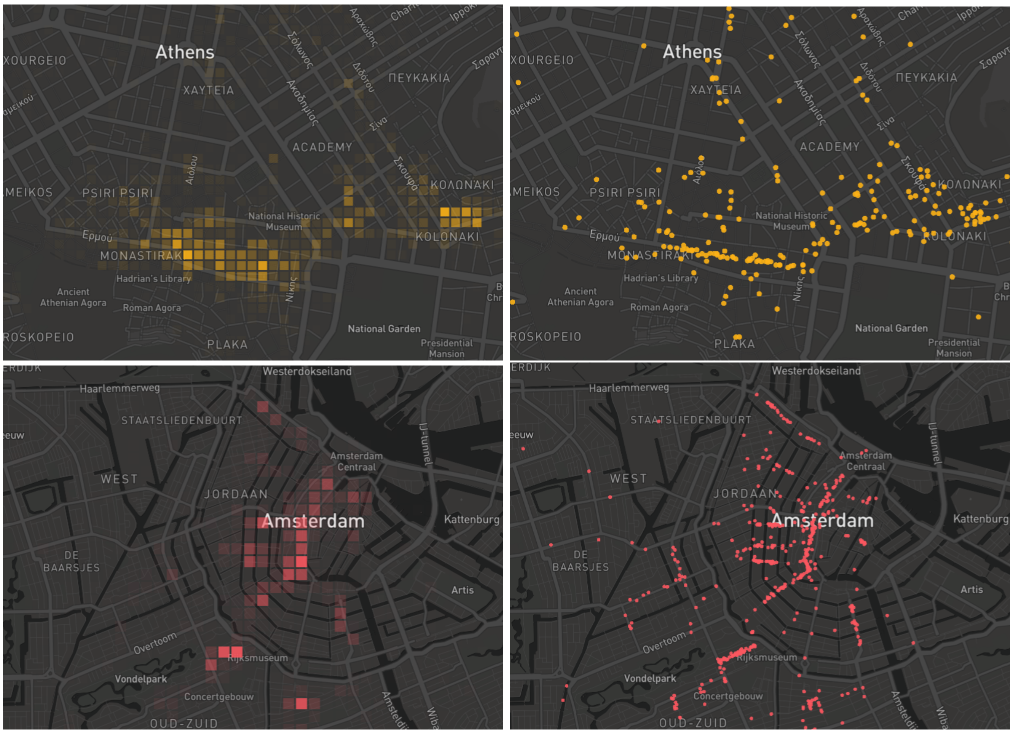


Figure 4.22: Athens (up) and Amsterdam (down) - Nearby clothing stores (radius 100m - left) and actual clothing stores (right)

Least Important Feature Sets Analysis

In this last part special the focus is given on the features which seem to be the least significant: the **visual** and the **socio-topical**. Since those features did not prove to be good POI types indicators when all the POI types are used, only three POI types are selected to be investigated in this section, which seem to be quite distinguishable both by the above-mentioned results (figures 4.14 and 4.15) and by nature: "**Hotel**", "**Clothing Store**" and "**Restaurant**". In that way, it is easier to understand the most important visual and socio-topical features and if they include any significant information for the classification of the POI types. As stated in the Experiment Design chapter, the visual features are extracted from four Google Street View images, which are taken from as close as possible to the outside of each place, using scene recognition and object detection algorithms. Then, the extracted entities from each image are aggregated per each POI. Thus, if for example for in each of the four images of a place the scene "shopfront" has been recognized the total amount of "shopfront" for this place is counted equal to 4. The socio-topical features are extracted from the tweets sent around each place and include attributes such as total amount of tweets, sentiment and the topics extracted from the tweets using the LDA method.

When predicting those three POI types the F1-score of the classifier reaches the values of **0.912** and **0.92** for Amsterdam and Athens, respectively. When using only the **visual** feature set the F1-score of the classifier is **0.47** for Amsterdam and **0.4** for Athens. The most important visual features in this case are presented in figures 4.23 and 4.24. As could bee observed, for both cities the scene recognition features seem to be more important than the object detection ones. Particularly, the three most important features are, for Amsterdam, the "arcade", "tree" and "shopfront" and for Athens the "shopfront", "residential neighborhood" and "downtown". From all those visual features only the "tree" has been extracted using the object detection algorithm.

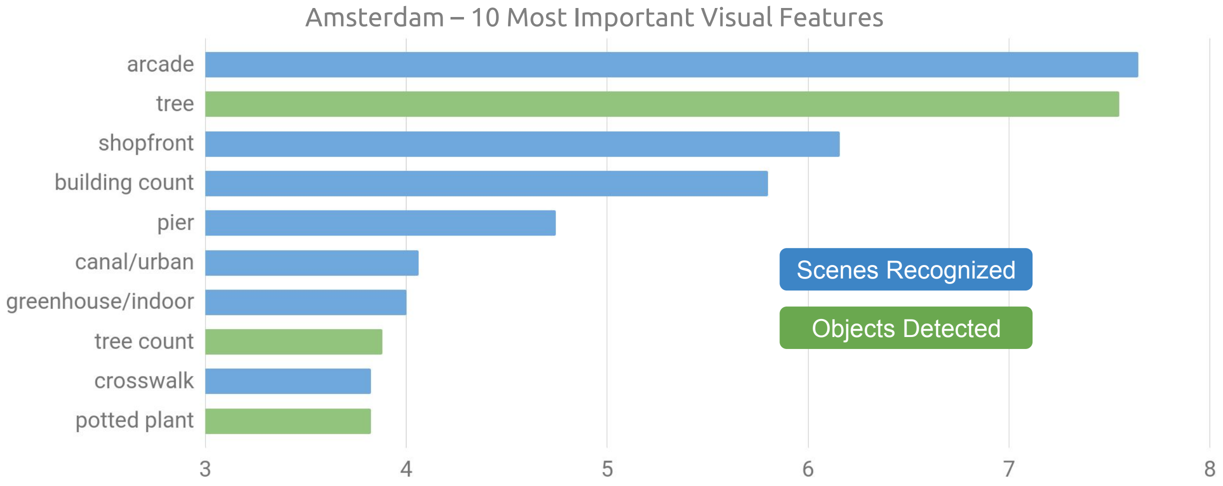


Figure 4.23: Top 10 most important visual features - **Amsterdam** - Classes: Hotel, Clothing Store, Restaurant

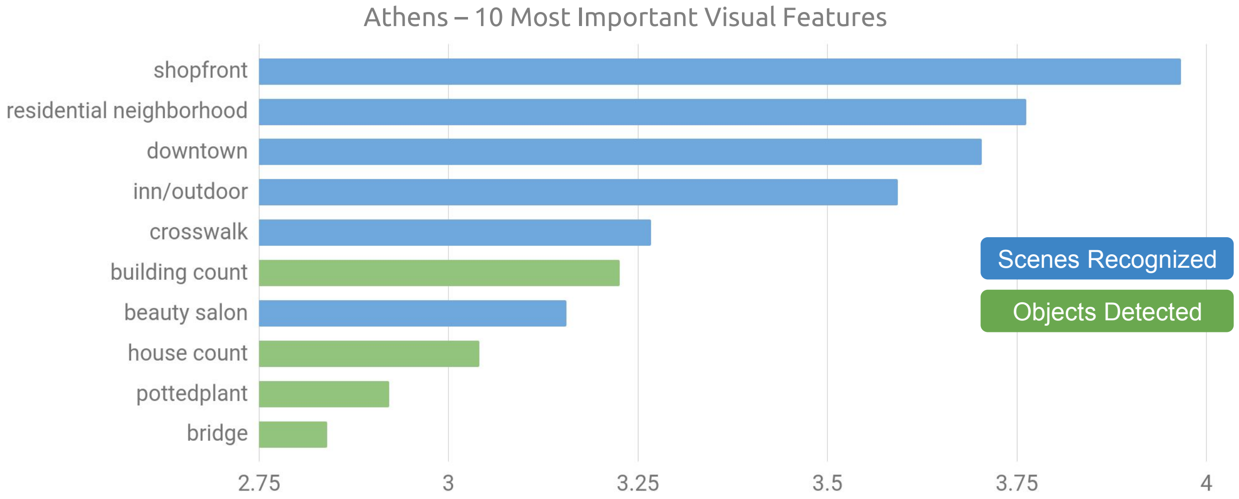


Figure 4.24: Top 10 most important visual features - **Athens** - Classes: Hotel, Clothing Store, Restaurant

In figures 4.25 and 4.26 the average amount of the recognized scenes (in the Google Street View images), which seem to be among the most important **visual features**, per each POI type is presented for Amsterdam and Athens, respectively.

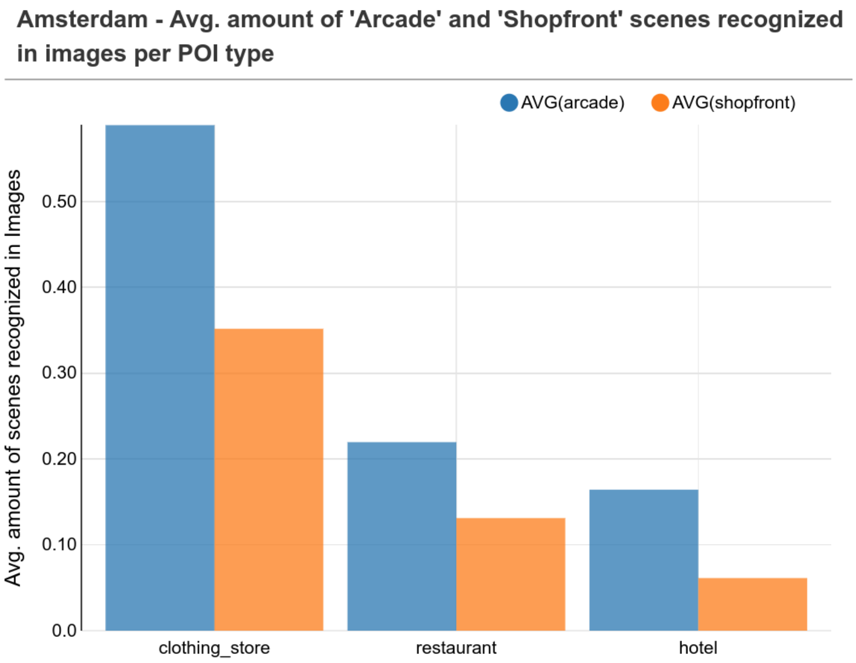


Figure 4.25: Average Frequency of the most important recognized scenes (visual features) per type for **Amsterdam** - Predicted Classes: Hotel, Clothing Store, Restaurant

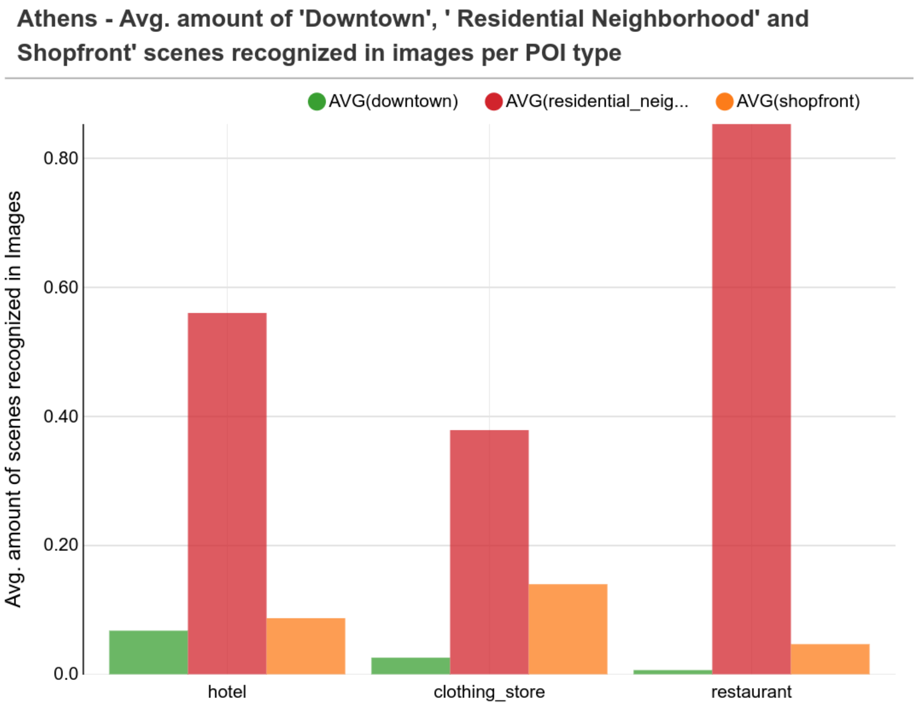


Figure 4.26: Average Frequency of the most important recognized scenes per type (visual features) **Athens** - Predicted Classes: Hotel, Clothing Store, Restaurant

As expected, figure 4.25 supports that both the "arcade" and "shopfront" scenes are more often recognized in the pictures of clothing stores than those of restaurants and hotels, **for Amsterdam**. Similarly, the "shopfront" also occurs more often for the clothing stores in **Athens** than for the other two POI types. On the contrary, again for Athens, the pictures that have been characterized as "residential neighborhood" belong mostly to restaurants and pictures that have been characterized as "downtown" belong mostly to hotels. Those results support that there is indeed valuable information included in the extracted visual features.

When using only the **socio-topical** features such as the amount of tweets sent near a place or the topics extracted from those tweets, the F1-score of the classifier is **54%** for Amsterdam and **45%** for Athens.

The most important socio-topical features for **Athens** are the three topics which are present in table 4.6, one of which is written in Greek. As could be observed, those topics are not as interpretable as the ones extracted from the Google reviews and this is reasonable due to the nature of the tweets (very unstructured text).

For **Amsterdam** the most important socio-topical features are the "total amount of tweets sent" around each POI, the average time difference of the consecutive tweets sent and some of the extracted topics which could be also seen in table 4.6.

Overall, it is important to notice that the least important features seem to also include some valuable information for distinguishing the POI types. Both the tweets and the images contain unstructured data from which the extraction of information is complex and not straightforward. Even if the rest of the feature sets proved to be more important the results support that there is still a relatively large amount of information provided indirectly by those two data sources.

City	Topic	Lang.	Main Words (Translated)	Interpretation
Amsterdam	3rd/10	Dutch	0.047*noordholland (north holland) + 0.031* goed (good) + 0.019*man + 0.018*cafe + 0.012*north	Cafe
Amsterdam	4th/10	Dutch.	0.022*brand (fire) + 0.017*drinking + 0.017*new + 0.017*cafe + 0.014*hotel	Drink/Cafe/Hotel
Amsterdam	1st/10	Eng.	0.083*old + 0.051*hotel + 0.016*art + 0.015*city + 0.012*place	Hotel/Art
Athens	7th/10	Eng.	0.091*day + 0.081*love + 0.023*hotel + 0.020*coffeeislandco + 0.019*thissio	Hotel
Athens	3rd/10	Greek	0.072* μία (one) + 0.054*μόλις (just) + 0.053* δημοσίευσε (publish) + 0.050* φωτογραφία (photo) + 0.016*face_with_tears_of_joy (emoji)	Photo
Athens	4th/10	Eng	0.046*yon [1] + 0.026*issue + 0.023*fashion + 0.016*love + 0.011*studio	Fashion

Table 4.6: Main words of the topics extracted from **tweets** using the lda method. The "Interpretation" column is a subjective interpretation of which subject the specified topic seem to describe.

4.3.4 Analysing POI types

This part aims to discover the features which "make" a place in the sense that those features distinguish one specific POI type from all the rest. The most correctly predicted POI types according to the above presented confusion matrices (figures 4.14 and 4.15) are, for both cities, the "**hotel**", "**clothing store**" and "**restaurant**". Thus, since the extracted features seem to be able to capture the dimensions of those three POI types more accurately than the rest, those are the POI types which are selected to be analyzed. In other words, in the following paragraphs three cases are being studied: predicting if the type of a POI is (1) a hotel, (2) a clothing store, (3) a restaurant.

To be able to compare the results of this part with the previously presented results once again the XGB classifier is used and the F1-score for the evaluation of its performance. Unsurprisingly, the amount of one specific POI type is much less than the amount of all of the rest POI types combined. To avoid having such an unbalanced dataset the following process is followed: let p be the amount of instances of the type to be predicted and n the total amount of types. Then, from each type a random sample of instances equal to $\frac{p}{n-1}$ is retrieved and used so that their sum is equal to the amount of the instances of the POI type to be predicted. Thus, the dataset is balanced and consists of two classes: one representing the type to be predicted and another one representing the all the rest for which the instances are equally distributed among the types and combined they lead to an amount of instances equal to the one of the predicted class.

Feature Sets Importance

In table 4.7, the F1-scores of the XGB classifier when trained on the different feature sets, for each binary classification problem (hotel, clothing store and restaurant) and for both cities are presented. As before, it is important to notice that for each POI type and for both cities the **combination** of all feature sets leads to the **highest F1-scores**. Moreover, as could be observed, there are some differences in the overall F1-score among the POI types. In addition, even if a feature set could lead to a relatively high F1-score when predicting one POI type it might lead to much worse F1-score for another POI type. An example as such is the **socio-topical** feature set for Amsterdam (**77%** for the clothing store and **58% for Restaurant**). This implies that different POI types are "special" for different reasons.

Regarding **Amsterdam**, it seems to be easier to predict if a POI represents a clothing store (**92%**) than a hotel (**89%**) or a restaurant (**0.877**). Again, the most important feature sets seem to be the functional, experiential and locational while the socio-topical and visual tend to be less important. However, for each POI type the results are quite different and, therefore, overall statements are harder to be made.Hotel

For **Athens**, the best results are obtained for the **hotel** followed by the **restaurant** and then the **clothing store**. It is worth mentioning that in the cases of hotel and clothing store the locational feature set leads to better results than the functional or experiential feature sets. On the contrary, for the restaurant the locational feature set does not work that well.

Overall, the results obtained when the classifier is trained on the **functional** and experiential feature sets are quite consistent for both cities and always relatively high. In contrast, the socio-topical and visual feature sets tend to be both not that important in almost all of the cases. Finally, the locational feature set is very indicative of the POI type in some cases while not that much in other.

XGB Classifier - Classes set: A

Amsterdam Data

Feature Set	Hotel F1-score	Cl. Store F1-score	Restaurant F1-score	Data Sources Used
Functional	0.832	0.857	0.823	Google & FSQ POIs
Experiential	0.889	0.784	0.853	Google-FSQ POIs & G. Rev.
Locational	0.633	0.783	0.600	Foursquare POIs
Socio-topical	0.610	0.774	0.579	Tweets
Visual	0.550	0.650	0.558	Google Street View
All	0.892	0.924	0.877	All

Athens Data

Feature Set	Hotel F1-score	Cl. Store F1-score	Restaurant F1-score	Data Sources Used
Functional	0.790	0.776	0.775	Google & Foursquare POIs
Experiential	0.881	0.661	0.774	Google-FSQ POIs & G. Rev.
Locational	0.818	0.781	0.580	Foursquare POIs
Socio-topical	0.65	0.6	0.558	Tweets
Visual	0.550	0.601	0.567	Google Street View
All	0.920	0.820	0.831	All

Table 4.7: Prediction performance based on different feature sets. The columns "Hotel", "Cl. Store" and "Restaurant" represent the three different classification problems that are analysed. For instance, the column "Hotel" includes the results when the classifier tries to predict which POIs are hotels and which are not.

Features Importance

To further understand what are the most important features which, in a sense, define each of the studied POI types, the 15 most important features for each city and each case (i.e. hotel, clothing store, restaurant) are presented in figures 4.27 - 4.29. Generally, the results presented in table 4.7 quite agree with the results of those figures, meaning that the most important features indeed belong to the feature sets that lead to the higher performance of the classifier. For instance, for **Amsterdam** in the third case (**restaurant**) the most important feature sets are clearly the functional and the experiential and in figure 4.29 (for Amsterdam) all the 15 most important features are either functional or experiential. This consistency, however, is not observed in all the cases.

Regarding the **hotel**, figure 4.27, the most important features tend to be part of the **functional** (e.g. opening/closing times, popular times) and **experiential** (e.g. features extracted from reviews) for both cities. Not surprisingly, the extracted topics from the Google reviews which have been interpreted as "Hotel", again for both cities, are present in those figures among the three most important features. This was expected since the fact that the main subject of the reviews is the "Hotel" is reasonable to indicate that the corresponding POI represents indeed a hotel. Other topics which are showed to be important are the "Gym" (for Amsterdam) and the "Food Place" for Athens which are also amenities which could be offered by hotels. In addition, some of the locational features are also among the 15 most important features. The most important **locational** feature is the "nearby hotels" and the "nearby Art Galleries" for Amsterdam and Athens, respectively. For **Athens**, there is also a visual feature among the 15 most important features: the amount of benches. It is interesting to notice that the visual feature set lead to the worst performance of the classifier (with a quite high difference from the rest) and even the **socio-topical** feature set performs better, none of the features included in this set is among the top 15 features.

Thus, the features which mostly "define" a hotel seem to be the opening/closing times, the amount, size and topics of the written reviews and their location in respect to the other places.

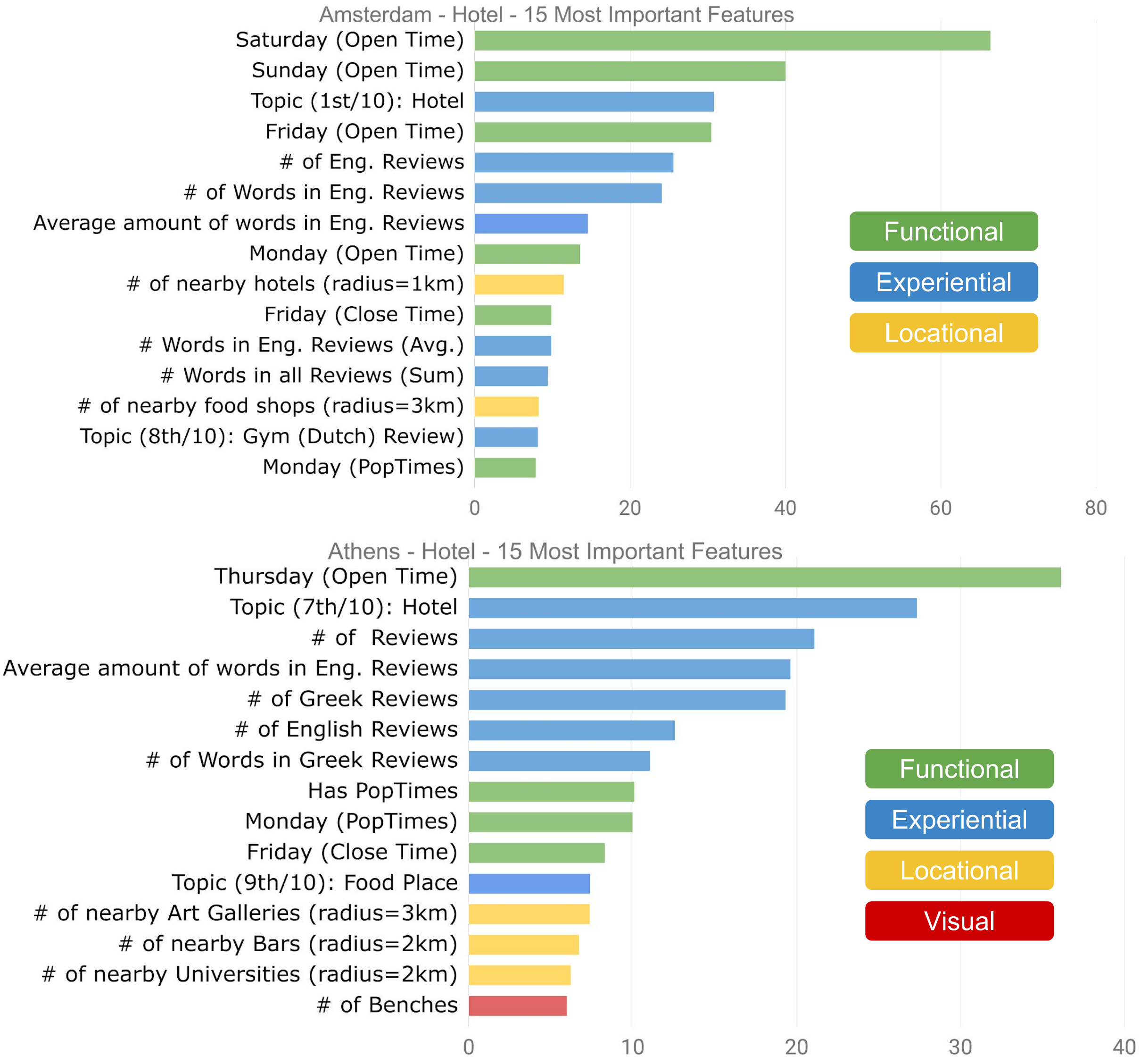


Figure 4.27: The 15 most important features for predicting if a POI is a **Hotel** in **Amsterdam** (up) and **Athens** (down). The "Topic" features refer to the extracted topics from the Google Reviews.

For predicting if a POI's main function is **clothing store**, figure 4.28, the " amount of nearby clothing stores" is the most important feature for both Amsterdam and Athens but the similarities between the two cities are less. The **locational** features seem to be more important for **Athens** than for Amsterdam while for the **socio-topical** the opposite occurs.

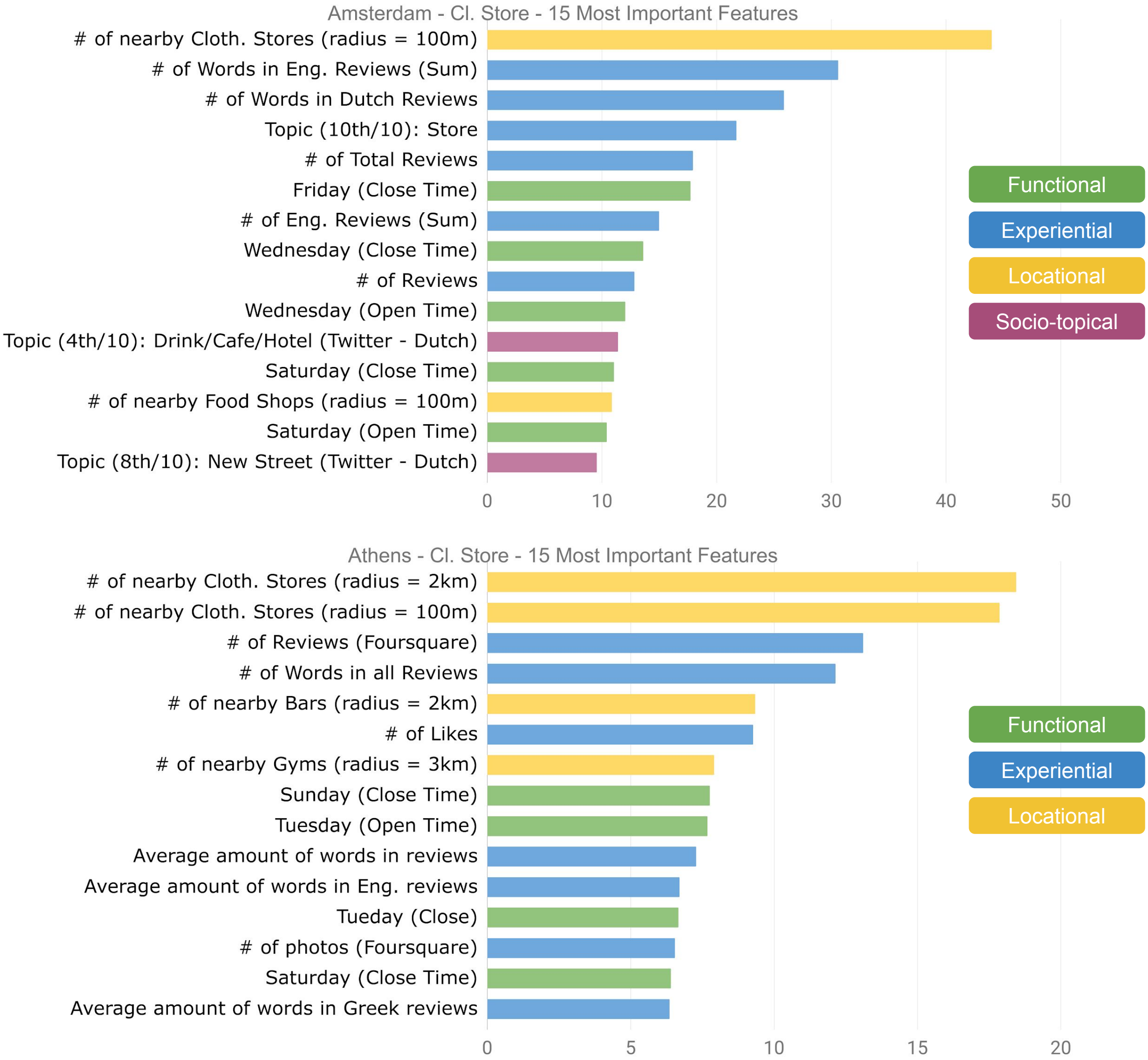


Figure 4.28: The 15 most important features for predicting if a POI is a **Clothing Store** in Amsterdam (up) and **Athens** (down). The "Topic" features refer to the extracted topics from the Google Reviews except for the cases where "Twitter" is written in which the features are extracted from the tweets.

As could be also seen in table 4.7, the case of clothing stores for Amsterdam is the only case in which the **socio-topical** feature set lead to comparable performance to the **experiential** and **functional** (e.g. the topic "Brand" which is extracted from Twitter is the 11th most important feature for Amsterdam).

Overall, the features which seem to better describe the clothing stores are the opening and closing times, the amount, size and topic of the written reviews and their location especially in respect to the other clothing stores which is a sign of spatial clustering.

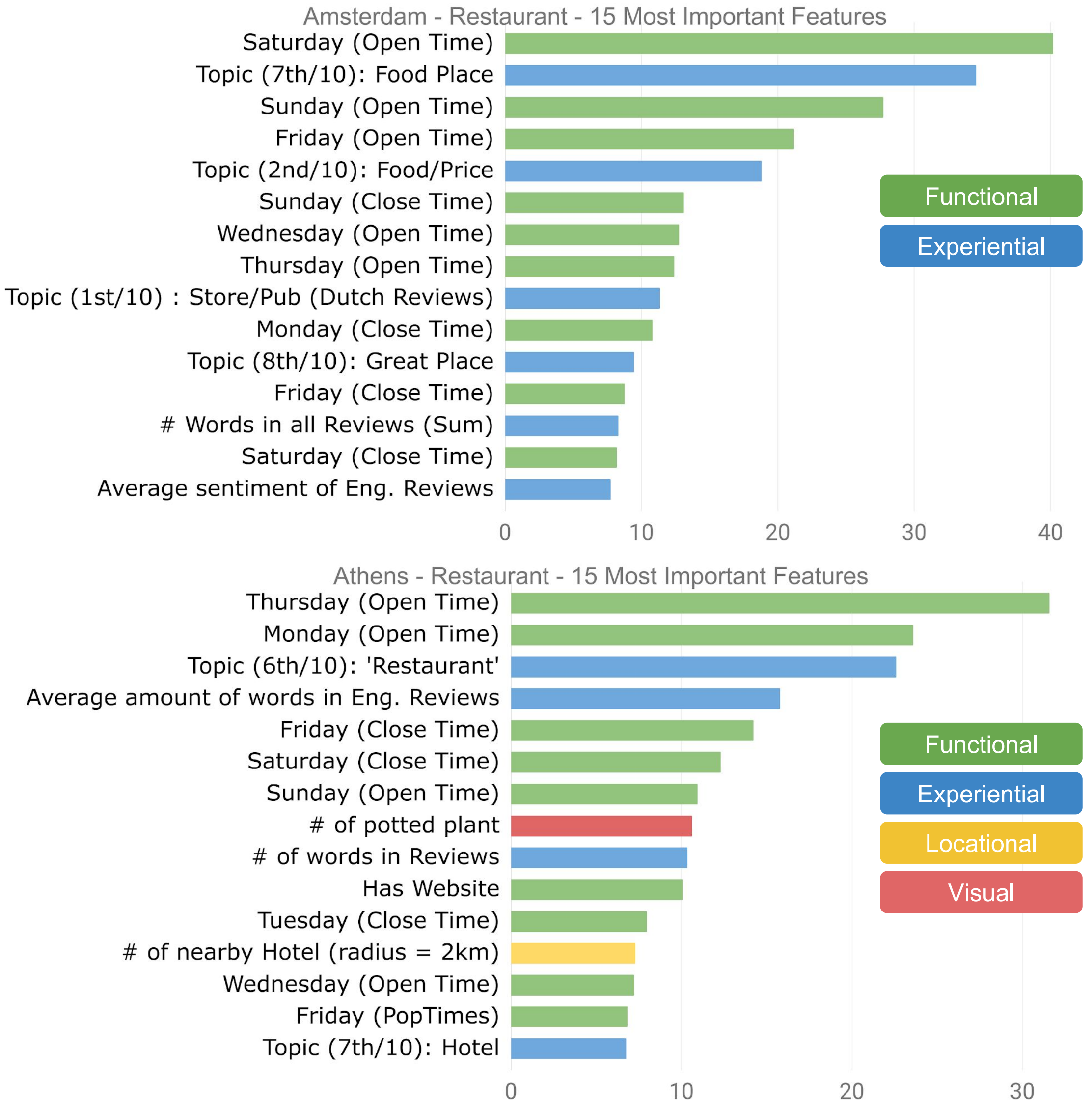


Figure 4.29: The 15 most important features for predicting if a POI is a **Restaurant** in **Amsterdam** (up) and **Athens** (down). The "Topic" features refer to the extracted topics from the Google Reviews.

Finally, the most important features for predicting if a POI's type is "restaurant". figure 4.29, consist mostly of features included in the "experiential" and "functional" feature sets. Once again, topics which are related to restaurants and extracted from reviews such as "Food Place" for Amsterdam and "Restaurant" for Athens tend be among the most important features. It is worth mentioning that for Athens, features from every feature set are present among the most important features. Quite unexpectedly the amount of "potted plants" seem to be quite indicative for restaurants.

The results support that the most indicative features for predicting if a POI represents a restaurant are part of the places' functional characteristics such as opening/closing times and of its experiential characteristics such as amount, size and topic of the written reviews.

The above results suggest that depending on the POI type to be predicted, the features which perform the best as indicators of the type differ. However, even if the individual features vary, in every case the functional and experiential characteristics of the places are dominant among the most important features. The locational features seem to be also quite important in most of the cases but they tend to work better for Athens than for Amsterdam. In contrast, the experiential and socio-topical features lead to better performance for Amsterdam.

For both cities and for all three cases, the **combination** of the features improves the results of the classification. Thus, the results support that indeed capturing the various dimensions of places is done better when multiple and diverse sources are used.

4.4 Chapter Summary

In this chapter, all the implementation details for the realization of each step of the pipeline showed in figure 3.1 in the Experiment Design chapter are explained. In addition, the results of the POI type classification are analyzed towards two main goals: understand which features are good indicators in distinguishing the POI types as a multiclass problem (e.g. discover if a POI represents a restaurant, a bar, a cafe or a hotel) and understand which are the main features which distinguish a specified POI from all the rest (i.e. what "makes" a specific POI). Overall and in the most cases the most important features are the functional characteristics of a place (e.g. opening times) and its experiential characteristics which are mostly extracted from the reviews written for a place.

Chapter 5

Discussion

In this chapter, the results presented in the *Implementation* chapter and the threats to the validity of this study are being discussed.

5.1 Discussion of obtained results

5.1.1 Data collection & Matching

Interestingly enough, even though the data are collected from a same size and shape area from the centre of Amsterdam and Athens, the amount of POIs retrieved from Amsterdam is quite larger than from Athens. Since this difference is proportionally the same when using both Google and Foursquare, meaning that for both sources the amount of POIs retrieved from Athens is almost equal to the 70% of the amount of POIs retrieved from Amsterdam it is reasonable to assume that Amsterdam contains more POIs than Athens. This fact does not necessarily mean that there are more physical places in Amsterdam than in Athens because it is possible that Amsterdam has been better digitized and, therefore, various physical places in Athens have not been digitally represented (at least in Google and Foursquare). However, this difference could also be explained from the fact that Athens contains some hills which cover a quite large area and in which no places and, consequently, points of interest exist.

From the retrieved POIs, 29% are matched using the discussed matching algorithm for both cities, despite the fact that the algorithm was tuned based mostly on the data coming from Amsterdam. This similarity, supports that the matching algorithm could possibly work equivalently for other regions as well. The used matching algorithm is based on the combination of the tree structure logic of [23] and various similarity metrics presented in [36]. The obtained results are similar to both of those works. However, a core difference of this study's approach with [36], is that in their case the authors deal with two sets of POIs for which they already know that for every POI of the one set a match exist in the other set. In this work, some places' digital representation is not included in both sources, thus, some POIs do not have a match. In addition, a main difference with [23], is that the authors only consider to match POIs with a geographical distance smaller or equal to 80m. In this work's matching algorithm, this initial distance is set to 300m and multiple matches were found to be geolocated in a distance more than 80m. It is important to clarify that the focus of this work is not the development of a novel, improved matching algorithm. The matching algorithm is developed for practical reasons meaning the realization of the combination of the two POI data sources.

The overall combination of the four sources is based on the POI data matching of Google and Foursquare and the geolocation-based matching between Google - Google Street View and Foursquare - Twitter. In the literature, it is common to use geo-coordinates from one source as parameters to get data from another source (e.g. [10]). In this work, the sources were selected in such a way so that there are no differences in the geo-location systems among the different data sources (same geolocation systems per pair). Knowing that the quality of the data would set the limitations of this study the POI data collection is meant to be as precise as possible.

Moreover, the amount of **tweets** collected for each city is proportional to the amount of the retrieved POIs which supports the correctness of the tweets collection. The amount of **street-level images** is straightforward as four images are retrieved for every POI. Once again, the amount of POIs for which no street-level image was found is the same, relatively to the total amount of POIs, for Amsterdam and Athens.

5.1.2 POI features & type classification

Throughout the POI type classification, various insights are gained both for the importance of the features and for the particularities of each city. One of the key-results of this study is that for both cities, training the classifier with all the extracted features lead to its best performance. Thus, the use of different sources, data of different nature and features which reflect different dimensions of places indeed assisted towards the better understanding of the POI types.

Furthermore, the comparison of the various features and feature sets provides some important insights. Most of the relevant previous studies focus on a specific data source or a specific type of features and investigate to what extent they could extract valuable POI-based information from them (e.g. [21], [37]). Generally, one could deduct what features to use for a similar purpose according to the results and conclusions of other studies. However, since each study uses different data, gathered in a different period of time and containing different POIs the comparison of the importance of the different data sources and features is implicit and could be considered quite biased. In this thesis, diverse data sources are used and diverse features are extracted while the data remain the same to provide insights which could help in a more straightforward comparison of the important features.

In addition, the initial intuitions for the selection of which type of classifier would be able to better learn from the extracted features mostly proved to be correct. For instance, the tree-based classifiers tend to perform the best for the classification of POI types and this comes in agreement with the fact that those classifiers work well with missing data and with correlated features. This is also supported from [52] in which the random forest classifier is used. On the contrary, the k-nearest neighbor classifier performs the worst among the used classifiers which was also expected as the "distance" in high dimensional spaces does not vary a lot. The only unexpected result was for the linear classifier which performed quite well even though multiple features are strongly correlated. It is also worth noticing that the order of the classifiers by their performance is exactly the same for both cities.

Another important aspect to be considered for the POI type classification is that since the POI type labels have been extracted from Foursquare, which is a user generated platform, the main function of the selected POIs has been basically defined by the people and by how people experience a place. In addition, the various extracted features are also based on people's perception. For instance, the tweets and reviews used have been written by people and express their own thoughts and experiences. The scene recognition algorithm has been trained on a dataset which has been labeled by people and, therefore, basically describes how people visually perceive a place. Thus in this work, an effort has been made to explore places as a combination of characteristics which are perceived by people. The need of this human-perception based digitization of "place" is also explained and emphasized in [39].

Feature Sets Importance

For the POI type classification two main classification problems are studied: (1) a multiclass classification problem in which each POI had to be classified among a set of 10 POI types [1] and (2) a binary classification problem in which the goal is to predict if the POIs' type is a "target" type [2] (e.g. hotel) or not. For both cities and both classification problems the performance of the classifier when trained on different feature sets fluctuates relatively consistently. For instance, the functional feature set in all the cases proved to include good indicators of the POIs' type in contrast with the visual feature set.

Thus, some general remarks about the importance of each feature set, which refer to both classification problems, are discussed in this part.

Starting with the **functional** features the results support, unsurprisingly, that they tend to work well for both cities and cases. However, what was not expected is that they would work better for Amsterdam than for Athens, mostly due to the fact that in Athens the opening times of places are more distributed in the 24 hours of a day. For instance, bars, tend to close really late while restaurants usually close much earlier. Even so, the functional feature set in every single experiment leads to better results for Amsterdam. This difference could be explained by two possible reasons: one that the amount of POIs collected from Amsterdam and used to trained the machine learning algorithm is higher than the amount collected from Athens and consequently the functional characteristics are captured better for the Amsterdam POIs or second that for Amsterdam the percentage of the POIs for which the feature "popular times" existed is higher than for Athens (43% for Amsterdam and 34% for Athens) and this feature proved to be quite important and one of the core functional features.

In addition, a feature set which proved to be truly valuable is the **experiential**. The experiential features consist mostly of the amount, length and topics of the Google reviews per POI. The structured nature of the reviews lead to the extraction of meaningful and interpretable topics. Since the reviews are type-specific they tend to include valuable information about the POIs' type. For instance, it is reasonable that the bar-reviews include different characteristics/topics from the restaurant-reviews as the people who write them focus on different qualities of each place-type to describe it (e.g.

[1] Set A for example includes Art gallery, bar, cafe, clothing store, coffee shop, college and university, food/drink shop, gym, hotel and restaurant.

[2] The studied types for this case are: clothing store, hotel and restaurant.

for bars they might focus on the drinks and for restaurant on the food). Thus, the analysis of the reviews could be considered a way to understand not only how do people experience a place but also what do they expect from it and what are the qualities that, according to them, make a "bar" or a "restaurant".

Admittedly, for both cities the **socio-topical** and **visual features** contribute the least. The small contribution of the socio-topical features was quite expected due to the fact that the social aspects of a place do not seem specialized enough to the POI type (e.g. bars from restaurants) as they do not include so much information about the places' function. Intuitively, the socio-topical features would be better in further characterizing "similar" POIs of the same type as for example "artistic" or "sport" bars, or in identifying characteristics of neighborhoods as in [14]. Nevertheless, still some socio-topical features proved to be of great importance as the amount of tweets in the multiclass classification problem (figure 4.16) or the extracted topics in the binary problem (figure 4.28), for Amsterdam. Therefore their inclusion to the selected features is supported by the results. On top of that, the difference of the contribution of the socio-topical features for the two cities is also notable as for Amsterdam they contribute quite more than for Athens for which two possible explanations are given: one this difference exist due to the fact that in total more tweets have been gathered from Amsterdam than for Athens and second the used natural language processing libraries tend to work better with the english language than with the dutch or greek and more people express their thoughts on Twitter in english in Amsterdam (58% of tweets written in engish) than in Athens (49% of tweets written in english).

Regarding the **visual features**, it was expected that they would contribute more to the POI type classification and that a larger difference would be found among Amsterdam and Athens as Athens is considered to be more visually diverse. The results suggest that indeed the visual features work better for Athens but the difference is so small that is almost insignificant. On the other hand, this difference might have been larger if the amount of training data for Athens were not less than for Amsterdam. Moreover, the non-importance of the visual features could be based on the fact that the selected POI types are quite similar (e.g. bar, cafe, coffee shop and restaurant could be all visually very similar.

Lastly, the largest difference between the two cities is found when using the **locational** feature set for predicting the POI types. Particularly, the locational feature set is more important for Athens than for Amsterdam. This difference exists probably due to the fact that in Athens, there is a tendency of places to spatially cluster based on their type. In addition, it is worth mentioning that the spatial clustering of the places seems to be more dependent on the places' function than on the city. For instance, the clothing stores seem to cluster more than the rest of the POI types.

Distinguishing POI Types

The overall performance (F1-score) of the classifier when predicting the POI types included in set A , 60% for Amsterdam (73% accuracy) and 60% for Athens (64% accuracy), are quite high if one considers that the selected POI types are relatively similar by nature and even people could disagree on what should be considered as the "main function" of those places. For instance, a place which serves food, coffee and drinks could be labeled either as restaurant, cafe or bar from different people.

The obtained results for predicting POI types among 10 different types suggest that indeed the classifier managed to understand the features which distinguish the POI types at a satisfactory level. For this case, a feature is seems to be important if it is able to distinguish one class from all the rest. Thus, the most important features are quite "specialized" to a specified POI type and mostly belong to the functional and experiential feature sets.

Confusion Matrices - Distinguishing POI Types

The **confusion matrices** (figures 4.14 and 4.15) are computed to reveal which POI types could be considered as more "similar". For instance, for Amsterdam, "cafe" is constantly misclassified either as "restaurant" or as "bar". This misclassification is quite reasonable as for Amsterdam cafes, restaurants and bars are indeed very similar (e.g. in terms of opening times and provided services). It is however relatively unexpected the fact that cafes are not considered to be coffee shops and vice versa. A possible explanation could be that coffee shops have different closing times, since they are not open until as late as the rest and they could be distinguished based on this information. For Athens, cafes are again the most misclassified class and in this situation they are indeed mostly misclassified as coffee shops. Another important insight gained from the confusion matrices, as discussed in the previous chapter, is that the misclassifications are more equally distributed among the different POI types for Athens than for Amsterdam. A possible explanation for this is that for Amsterdam some POI types are indeed very similar and hard to be distinguished even for humans, while for Athens they tend to differ more. For example, cafe and bars for Amsterdam often provide very similar services and have similar opening and popular times. For Athens, however, those two POI types differ quite a lot.

Features Importance - Distinguishing POI Types

Focusing on each **individual feature**, the ones which were computed as the **most important** (figures 4.16, 4.17) support the results of the importance of each feature set. Almost all of the 15 most important features for both cities and cases are included in the respective two most important feature sets. Additionally, by interpreting the topics which were extracted from the reviews and ended up being among the top indicators of the POIs type (table 4.5), one could understand why they are so valuable as they are almost describing specific POI types (e.g. "restaurant" or "hotel"). Interestingly enough, again for both cities, there are not features extracted from reviews written in greek or dutch among the most important features. This could be due to the small amount of reviews written in greek and dutch, relatively to the amount of the english reviews, or due to the software libraries used which are better in processing english text than greek and dutch.

A key insight is that the important features for distinguishing the POI types seem to be features which are "specialized" on specific POI types. For example, the extracted topic "Hotel" is presented as an important feature for both cities while in the most important features analysis it is suggested that this feature indeed distinguishes hotels from all the rest. The "Sunday Popular Times" feature also depicts as "different" some specific POIs (e.g. for Amsterdam the restaurants are the only places which are most popular during the evenings).

Moreover, the results from the least important features analysis supported that even the features which are not present among the most important features, contain relatively valuable information about the POI types. The average frequency of the most important visual features in the Google Street View images correlates reasonably with the respective POI types. For instance, the "shop front" and "arcade" scenes are more frequently present to images of clothing stores than of restaurants and hotels, exactly as one would expect. Additionally, the most important socio-topical features include topics extracted from the tweets which are also quite interpretable and specific to certain POI types (unsurprisingly less interpretable than the topics extracted from the Google reviews). Overall, the "least important features analysis" proved to be valuable as it depicted the importance of some features which are not present in the rest of the results and their value would have not been acknowledged.

Analysing POI types

When dealing with the binary problem of predicting if a POI is of a certain type or not, the most important features seem to be more diverse and tend to include the different dimensions of the target POI type. A clear example of this could be seen in figure 4.28 (for Amsterdam) in which features from every feature set are among the most important features for predicting if the type of a POI is "clothing store" or not. Those features include the "nearby clothing stores", the extracted from the Google reviews topic "store" and the extracted from Twitter topic "Drink/Cafe/Hotel". Thus, the features seem to be indeed focused on the different dimensions of clothing stores while in the previous case of distinguishing POIs, several experiential features could be seen, each one of them focused on a different POI type. In other words, since the experiential and functional features tend to express better the POI types, when dealing with multiple different types only those features are among the most important (the functional and experiential features of each type). However, when dealing with one POI type only the experiential and functional features relative to this type seem to be important, and then features from other feature sets follow.

5.2 Threats to validity

The threats to the validity of this research are the following:

- Amount of Data for Amsterdam and Athens

  The amount of POIs collected from Amsterdam is higher than the amount of POIs collected for Athens and this leads in having more training data for Amsterdam than for Athens. This fact, could have slightly biased the results (e.g. the classifier might perform better for Amsterdam due to the amount of the training data). Nevertheless, the differences in the classification results between the two cities are not that large, apart from when using the locational feature set in which case the results are better for Athens. In addition, the amount of POIs could be also considered as a characteristic or a particularity of each city.

- Used Models

  The quality of the extracted topics from the unstructured text and the image-annotations of the street-level images is inextricably linked with the quality of the used techniques and models. The models used are the currently state-of-the-art ones. However, it is possible that in the next years other models will be developed which are going to be superior than the ones used. In this case, the quality of the extracted features is going to be improved and the results based on the new models might differ. For instance, the reason that the visual features' contribution to the prediction of POI types is low, could be because the current algorithms are not able to extract features which are important for the success of this kind of prediction and not because the exterior of a place does not include enough information for predicting the places main function.

- Used Classifier

  The focus of this thesis was given on discovering which are the POI features that make a POI type distinguishable from others or in other words what are the qualities that "make" each POI type and not on achieving the highest possible performance in predicting the POI types. Thus, even though during the selection of the classifier multiple classifiers and techniques were tried, it is possible that even higher performance could have been obtained if more focus was given on the tuning of the classifiers and on trying different ensembles. Consequently, even if the presented results are valid, it is possible that even higher results could have been obtained.

Chapter 6

Conclusions

This chapter consists of two sections: the *Conclusion* in which the main research question of this thesis is being answered and the *Future work* in which possible ways to continue and extent this work are being discussed.

6.1 Conclusion

The answer of the main research question of this thesis, "*How to combine and extract multidimensional POI features from various web sources to classify urban place types?*", is provided by answering the four research sub-questions.

RQ1: *What are the current state of the art methods for the extraction of POI features from place-related data?*

Regarding the first research question, a literature survey is conducted and divided in three main parts. The first includes studies which focus on proposing methods on how POIs could be used towards gaining a deeper understanding of urban environments. The second and third parts include studies which focus on studying places through the extraction of POI features from a single and multiple data sources, respectively. The studies included in the first part, depict the importance of the POIs as well as the large currently existing amount of POI data, while the other two parts include various state of the art techniques which deal with the extraction of POI features from different kinds of data. For instance, the LDA topic model is found as one of the most used and suggested techniques for extracting topics from text and the CNN models are found promising for extracting information from images.

In addition, several state of the art matching approaches are discussed which, despite being different, they share some fundamental and useful conclusions such as that the geographical distance between two POIs, might be misleading as to if the POIs should be matched. Moreover, the analysis of the related work shows the connection between extracted features and data sources, meaning which sources should be used for the extraction of specific features. For example, Twitter is often used for the extraction of semantic features, such as topics discussed in a place, while Foursquare is used for extracting features of actual place entities.

Overall, the need of combining different sources towards the better digitization of physical places is emphasized from the relevant studies as well as the need of using different methods and techniques to extract POI features coming from data of different nature (e.g. images and text). Even if the relevant studies depict some of the advantages and disadvantages of various data sources and extracted features, the comparison of those studies is hard since different data, collected from different cities and time-slots are used. For that reason, in this thesis an effort has been made to extract and combine data of different nature in order to compare their usefulness towards the deeper understanding of places.

RQ2: *How to combine POI features from various web data sources?*

The combination of POI features from various sources involves the selection process of the data sources to be used and the method to link the data with the corresponding POI entities. The data sources could directly or indirectly linked to the POI entities and for each case a different approach is followed.

Particularly, to combine data which are **directly** linked to the POIs a "matching" process is suggested in which the different features of the POIs belonging in the different sources (e.g. name, location, function) are compared in order to discover the POIs that represent the same physical place.

On the other hand, for the sources which include place-related information but are **not linked directly** to any POIs (e.g. Twitter, and Google Street View) the combination is realized using the geolocation (e.g. collect tweets which are geolocated within a specified distance from a place) for which, to be as precise as possible, it is suggested to combine sources which use the same geolocation system (e.g. Twitter and Foursquare or Google and Google Street View).

For the two use cases of this work the used sources are: Google, Google Street View, Foursquare and Twitter. The combination of those four sources is being realized in the following way: for each Google POI the street-level images geolocated in the same location are being retrieved from Google Street View (Google and GSV use the same geolocation system) and for each Foursquare POI tweets sent in a distance of 50m are retrieved from Twitter (Foursquare and Twitter use the same geolocation system). For the combination of all four sources, Google POIs and Foursquare POIs are matched based on the matching algorithm depicted in 3.2. In that way, POIs which include data from all the four sources are created.

This combination seem to be essential so that the collected data for each POI are able to reflect, directly or indirectly, the high dimensional nature of places.

RQ3: *How to extract multidimensional POI features from combined data sets?*

Regarding the third question, the multidimensional POI features to be extracted are firstly defined and then the techniques used for this extraction are analysed. To define which multidimensional features could be considered valuable for better representing a place, focus is given on the term "sense of place" which comes from various theoretical studies that aim to explain what "place" truly is.

The selected extracted features are divided in five feature sets: functional, socio-topical, experiential, visual and locational. Several of features belonging to those sets are extracted directly from the POI data while others required more specialized techniques. Particularly, when dealing with unstructured text such as tweets and Google reviews, the **LDA** model proved to be very useful for extracting interpretable topics.

When dealing with images two state-of-the-art and openly shared **CNN models**, one for scene detection and another one for object recognition, are used for the extraction of information from the street-level images. Since the data sets are already combined, the extracted POI features are also combined and as a result each POI is eventually digitally described by those features.

RQ4: *Which POI features contribute the most to the classification of urban place types?*

The answer to the fourth research question is based on the implementation of the pipeline depicted in figure 3.1. The data collection, the POI features extraction and the POI types prediction is realized using data coming from two cities, Amsterdam and Athens. The POI types prediction is divided into two main cases: (1) predicting the type of the POIs among a set of types [1] (multiclass problem) and (2) predicting if a POI is of a certain type [2] (binary problem). By training a machine learning algorithm using different combinations of POI features, the features' contribution to the classification of POI types, in both cases, is quantified. Therefore, the contributions of the various features are compared and some light is shed on the question "what makes a place" for the selected ten POI types .

For **both cases and cities**, the results suggest that the features that contribute the most to the classification of urban place types are the ones relative to the functional(e.g. opening/closing times) and the experiential (e.g. topics extracted from the reviews) characteristics of the places. This is an indication that the importance of these features could be context-agnostic, meaning that it would be equally important for the POIs of other cities and other POI types as well. Of course, since this is a strong statement, further analysis should be conducted to ground this indication. Moreover, the comparison of the two cities reveals that the reason of the features to be important is influenced by the city's characteristics. A good example of that, is the features included in the locational feature set which tend to be more important for Athens probably due to the fact that in Athens the places are indeed more clustered by type.

Overall, the need of combining data sources, which is emphasized throughout this thesis, is supported from the obtained results. Particularly, the functional, experiential and locational features seem to contribute the most to the classification of urban place types. The least important feature sets are the socio-topical and the visual for which the results of the least important features analysis suggested that they also include valuable information about the POI types.

6.2 Future work

This study consists of several different parts which could be further extended or improved. Firstly, more cities could be taken into account so that the extent to which the obtained results are city-specific could be further studied. The inclusion of more and more diverse cities, such as cities belonging in different continents, could provide stronger support of the existing results.

[1] Most of the focus is given on: restaurant, clothing store, bar, hotel, food and drink shop, cafe, gym,coffee shop, college and university and art gallery

[2] This analysis is realized for clothing stores, hotels and restaurants

Furthermore, more focus could be given on the matching algorithm so that a more sophisticated solution is provided. This could be accomplished by the creation of ground truth. The creation of a relatively large dataset which includes the matching information for two or more POI data sources would be very beneficial towards this goal. A dataset as such could be for example created through crowdsourcing.

In addition, it would be insightful to include interior images of the studied places. However, it is important that those images are consistent among the different places and objective. This is challenging as the majority of the place-related images which exist in the POI data sources are included either for marketing purposes or because users have uploaded them. In both cases, the images could be extremely biased and not consistent and as a result the extracted features might express other characteristics than the visual appearance of each place. Those features, might improve the accuracy in the prediction of POI types but the places dimensions which those features express might be different than the ones studied in this work. For instance, an important question to be asked is "Is a place considered to be a bar due to how it is being visually advertised?". Thus, to include images of the interior of places one has to be conscious about what dimensions of the places she wants to study and if the selected images reflect those dimensions.

Lastly, it is very intriguing to study places not only according to their type but also based on other qualities. It is very possible that the contribution of the extracted features would change if instead of distinguishing, for example, bars from restaurants one tried to find which are the "similar" bars or restaurants and even more importantly what makes them similar.

Bibliography

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Albert Acedo, Marco Painho, and Sven Casteleyn. Place and city: Operationalizing sense of place and social capital in the urban context. *Transactions in GIS*, 21(3):503–520, 2017.

[3] Benjamin Adams and Krzysztof Janowicz. Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science*, 29(4):556–579, 2015.

[4] Pragya Agarwal. Operationalising 'sense of place' as a cognitive operator for semantics in place-based ontologies. In *International Conference on Spatial Information Theory*, pages 96–114. Springer, 2005.

[5] Gideon DPA Aschwanden, Simon Haegler, Frédéric Bosché, Luc Van Gool, and Gerhard Schmitt. Empiric design evaluation in urban planning. *Automation in construction*, 20(3):299–310, 2011.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Antonio Torralba, Aude Oliva. Places website, 2017.

[8] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] Anne Cocos and Chris Callison-Burch. The language of place: Semantic value from geospatial context. In *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, volume 2, pages 99–104, 2017.

[11] Jennifer E Cross. What is sense of place? Colorado State University. Libraries, 2001.

[12] Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[13] J Nicholas Entrikin. The betweenness of place. In *The Betweenness of Place*, pages 6–26. Springer, 1991.

[14] Cheng Fu, Grant McKenzie, Vanessa Frias-Martinez, and Kathleen Stewart. Identifying spatiotemporal urban activities through linguistic signatures. *Computers, Environment and Urban Systems*, 2018.

[15] Song Gao, Krzysztof Janowicz, Grant McKenzie, and Linna Li. Towards platial joins and buffers in place-based gis. In *COMP@ SIGSPATIAL*, pages 42–49, 2013.

[16] Song Gao, Krzysztof Janowicz, Daniel R Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 31(6):1245–1271, 2017.

[17] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM, 2010.

[18] Lingzi Hong, Weiwei Yang, Philip Resnik, and Vanessa Frias-Martinez. Uncovering topic dynamics of social media and news: the case of ferguson. In *International Conference on Social Informatics*, pages 240–256. Springer, 2016.

[19] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, volume 4, 2017.

[20] John Brinckerhoff Jackson. *A sense of place, a sense of time*. Yale University Press, 1994.

[21] Krzysztof Janowicz, Grant McKenzie, Yingjie Hu, Rui Zhu, and Song Gao. Using semantic signatures for social sensing in urban environments. 2018.

[22] Andrew Jenkins, Arie Croitoru, Andrew T Crooks, and Anthony Stefanidis. Crowdsourcing a collective sense of place. *PloS one*, 11(4):e0152932, 2016.

[23] Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira Jr, and Francisco C Pereira. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46, 2015.

[24] Gunila Jive´ n and Peter J Larkham. Sense of place, authenticity and character: a commentary. *Journal of urban design*, 8(1):67–81, 2003.

[25] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[26] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *null*, pages 2169–2178. IEEE, 2006.

[27] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[28] Lin Li, Xiaoyu Xing, Hui Xia, and Xiaoying Huang. Entropy-weighted instance matching between different sourcing points of interest. *Entropy*, 18(2):45, 2016.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] Kevin Lynch. *The image of the city*, volume 11. MIT press, 1960.

[31] Alan M MacEachren. Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial Data Handling in Big Data Era*, pages 139–155. Springer, 2017.

[32] CD Manning, R PRABHAKAR, and S HINRICH. Introduction to information retrieval, volume 1 cambridge university press. *Cambridge, UK*, 2008.

[33] Grant McKenzie and Benjamin Adams. A data-driven approach to exploring similarities of tourist attractions through online reviews. *Journal of Location Based Services*, pages 1–25, 2018.

[34] Grant McKenzie and BT Adams. Juxtaposing thematic regions derived from spatial and platial user-generated content. 2017.

[35] Grant McKenzie and Krzysztof Janowicz. Openpoi: An open place of interest platform. 2018.

[36] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. Weighted multi-attribute matching of user-generated points of interest. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 440–443. ACM, 2013.

[37] Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015.

[38] Grant McKenzie, Krzysztof Janowicz, Song Gao, Jiue-An Yang, and Yingjie Hu. Poi pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2):71–85, 2015.

[39] Helena Merschdorf and Thomas Blaschke. Revisiting the role of place in geographic information science. *ISPRS International Journal of Geo-Information*, 7(9):364, 2018.

[40] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[41] Daniel R Montello, Alinda Friedman, and Daniel W Phillips. Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28(9):1802–1820, 2014.

[42] Christoph Mülligann, Krzysztof Janowicz, Mao Ye, and Wang-Chien Lee. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *International Conference on Spatial Information Theory*, pages 350–370. Springer, 2011.

[43] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[44] Lawrence Philips. The double metaphone search algorithm. *C/C++ users journal*, 18(6):38–43, 2000.

[45] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[46] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009.

[47] John W Ratcliff and David E Metzener. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46, 1988.

[48] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1):4, 2010.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[50] Richard C Stedman. Is it really just a social construction?: The contribution of the physical environment to sense of place. *Society &Natural Resources*, 16(8):671–685, 2003.

[51] Fritz Steele. *The sense of place*. Cbi Pub Co, 1981.

[52] Sabine Storandt and Stefan Funke. Automatic improvement of point-of-interest tags for openstreetmap data. In *Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings*, volume 15, page 56, 2015.

[53] Daniel Sui and Michael Goodchild. The convergence of gis and social media: challenges for giscience. *International Journal of Geographical Information Science*, 25(11):1737–1748, 2011.

[54] Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

[55] Yi-Fu Tuan. Space and place: humanistic perspective. In *Philosophy in geography*, pages 387–427. Springer, 1979.

[56] Gustav von Zitzewitz. Survey of neural networks in autonomous driving. 2017.

[57] Wikipedia contributors. List of virtual communities with more than 1 million users — Wikipedia, the free encyclopedia, 2018. [Online; accessed 13-October-2018].

[58] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[59] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 35. ACM, 2017.

[60] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu. xnet+ sc: Classifying places based on images by incorporating spatial contexts. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 114. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[61] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[63] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

Appendix A

Glossary

In this appendix we give an overview of frequently used terms and abbreviations.

POI: Acronym which stands for "Point Of Interest". POIs are the digital representations of places.

POI Types and extracted features: The POI types and the extracted features are used as proxies of real-world-places.

Matching: Identifications of POIs which belong to different sources and represent the same physical space.

Functional feature set: Functional characteristics of places such as instance, opening/closing times, if they have website or popular time-slots.

Experiential feature set: Features about how people experience places such as likes, ratings and feetures extracted from the Google reviews.

Socio-topical feature set: Features gathered from Twitter which aim to express social aspects of a place such as what topics are discussed or the sentiment of the people around it .

Visual feature set: Features relevant to the visual appearance of the exterior of places.

Locational feature set: Features relevant to the location of a place such as what is the type of the nearby places within several radius.

Set A (classification): art gallery, bar, cafe, clothing store, coffee shop, college and university, food and drink shop, gym, hotel and restaurant.

Set B (classification): shop and service, food, travel and transport, arts and entertainment,nightlife spot and college and university.

Appendix B

Figures

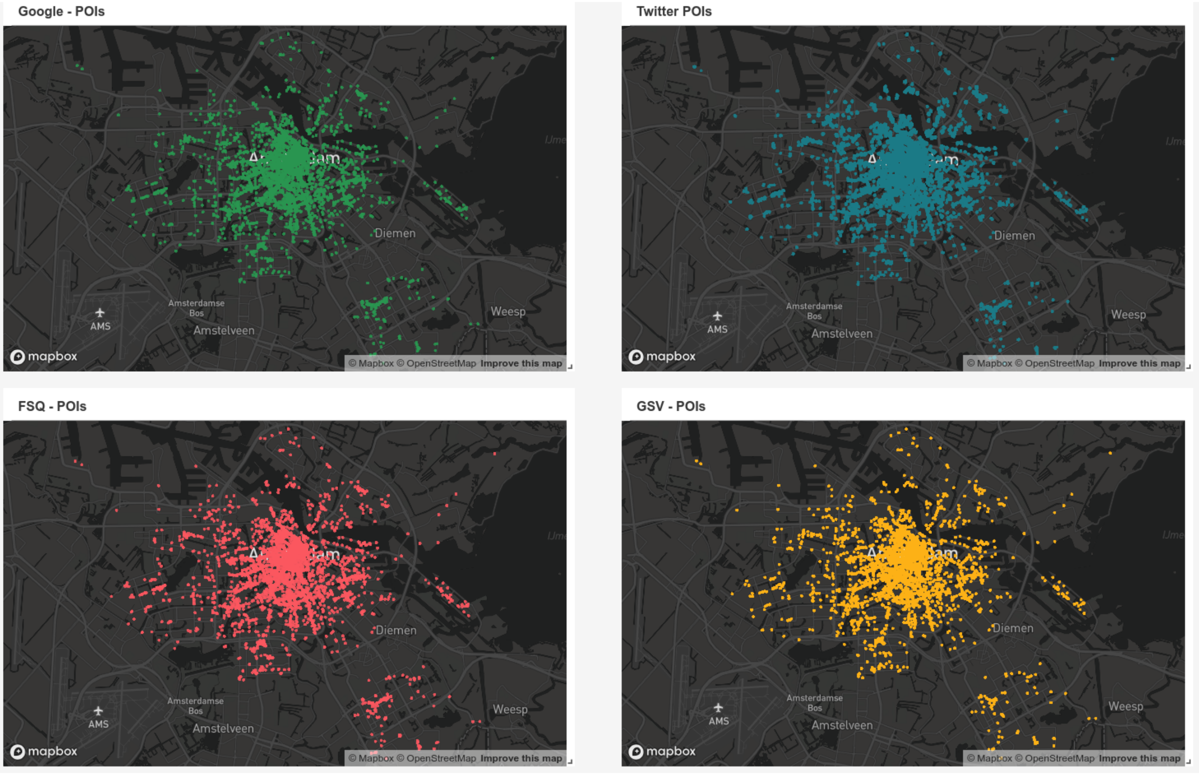


Figure B.1: Example of the data gathered for Amsterdam. Obviously, the collected data from each source seem to be almost identical as their geolocation is almost the same.

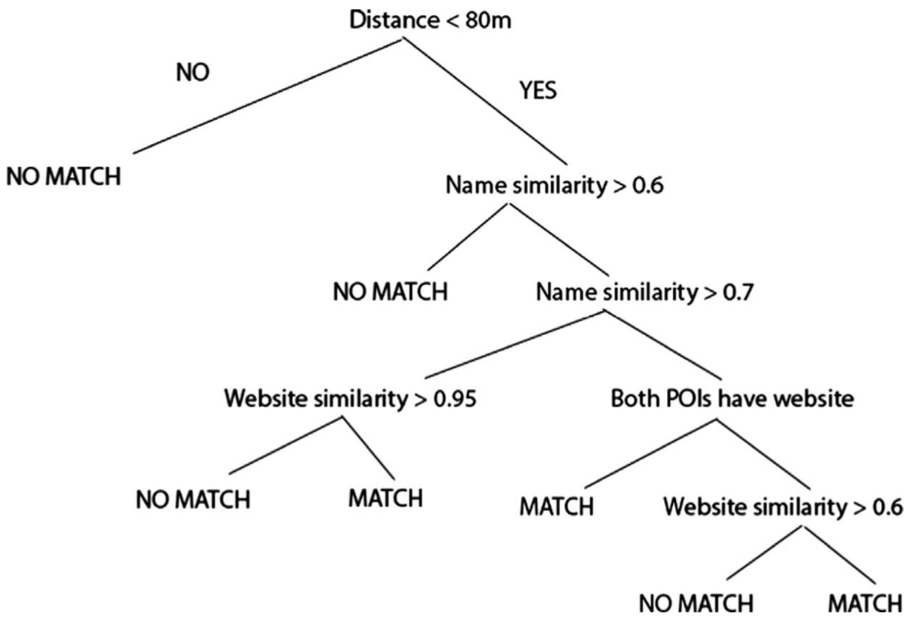


Figure B.2: POI Matching Algorithm from [23]

Appendix C

Source Code Repository

The code could be found in the following Github repository:
https://github.com/MiliasV/poi-feature-mining.