BIG DATA ANALYTICS AND KNOWLEDGE DISCOVERY THROUGH LOCATION-BASED SOCIAL NETWORKS (LBSN)

by

IOANNIS MAKRIDIS

BSc. Technological Educational Institute of Crete, 2015

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF INFORMATICS ENGINEERING

SCHOOL OF ENGINEERING

TECHNOLOGICAL EDUCATIONAL INSTITUTE OF CRETE

Approved by:
Professor Ioannis Kopanakis

Heraklion, Crete
2018

Copyright

Abstract

Current research argues that the use of social networks can be a dominant resource for acquiring valuable knowledge about tourist destinations through the collection of data from Location-Based Social Networks (LBSN). This approach can provide solutions to problems and significant benefits to enterprises and local authorities. One of the most important problems is the lack of knowledge about the visitor's opinion for a destination, as well as the fact that the visitors' behavior, needs and preferences are not visible.

Nowadays, enterprises and local authorities are still using the traditional method for acquiring valuable knowledge to make strategic decisions, using data from surveys conducted using questionnaires. In this process, despite the benefits it can offer, it can be observed that it is not continuous, it is not adaptable, and the number of the participants in most of the cases is quite small compared to the total number of visitors.

This thesis presents the methodology for extraction, association, analysis and visualization of data derived from Location-Based Social Networks, providing valuable knowledge about visitors' behavior, impressions and preferences for tourist destinations. In that context, using a case study of the region of Crete (Greece) as tourist destination the overall insights of visitors are identified.

Utilizing state of the art web technologies for the development of this monitoring framework, we will collect publicly available social data for the selected tourism destinations. Some of the data that need to be collected for the process of analysis are the visitors' posts and reviews, publication's date and time, published photos, places' rankings, user's profile, nationality and engagement. After the completion of data collection, the correlation and analysis are followed, presenting data visualization in a proper and efficient way to the stakeholders.

Having acquired this valuable knowledge about the visitors' behavior and preferences, the stakeholders such as local authorities, enterprises etc. can make a more efficient promotion of tourist destinations based on the needs of visitors, improvement of the existing facilities and creation of new experiences attracting the interest of more potential visitors.

Keywords: big data, tourism, data analytics, social media analytics, location intelligence, location-based social networks (LBSN), data-driven decision making

Περίληψη

Τρέχουσες έρευνες υποστηρίζουν ότι η χρήση των κοινωνικών δικτύων μπορεί να αποτελέσει κυρίαρχο μέσο για την απόκτηση πολύτιμης γνώσης σχετικά με τους τουριστικούς προορισμούς μέσω της συλλογής δεδομένων από τα κοινωνικά δίκτυα που περιλαμβάνουν τοποθεσία (Location-Based Social Networks - LBSN). Αυτή η προσέγγιση μπορεί να προσφέρει λύσεις σε προβλήματα και σημαντικά οφέλη για τις επιχειρήσεις και τους φορείς του δημοσίου. Ένα από τα σημαντικότερα προβλήματα είναι η έλλειψη γνώσης σχετικά με τη γνώμη και τις εντυπώσεις του επισκέπτη για έναν προορισμό, καθώς και το γεγονός ότι η συμπεριφορά, οι ανάγκες και οι προτιμήσεις των επισκεπτών δεν είναι ορατές.

Στις μέρες μας, οι επιχειρήσεις και οι τοπικές αρχές εξακολουθούν να χρησιμοποιούν την παραδοσιακή μέθοδο για την απόκτηση γνώσης ως προς τη λήψη στρατηγικών αποφάσεων χρησιμοποιώντας δεδομένα από έρευνες που διεξάγονται με τη χρήση ερωτηματολογίων. Σε αυτή τη διαδικασία, παρά τα οφέλη που μπορεί να προσφέρει, μπορεί να παρατηρηθεί ότι δεν είναι συνεχής, δεν είναι προσαρμόσιμη και ο αριθμός των συμμετεχόντων στις περισσότερες περιπτώσεις είναι αρκετά μικρός σε σύγκριση με τον συνολικό αριθμό των επισκεπτών.

Η παρούσα εργασία παρουσιάζει τη μεθοδολογία άντλησης, συσχέτισης, ανάλυσης και απεικόνισης δεδομένων που προέρχονται από κοινωνικά δίκτυα που βασίζονται στην τοποθεσία, παρέχοντας πολύτιμες γνώσεις σχετικά με τη συμπεριφορά, τις εντυπώσεις και τις προτιμήσεις των επισκεπτών για τους τουριστικούς προορισμούς σε πραγματικό χρόνο. Στο πλαίσιο αυτό, χρησιμοποιώντας ως μελέτη περίπτωσης την περιφέρεια της Κρήτης ως τουριστικό προορισμό, εξάγονται χρήσιμες γνώσεις σχετικά με τις δυο μεγαλύτερες πόλεις του νησιού, το Ηράκλειο και τα Χανιά.

Χρησιμοποιώντας σύγχρονες τεχνολογίες διαδικτύου για την ανάπτυξη αυτού του πλαισίου παρακολούθησης, συλλέξαμε δημοσίως διαθέσιμα δεδομένα από τέσσερα κοινωνικά δίκτυα για τους επιλεγμένους τουριστικούς προορισμούς. Μερικά από τα στοιχεία που πρέπει να συλλεχθούν για τη διαδικασία ανάλυσης είναι οι αναρτήσεις και τα σχόλια των επισκεπτών, η ημερομηνία και η ώρα των δημοσιεύσεων, οι δημοσιευμένες φωτογραφίες, η βαθμολογία των επιχειρήσεων και σημείων ενδιαφέροντος, το προφίλ των χρηστών, η εθνικότητα και η διάδρασή τους. Μετά την ολοκλήρωση της συλλογής δεδομένων, ακολουθείται η συσχέτιση και η ανάλυση τους και εν συνεχεία τα αποτελέσματα οπτικοποιούνται μέσω γραφημάτων στους ενδιαφερόμενους.

Έχοντας αποκτήσει αυτή την πολύτιμη γνώση σχετικά με τη συμπεριφορά και τις προτιμήσεις των επισκεπτών, οι ενδιαφερόμενοι όπως οι τοπικές αρχές, οι επιχειρήσεις κ.λπ. μπορούν να

προωθήσουν αποτελεσματικότερα τους τουριστικούς προορισμούς με βάση τις ανάγκες των επισκεπτών, να βελτιώσουν τις υφιστάμενες εγκαταστάσεις και υπηρεσίες και να δημιουργήσουν νέες εμπειρίες που θα προσελκύσουν το ενδιαφέρον περισσότερων πιθανών επισκεπτών.

Λέξεις-Κλειδιά: μεγάλα δεδομένα, ανάλυση δεδομένων, ανάλυση κοινωνικών δικτύων, location intelligence, location-based social networks (LBSN), λήψη αποφάσεων βάσει δεδομένων

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my appreciation to Dr. Ioannis Kopanakis, my thesis supervisor, for accepting me as his master student. He has definitely been a source of inspiration from the beginning and his contribution was very valuable in conceiving the idea of this dissertation and hence its implementation.

My thanks also go to my co-advisor Dr. George Mastorakis for his guidance and the advices that he offered. Also, I would like to thank my colleague Konstantinos Vassakis from the research laboratory e-Business Intelligence Lab for his willingness to provide me his knowledge in several steps of my dissertation.

Ioannis Makridis

2018

Dedication

My bottomless appreciativeness for her love and many years of supporting goes to my precious mother, Pelagia and to my beloved father George for his encouragement all these years. So countless thanks for being patient, always next to me and encouraging me during my entire journey in life. Also, I would like to express my gratitude to my brother Nick for supporting me in all my life.

Dedicated to the person who was the source of inspiration for me and left early from life
My beloved brother Konstantinos …

Chapter 1 - INTRODUCTION

The development of data science is comparable to the different stages of development for the internet in businesses. The Big Data era 1.0 was focused on establishing the means for data collection on an enormous scale. 'Building' or purchasing data warehouses and implementing the rudimental analysis skills and operational changes in companies was the first step of working with big data for many companies [1]. With Web 1.0, companies were able to build a presence online, introducing their products and services to a wider range of customers. Web 2.0 made interaction possible for businesses and customers alike. The online communication did not remain one-sided. Instead of only being able to inform themselves about products and services, (potential) customers were now able to communicate their issues, preferences, and experiences via websites, social media and online surveys. The widespread of social media turned the market online into a sort of conversation, where direct feedback was made possible [2].

A social media or online social network is a social structure that consists of individuals that are connected by one or more specific types of interdependency, such as friendship, common interests, and activities, and shared knowledge. In general, a social networking service (SNS) relies on and reflects the real-life social networks among people through the online web and mobile applications, allowing users to communicate and share ideas, activities, events, and interests over the internet. All the data generated by these 'conversations' constitutes the base for the era of Big Data 2.0.

The interaction between companies and clients gave an incentive for analysts to ask themselves what more they could do with the available information. By being able to collect and evaluate a user's movements online, from their social media presence to entries in forums or customer reference sites, to their online shopping behaviors and purchasing patterns; companies and their market analysts were now able to create a personalized profile for every customer, existing or potential. To which extent this is possible, and how well these generated profiles match the client's depends on the availability of data, the needed technology, and analytics knowledge, and the capable data scientists to develop methods of data analysis.

The development of the internet into Web 3.0 happened with the rise of mobile devices like smartphones, tablets and laptops. The potential that Big Data 3.0 has by adding location-based information to the profiles of customers is enormous. Making it possible to tailor marketing efforts not only to the preferences of people, but also to their physical position brings a variety of advantages to marketers [3].

The increasing availability of location-acquisition technologies (e.g. GPS, Wi-Fi, and 4G) allowing users to publish their content along with their location as location-tagged media content, such as photos, video, and texts making social networks geosocial or location-based social networks (LBSN). This is increasing the use of LBSNs by providing users the ability to express opinions, report events, share reviews and sentiment, and generating more and more user-generated content (UGC) while they are connected with others, which was unthinkable in the pre-Internet age [4].

A LBSN does not only mean adding a location as a new feature to an existing social structure so that people in the social network share information with a position tag but also consists of the new social structure made up of individuals linked by interdependence derived from their physical world locations [5]. In a LBSN, the location of a user is represented as a place, such as a street, shop, park, beach, point of interest, building etc. [6] which is tagged among a media content such as a post with text, picture, or video, in order to inform other users of the network about the location of the specific post. Examples of location-based social networks include Facebook, Foursquare, Google+, Instagram, Twitter, and Flickr. These networks have been characterized as major tools for communicating, spreading ideas, information and knowledge among their users. Each has its own peculiarities and features, but they all support the addition of a location in a post.

Recently, we have observed the explosive growth of LBSNs. For instance, the market leader Facebook was the first social network to surpass 2 billion registered accounts and currently sits at 2.23 billion monthly active users, while Instagram follows with 813 million active users, Tumblr with 794 million users and Twitter with more than 330 million users as of April 2018 [7]. This growth over the past years has resulted in the definition of the term Social Big Data which represents the enormous volume of data derived from online social networks and their hundreds of millions of monthly active users worldwide. Social big data usually are stored in two formats: the structured data includes users' personality and social attitude (number of posts, likes, comments, etc.), and the unstructured data is related to user-generated content ranging from websites and microblogs to rich content including both audio and visuals [8]. According to International Data Corporation (IDC) the global datasphere will grow from 16.1 zettabytes (ZB) generated in 2016 to 163 ZB by 2025, having as main factors for this increase the further growth of online social networking and the establishment of the Internet of Things (IoT) [9].

The emergence of big data and the progress in the field of data science have changed the way companies observe, analyze and influence the market. The principle of marketing is a discipline that combines many fields, such as statistics, psychology, economics, mathematics etc., is still true. But

whereas in the past the traditional market research tools served mainly to observe the market the way it was, using past data, big data analysis has added a tool that enables the prediction of market developments and building of complex models and customer profiles.

Social networks now affect digitally every aspect of social sciences and business research by changing the dynamics of the relationship between brands, employees, customers, and stakeholders. In that context, summarizing, we present a data analytics framework that follows the knowledge discovery in databases (KDD) process to acquire, transform, store, analyze, and visualize publicly available user-generated content from location-based social networks.

Our implementation consists of a cloud-hosted web application that uses modern cloud computing technologies for processing, storing, and analyzing data in real-time. Using geospatial data that was obtained through the communication with the web APIs (Application Programming Interfaces) of four popular location-based social networks including Twitter, Foursquare, Instagram, and Flickr, and stored into a data warehouse, we applied descriptive analytics to provide valuable knowledge about visitors' behavior, impressions, and preferences for the two most popular tourist destinations of the region of Crete (Greece).

1.1 Thesis outline

Chapter 1 gives a brief description of the fields that are involved in this thesis and our framework. The rest of the thesis is organized as follows. In Chapter 2 a review of the literature about the topics that are involved with this thesis including big data, data analytics, and location intelligence, as well as the related work about social media analytics are described in detail. Chapter 3 presents the proposed framework for acquiring, cleaning, storing, analyzing and visualizing data derived from location-based social networks in geographical regions of interest. Furthermore, the case study of the region of Crete (Greece) that is used to apply our framework and its insights are presented in Chapter 4. Finally, Chapter 5 presents the conclusions, findings, and directions for future work for further improvements and extensions of the proposed framework.

1.2 Motivation and research questions

The aim of this thesis is to design and develop a web application for discovering knowledge through actual user-generated content on location-based social networks and emphasizing the significant impact of "where" in business operations. Through the acquisition of geospatial data from four

popular LBSNs including Twitter, Foursquare, Instagram and Flickr we wanted to answer the following research questions:

- "What are the visitors' and local people's behavior, impressions, and preferences for tourist destinations?"
- "What decisions local authorities and businesses can take to make a more efficient promotion of the tourist destinations, to improve the existing facilities and activities, and to create new experiences for attracting the interest of more potential visitors?"

In order to answer these questions and find out what insights can be extracted from the analysis of LBSN data, we used the case study of two cities - Heraklion and Chania - from Crete (Greece).

Chapter 2 - THEORETICAL BACKGROUND

2.1 Big data

In 1997, Michael Cox and David Ellsworth were referred to the term “big data” which defined as the challenge of storing large datasets for visualization purposes [10]. Nowadays, with the large growth and fast development of web-based technologies, we are observing an exponential growth in the volume of datasets, known as the big data era. Big data is a process of collecting, managing and analyzing large amounts of data to generate knowledge and expose hidden patterns. In big data approaches, there are several challenges to be addressed including the collection, storage, and processing of data, as well as to gain valuable knowledge by analyzing and visualizing them in a proper way.

Exploiting the obtained data has always been considered by researchers and practitioners, but the large volume, diversity, and high velocity of real-time data streams are pushing the limits of the current processing, storage, and management capabilities. Using a set of storage and computational resources called cluster, we can analyze big data concerning their amount and speed, satisfying in that way the main principle in big data analytics, the need of high performance in computations.

Big data is one of the most representative examples of the “knowledge economy” and represents an emerging field of research for researchers and practitioners. It can provide organizations and businesses a huge and varied amount of data, as well as it is possible to provide invaluable insights about customers views, preferences, needs, attitudes etc. [11]. In addition, it is accepted that big data is a key source of value creation [12] since it is considered to result in more efficient and effective procedures by optimizing supply chain flows, setting the most profitable product price for products and services, selecting the right people for specific tasks and jobs, addresses issues, minimizing errors and quality problems, and improving customer relationships and experiences [12], [13].

2.1.1 The characteristics of big data

Big data can be characterized by the seven Vs.: *volume*, *variety*, *variability*, *velocity*, *veracity*, *visualization* which are based on large volumes of extensively varied data that are generated, captured, and processed at high velocity, and recognized as a key source for creating *value* [3], [14].

Volume refers to the large size of datasets generated through the fast development of technologies such as Internet of Things (IoT), Artificial Intelligence (AI), and user-generated content (UGC) through the use of online social media platforms. In the big data era, the volume of data that is processed is a lot bigger than traditional datasets. The generation of data is continuously increasing, and thus new algorithms need to be developed for reading, processing and analyzing them, as well as new infrastructures, need to be adopted for storing them. The storage of those big datasets has been made possible by using data warehouses which are repositories of integrated data from one or more different sources.

The **variety** of available data represents the diversity of data sources and formats. Nowadays, it is on a scale that has not been available before since the combination of customer data collected automatically by the company and the data the customers themselves create with their activities online surpasses the traditional data by great lengths.
Data can come in the form of images, text posts, log data, sound snippets and many more. The diversity of this information helps to gain a more complex insight into customers and their habits but also brings with it the problem of unstructured and hard to handle bundles of data. More specifically, the data types are:

- **Structured data** referred to data that is arranged in a definite pattern. One significant characteristic of data type is that all entities of the same category have the same attributes, length, and format, as well as it follows the same order, e.g. Relational Database Management Systems (RDBMS).
- **Semi-structured** in which different entities can have different structures with no predefined structure. This type of data can be supported by graph data structures and includes scientific data, bibliographic data, etc.
- **Unstructured data** referred to data that has no standard structure, e.g. videos, images, documents, etc.
- **Multi-structured data** is often an association of structured, semi-structured, and unstructured data, e.g. sensors data, operating system's log files, and customer interaction streams [15].

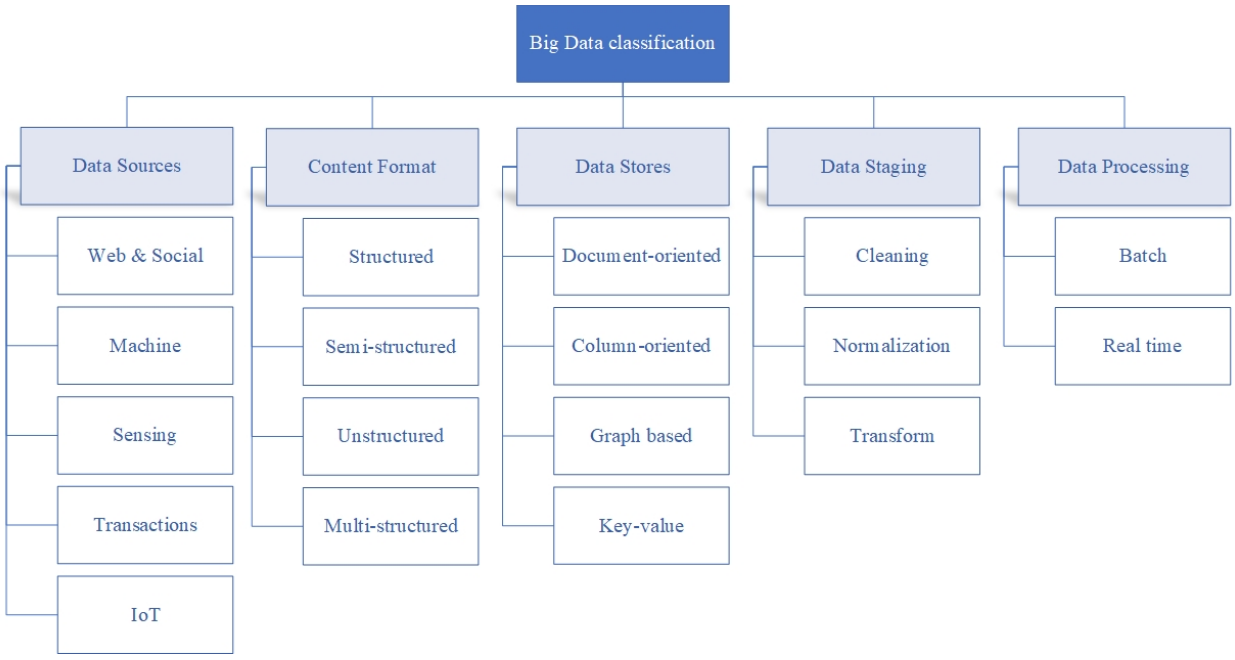


Figure 1 - Big Data classification [16]

Variability is related with the data whose meaning can vary significantly in context. It is often confused with variety; however, variability is related with fast change of meaning. For instance, words or phrases in a text can have a different meaning according to the context of the text, thus for accurate sentiment analysis, algorithms need to recognize the meaning of the phrase taking into account the whole context.

Velocity refers to the high speed of data generation since more and more devices are interconnected, resulting in generating enormous amounts of data in real-time. The velocity of big data makes it possible for the first time to work with real-time or near-real-time data, reacting immediately to customer's actions and needs. The throughput speed of data online has continuously increased. The speed of data is surpassing Moore's law and the volume of data generation introduced new measures for data storage i.e. exabytes, zettabytes and yottabytes. This means that now a direct reaction to a customer's action can take place, whereas previously data had to be collected and analyzed after transactions had taken place, delaying the response time.

Veracity refers to the data uncertainty and the level of reliability correlated with some type of data [3]. The veracity of the collected information can be an issue for anyone trying to come to conclusions after analyzing huge datasets with no accurate data. Often there is no guarantee that

the data set is complete, or how much of it is genuine. Verifying the integrity of datasets is an important step for finding valuable conclusions from evaluating them.

Data **Visualization** is the science of representing data in a manner that's easily readable, accessible and understandable with visuals or pictures. In contrast to other forms like text and tables, data visualization presents quantitative and qualitative information in a schematic form, in ways that trends, patterns, anomalies, etc., can be observed and understood.

Big data is recognized as a key source for creating **Value** since it is believed to result in more efficient and effective operations such as setting the most profitable price for products and services, optimizing supply chain flows, minimizing errors and quality problems, and improving customer relationships [14]. At first sight, it seems curious to put a monetary value on information, at least this is the thought that would have been dominant a few decades ago. But considering the big net worth of data-collecting corporations like Facebook or Amazon or Google, it should be clear that personal information has financial worth and therefore, data is valuable. In addition, businesses that overcome challenges and exploit big data efficiently have more precise information and are able to discover knowledge by which they can improve their strategy and operations regarding well-defined targets like productivity, financial performance, and market value, while big data is considered a significant factor in the digital transformation of enterprises introducing innovations [3].

2.1.2 Benefits from big data

In recent years, the topic of big data has created a huge hype and nowadays is more popular than ever. The most popular and highest worthed enterprises like Google, Facebook, and Amazon stressing that the integration of big data analytics into decision-making process can provide significant benefits. There is a widespread belief among executives, managers, and analysts that big data can provide value. [17] state that the capability to swiftly condense, analyze and distribute crucial information to the main decision-makers is the first stage of an efficient data-driven culture. Building this data-driven first stage is the most important to enhance performance and propel the organization forward. In addition to that developing and improving previously mentioned capabilities can empower enterprises to introduce improvements across all the departments of enterprises and lead to more attractive returns on investments.

[18] describe several benefits coming from big data, generally due to big data enterprises can make better predictions and smarter decisions based on evidence (data) instead of intuition and also target more effective interventions. The result is gaining a competitive advantage against rivals. In addition to that, they highlight that enterprises were born digital such as Google are already familiar with big data and exploit its power. [19] similar to the above state that the best performing retailers utilize big data not only for financial and operational purposes but also for strengthening their competitive advantage in any field such as marketing, customer care, and customer experience management. [20] summarized some ways that big data is creating value by the following: partitioning population groups in order to customize operations, allowing experimenting to spot needs, expose changeability and enhance performance, creating new business models, products and services.

All these result in competitive advantage and growth for enterprises, because of reducing prices due to transparency and better matching between consumer needs and product offerings. [18] found that data-driven enterprises perform better in finance and operations than their competitors. More specifically, data-driven enterprises present to be 5% more productive and 6% more profitable than their rivals.

Big data is much more than a buzzword. Big data is not about volume of data. Its analysis helps enterprises to better understand business environments, customers' needs and behavior and competitors' activities. Thanks to big data enterprises can customize their products exactly to their customer needs. This results in a competitive advantage for enterprises that leverage big data as well as customer surplus.

2.3 Big data analytics

Data analytics is a field which combining data science, business intelligence (BI) and business analytics for discovering, interpreting and communicating meaningful patterns in data [3]. The term data analytics became popular in the early 2000s and defined as "*the application of computer systems to the analysis of large datasets for the support of decisions*" [21]. It is a very multidisciplinary field that has followed aspects from many other scientific disciplines including statistics, operational research, pattern recognition, computational intelligence and machine learning.

In their majority, data analysis projects are following the traditional methodology (Figure 2) which consists of several phases including *preparation*, where data is assessed and selected, *preprocessing* where data are cleaned, transformed and filtered, *analysis* in which the data are classified, correlated, clustered and visualized, and finally the *postprocessing* phase which is related to the interpretation and evaluation of the results [21].

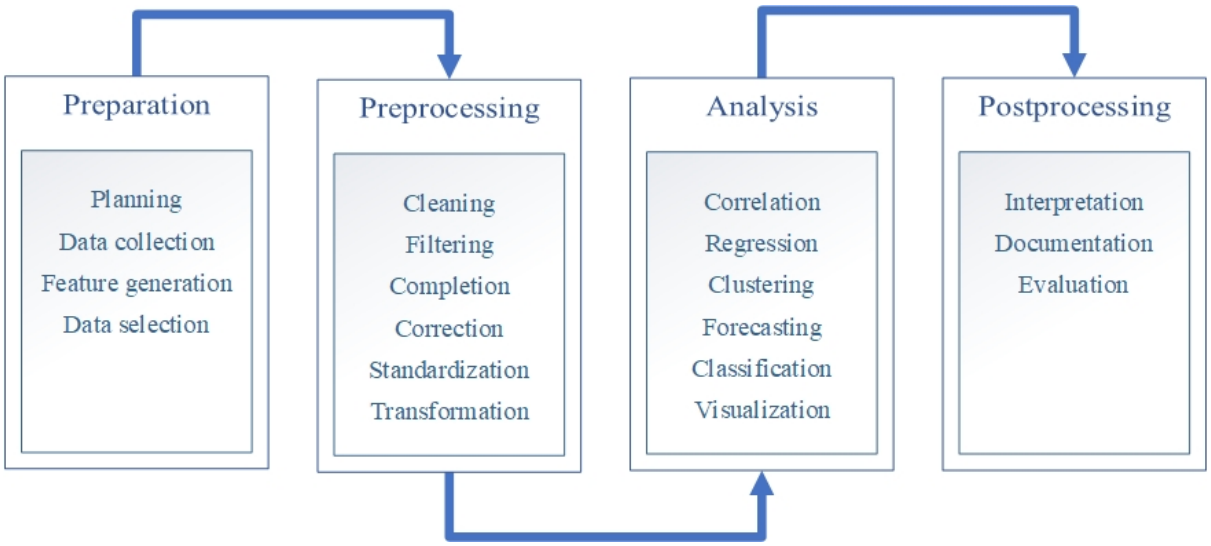


Figure 2 - Phases of data analysis projects

Nowadays, data is seen as a raw material with high value. Converting this raw material to information becomes a significant factor influencing the quality of services and products, as well as the decision-making process. In the big data era, businesses are performing data analytics over various types of business problems.

The most important questions that need to be answered in any data analytics project are (1) the problem the analysis is trying to solve and (2) which solution would create the higher impact. After the identification of the problem, data needs to be collected from various data sources which will be useful for the analysis. The attributes of data on these datasets can be defined according to the problem definition. Before providing data to algorithms and tools, a pre-processing should be performed to translate them into a fixed format. Then, data analytics can be performed having as main purpose the presentation of results with data visualizations to extract meaningful information [15], [22].

In contrast to the traditional analytics tools which are dealing with the existing and historical data and performing exploratory analysis, the advanced analytics techniques can deal with more

complex problems by extracting knowledge at a deeper level, as well as they can provide the capability to predict the future actions and reveal hidden patterns. Advanced data analytics can be accomplished through utilizing dedicated mathematical or statistical methods and using several algorithmic concepts including data mining, pattern association, classification, regression, forecasting, clustering and neural networks [23]. These methods and techniques provide the capability of processing both structured and unstructured data and producing descriptive, predictive, and prescriptive actionable results [24].

2.3.1 Data analytics types

In general, the goal of data analytics is to provide the organization with actionable insights for making strategic decisions in a smarter way and having better business outcomes. Since data analytics can provide different significant insights, there are different types of analytics that can be used according to the knowledge they provide. The three principal analytics types are: *descriptive* (what happened), *predictive* (what will happen) and *prescriptive* (what should I do) [25], [26].

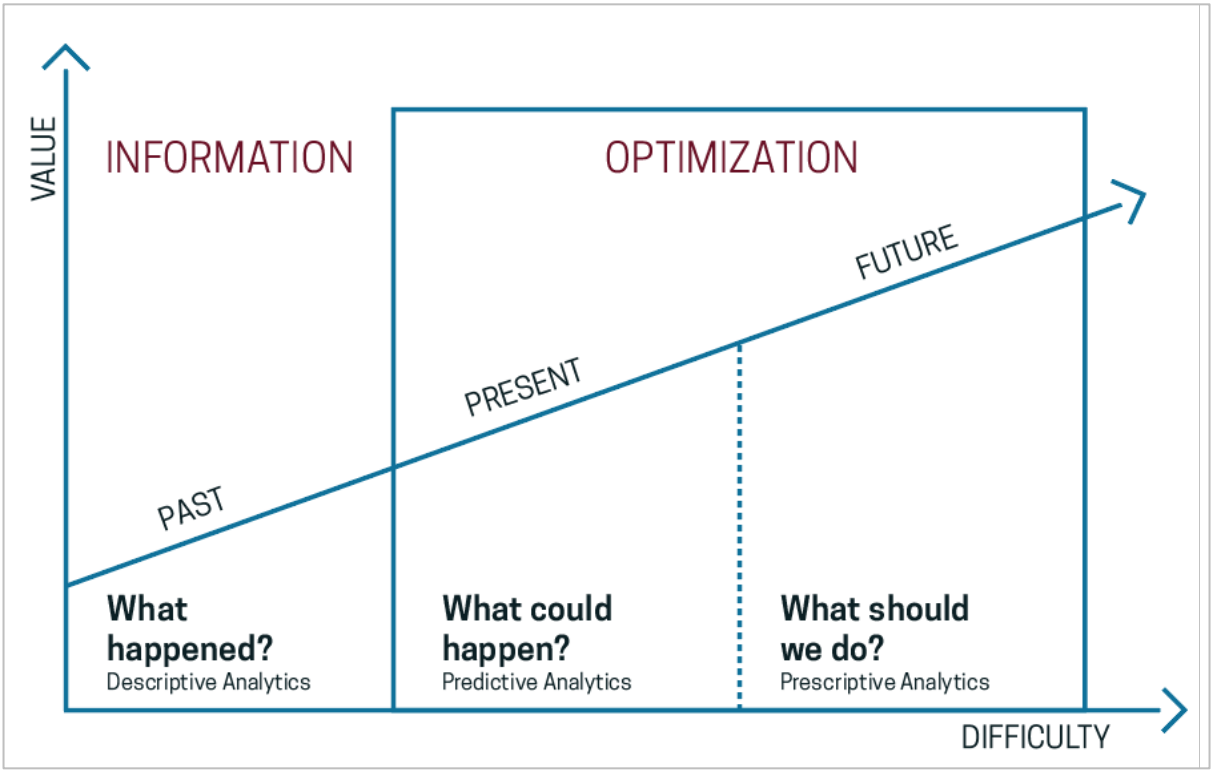


Figure 3 - The three principal data analytics types [28]

2.3.1.1 Descriptive analytics

Descriptive analytics is based on current and historical data to provide significant insights. Using techniques like online analytical processing (OLAP), probability analysis, trending and association of data that is already classified and categorized, descriptive analytics answers to what's happening in the organizations. Common examples of descriptive analytics include interactive dashboards, alerts, reports and data visualizations presenting key metrics of businesses sales, orders, marketing, supply chain, support, customers and financial performance [3], [15].

2.3.1.2 Predictive analytics

Predictive analytics is about making predictions about the future outcome. It provides information on what will happen in the future, while it also predicts the time frame for when it might take place. These analytics use historical and current data, which are analyzed and are found out the relationships. Next step is about extrapolating these relationships and using hypotheses into the future, which is of course limited by technologies and chosen techniques. Common techniques used in predictive analytics including data mining, neural networks (NNs), support vector machines (SVMs), decision trees, linear and logistic regression and clustering.

We can apply this analytic type to all the disciplines, from the prediction of failure in a production line, to predicting customer behavior based on what they are adding to their cart or buying. The benefit of this analysis is that reveals hidden patterns and detects relationships in the data. Using that type of analytics, it is possible for organizations to imagine, plan and make decisions about future operations while at the same time they can predict and manage their risks [3], [15], [27].

2.3.1.3 Prescriptive analytics

Prescriptive analytics is based on two previous analyses (descriptive and predictive) on complex data and suggests actions that could be taken. More specifically, it provides a prediction of the impact of future actions before they are taken, answering "what should we do" as the outcome of the organization's actions. Although the final decision is up to the managers, prescriptive analytics can provide a credible path to an optimal solution for business need or resolution of operational problems. Very useful is also a view for the different goals of an organization, which can be given as two extremes - to minimize or to maximize something. This analysis can give to give a picture of many questions for example about production, customers or profit. In fact, it is possible to find

the best choice among different options and make suggestions on how to continue reaching the wanted goal [3], [26].

2.3.2 Data analytics techniques

There are several techniques being used to analyze big datasets. Gartner lists several advanced techniques including among others data/text mining, sentiment analysis, semantic analysis, clustering, pattern matching, forecasting, and data visualization [23]. Nowadays, researchers continue to develop new techniques and improve the existing ones, since the need for analyzing new combinations of data is identified, thus we can observe that more than 25 techniques were developed [29].

In this chapter, we will focus on six of them including *text mining*, *clustering*, *visualization*, and *sentiment*, *social network*, and *spatial analysis*, since they used in the context of this thesis.

2.3.2.1 Text mining

Text analytics or text mining is a process that analyzes unstructured text, extracting relevant information, and transforming it into useful business intelligence. Examples of textual data that are analyzed in this process including social network feeds, emails, blog posts, news feeds, survey responses etc. Using techniques involved in computational linguistics, statistical analysis, and machine learning, this technique extracts meaningful information for the stakeholders and supports evidence-based decision-making [27].

2.3.2.2 Clustering

Cluster analysis or clustering is a data mining technique which is used to find data segmentation (items that are grouped/clustered based on their similarities) and pattern information. It is usually involved as the preprocessing of other KDD (Knowledge Discovery in Databases) operations and the result is important for the whole KDD process. Clustering is widely used in applications of financial data classification, spatial data processing, customers segmentation etc. The most important problem in a cluster analysis is the partitioning of the dataset into segments (clusters) so that intra-cluster data are similar and inter-cluster data are dissimilar [22], [30], [31].

2.3.2.3 Sentiment analysis

Nowadays, in a digital world where opinions and impressions are expressed and shared publicly in social networks, sentiment analysis or opinion mining is vital for businesses since discussions can be utilized for improving the quality of services, as well as it is an important technique adopted by those organizations which realize the importance of such process and analysis [32].

Sentiment analysis is referred as *"an important tool of data analytics that analyzes people's opinion, impression, evaluation, attitude, needs and emotions towards tangible or intangible objects, issues or attributes, like products, services, individuals, events, topics etc."* [33]. More specifically, sentiment analysis involves the acquisition of data about customers opinion and their analysis to find reviews about entities like products, services, individuals, events etc. It is mostly utilized in areas like marketing and the political and social sciences.

The process of sentiment analysis can be classified into three types: (1) *document-level* where the whole document is analyzed as it is assumed to express sentiments about a single entity, (2) *sentence-level* where a sentence analyzed to extract a single sentiment of a known entity and (3) *aspect-based* where all the features of an entity are identified, and all sentiments are recognized inside a document [27]. The sentiment on all the three types could be positive, neutral, negative or having more classes.

2.3.2.4 Social network analysis

A social network is a structure which consists of social entities (individuals and organizations) and relationships between them. It is an interdisciplinary field as it has emerged from fields including sociology, statistics, and graph theory. This social structure can be modeled through nodes and edges, representing the entities and relationships respectively, and it can be visualized as a graph consisted of these nodes and edges [34].

In the past few years, more and more organizations have recognized the importance of analyzing data derived from social networks (social data). Thus, a new data analytics type has adapted to their operations. Social network analytics or social network analysis (SNA) is referred as the analytics type that is concerned with the composition of the structural attributes of a social network and the extraction of insights from the relationships among the participating entities. This technique has become popular in our days since billions of those social entities are using the online social networking structures (online social networks - OSN) such as news websites, forums,

reviews sites, and social media (microblogs, media sharing platforms etc.). It is widely recognized that SNA can provide significant benefits to organizations. The most important benefits are among others the capability of organizations can understand the structure of the network, gain insights about how the network operates, and make strategic decisions either by investigating the relationships among the entities or by looking metrics derived from the whole network [34], [35].

2.3.2.5 Spatial analysis

Many objects in the real world have some spatial attributes such as location, time, as well as other types of attributes. Spatial analysis is the process of examining these locations, attributes, and relationships of characteristics in spatial data. Using other analytical techniques aims to provide useful knowledge by extracting or creating new information from spatial data [36].

A notable aspect of spatial data is the spatial autocorrelation, which means objects that are physically close tend to be similar in other ways as well [37]. This can be observed with the weather data, where temperature, pressure and other attributes have similar values in geographical locations that are close to each other. Except for meteorology, spatial analysis is used in several studies including vehicle tracking systems [38], agricultural innovation [39], electrical energy [40] and urban infrastructures management [41], disease treatment [42], and spatial social network analysis [43], [44].

2.3.2.6 Visualization

The human visual system is amazingly powerful in spotting trends and relationships. People reading text remember 10% of the information for three days while visualizations are more likely to be recalled at 65% after the same interval. This represents the Picture Superiority Effect, which is the sense that concepts that are learned from viewing pictures are more easily recalled than those learned by textual equivalents', as well as it can be recalled with higher frequency [45].

The human brain perceives via a "Gestalt" process which tries to make sense of a whole signal through pattern recognition [46]. Gestalt reasoning is a process which attempts to determine how the human visual perception is creating clusters of pieces of visual information together [47]. Data visualization allows to harness these capabilities and therefore is widely used in transferring large amounts of information (quantitative and qualitative data) in a digestible form by combining them with a visual representation [48], [49].

Therefore, a great visualization is a far better way of retrieving information than paragraphs of text. Knaflic [50] lists some of the most important characteristics of a good visualization including: (1) *targeted visualizations* where the visualization should be tailored for the audience, (2) *the use of the appropriate visual*, e.g. bar charts to compare categorical data, (3) *the avoidance of clutter*, e.g. align elements and remove information that does not add value, (4) *the use of color and size* to signify importance (5) *good design*, the design should be kept simple, and aesthetically appealing, and (6) *storytelling*, where the visualization should keep the audience's attention and provide them useful information by telling a story. Knaflic also notes that if these characteristics are missing, the generated visualizations will be ineffective, and they distract the audience's attention.

Data visualization offers several techniques that could be applied in both static or dynamic mode based on business needs and goals such as maps, treemaps, scatter plots, line charts, bar charts and word clouds. These visuals allow for the big data to unveil its potential and create a framework of understanding, while they also enable the decision makers to investigate a large amount of data to communicate and take effective decisions [51].

Nowadays, data visualization has become an essential tool for modern and advanced companies where traditional reports are not so effective to take decisions especially with the rise of big data. Companies like Amazon, Apple Facebook, Google, Twitter, and Netflix are utilizing a wide range of data visualization tools to ask better questions of their data and make better business decisions [52].

2.4 Social media analytics

Social media analytics is defined as *"the analysis of structured and unstructured data collected from various social media channels"* [27]. The main characteristic of social media analytics is its data-oriented nature. The collection of these data across the internet from some different online platforms allow us to analyze and visualize them to gain valuable knowledge. Usually, the social media platforms that are selected for analysis are social networks (e.g. Facebook and LinkedIn), microblogs (e.g. Twitter and Tumblr), media sharing (e.g. Instagram, Flickr and YouTube), social news (e.g. Digg and Reddit) and review sites (e.g. Foursquare and TripAdvisor).

The analysis of social big data involves fields such as mathematics, informatics, sociology, management science, etc. As [53] defines, social media analytics includes processes such as

sentiment analysis, natural-language processing (NLP), social networking analysis (influencer identification, profiling and scoring), and advanced techniques such as text analysis, predictive modeling and recommendations, and automated identification and classification of a subject, people or content. In general, these techniques are applied using social big data in operations such as opinion mining and analysis, intelligence collection and analysis, socialized marketing, online education, and decision-making support.

The applications of social big data are divided into two main types:

- **Content-based Applications** where the text and its language are the most important factors as we can identify the users' preference, emotion, interest, demand, etc.
- **Structure-based Applications** where we can identify the users' hobbies, interests, and relations into a clustered structure (community).

Both content-based and community-based analyses are of vital importance to provide valuable knowledge [54]. Nowadays, the main application area of the social media analytics is the tourism industry, where businesses are focused on gathering and analyzing social big data for creating value in marketing by developing tools such as recommendation systems. This strategy, is vital for every business and organization as it is allowed (1) to find out what people are thinking about their products and services, their brand or their competitors, (2) to gain a wealth of information about people behaviors and preferences, (3) to identify new market trends and opportunities to stay competitive, (4) to create contextualized offerings to their customers, (5) to predict their demands, and (6) to make life-critical decisions in real-time.

In that context, the real-time analysis of user generated content from social media platforms, becomes the major driver for value creation for almost all the industries [27], [12]. Current research confirms this hypothesis by applying social big data analysis in various sectors.

[55] researched to explore perceptions of Asian restaurants (Korean, Japanese, Chinese and Thai) using Twitter analysis. They applied techniques such as text mining, word frequency analysis, and sentiment analysis on 86,015 tweets that were collected in four months, and they found that (1) the average sentiment score of Chinese restaurants was significantly lower than the others, (2) the most positive tweets referred to food quality, and (3) many negative tweets suggested problems about the service quality or food culture.

[56] used the reviews site TripAdvisor to extract and visualize the ratings and reviews for Hilton hotel. They proposed a framework which is primarily based on sentiment analysis and natural

language processing and found the types of travelers that are giving the lower and higher ratings (business travelers and couples), the months with the lowest and highest rates (July and December), and the travelers' emotions according to the most frequently used negative or positive words.

[57] developed a methodology which integrates data from reviews site TripAdvisor, travel site Booking.com with territorial and tourism data to extract meaningful information for tourism planning and decision-making. Their case study demonstrates the value of social media data and computational social science techniques in tourism planning and explores three questions related to tourist preferences (1) which the most popular destinations are, (2) why people chose those destinations, and (3) what attracts tourists attention and what do they appreciate/disregard. They found that the success of tourist destination does not only depend on the quality of the tourist industry offer but also it depends on the territorial setting of the destinations, including the natural, cultural and the physical characteristics of the places, as well as infrastructure and services.

2.5 Location intelligence

Location intelligence (LI) or locational data analysis is defined as "*the use of locationally referenced information as a key input in business decision-making" [58]*. It is derived from the concept of Business Intelligence (BI), which aims to leverage business data to make strategic decisions. LI expands BI by adding a spatial perspective to business data analysis creating a critical context to the decision-making process with the incorporation of powerful data correlation and visualization methods [59]. For instance, data-driven maps and location-based applications created using LI reveal spatial relationships and correlations with other types of business data that otherwise may not have been visible.

Unlike traditional business analytics, which tends to present results in reports, spreadsheets, charts and graphs, data-driven maps allow businesses to see the underlying proximity relationships and trends in their data. Examples of locationally-referenced business data include such items as property locations, customer locations, or supplier locations. These data are usually found in the form of addresses, geographic coordinates, and region designations. Combining these data with other types of geographic data such as population, road networks, climate information, topography, etc. we can analyze various spatially-referenced phenomena and gain valuable knowledge for our business, organization and public authority.

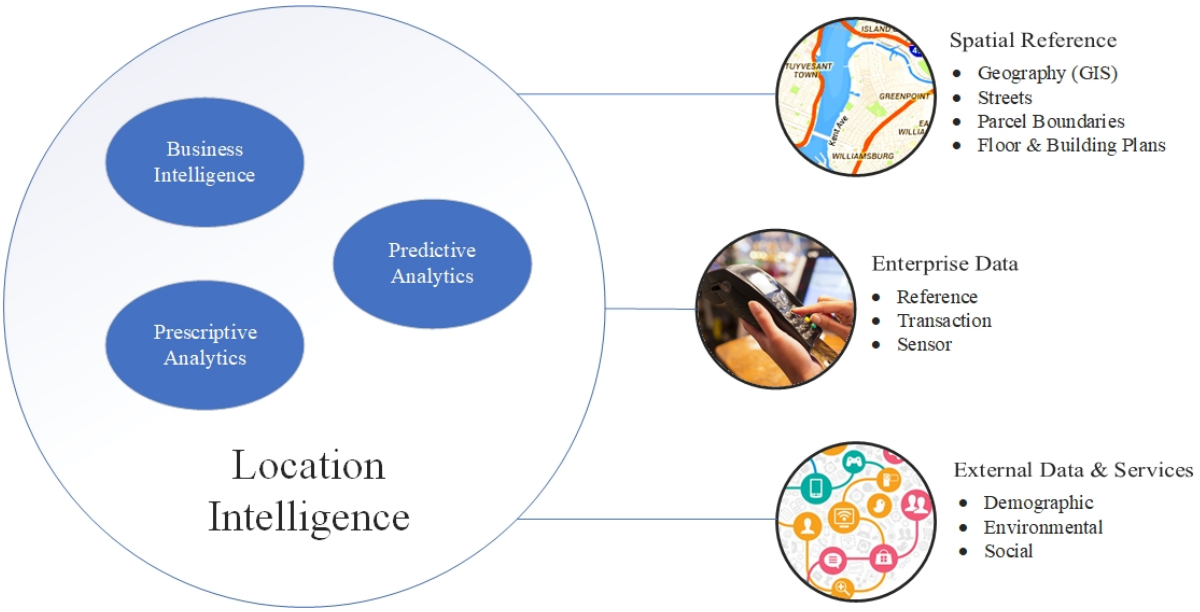


Figure 4 - Defining Location Intelligence [60]

Nowadays, businesses are being built on location data because they understand the business impact of "where". The main advantages derived by adding Location Intelligence as a part of the whole data analysis process are (1) businesses or public authorities can better understand external characteristics and how they affect their operations, and (2) gaining a much more complete picture for a phenomenon by integrating location and time dimensions to internal data [59].

Modern companies must follow three main steps to integrate LI into their business properly; these are *location discovery*, *location analytics*, and *optimization* [61].

2.5.1 Location discovery or Localization

To work effectively, enterprises must coordinate all their departments and branch offices. This also means the systematization and synchronization of various information from various sources into one database. Location Intelligence allows the collection and storage of convenient and comprehensible data throughout the entire enterprise. Additionally, it is the first step of a process which enables enterprises to see all the data from different numeric tables in the form of images or maps. This helps the companies to better understand the relationship between them, their data and their customers. Location Discovery is a process in which every significant item in the database is assigned at least one set of coordinates (for example: a and b). The assignment provides the items the ability to be shown on a map and be further examined in the following Location

Intelligence cores. Furthermore, Location Intelligence software permits the mentioned data to change according to real-life changes. For instance, if items change their location or quantification in real life, they can change coordinates or some other value in the system. After all items or objects get the location component in the system, the Location Intelligence data is combined with other data from Business Intelligence. This means that the next step (second Location Intelligence core) can begin.

2.5.2 Location analytics or Insight

Due to the simplified visualization from the first core of Location Intelligence, insight into available data is not difficult. For instance, companies can gain awareness and better understanding of their customers, especially their targeted group of consumers. In this case, Location Intelligence combines maps with enterprise data about sales, customer lifestyle, personalities, interests, opinions, motives, values, attitudes and so on.

The combination of all mentioned data provides companies with information about who their customers are, what they need/want/buy, how much they need/want/buy, where they need/want/buy it and so much more. The insight consists of three different processes - *visualization, analysis* and *forecast*. After the enterprises gain insight, it is much easier to optimize their strategies and be more competitive in today's market.

A. Visualization or mapping

Succeeding the localization of data or the specification of coordinates for each item, it is possible to begin the process of mapping. Mapping represents the portrayal of data on a map. In other words, all the objects with coordinates are colorfully depicted in interactive maps.

B. Analysis

This process allows companies to view data about their business on maps in different layers. They can, for instance, analyze and compare their different selling spots X and Y depending on different factors. This means they can define the number and characteristics of consumers for location X and location Y, as well as revenue for both. By showing the information in different map layers the companies can describe why one location has higher revenue even though the other one has more consumers and so on.

C. Forecast or prediction

The forecast is considered the key step of the analytical Location Intelligence core. This process enables the previously mentioned information to be merged with the frequency of certain kinds of circumstances occurring. This makes it simpler to predict what circumstance can or will arise and where.

2.5.3 Location optimization

Given that enterprises are completely aware of their target consumer characteristics and have information from previous cores, Location Intelligence allows them to identify areas with the highest number of likely and appealing to new consumers.

Chapter 3 - KnowLI: AN APPLICATION FOR KNOWLEDGE DISCOVERY THROUGH LBSNs

The continuous increase in the volume of data with the rise of online social networks, internet of things (IoT) and multimedia, has produced an overwhelming flow of data in either structured or unstructured form [16]. Digital economy through the tremendous use of web-based and digital services has transformed almost all the industry sectors including agriculture and manufacturing to more service-oriented like advertising and tourism industry.

Data-driven enterprises clarify the power of big data with creating more accurate predictions leading to non-obvious knowledge generation and better strategic decision-making [3]. This is achieved with the advancement in data processing and data storage technologies as well as the continuous development of advanced data analytics techniques which can be now adopted by any organization. Data analytics harnessing cloud computing and big data while they also involve revolutionizing techniques including data mining, machine learning etc. [12], [13], [16]. Currently, many researchers aimed to answer how big data analytics can be adopted into the decision-making process by developing several frameworks and using established analytics models and techniques [22].

In that context, we developed KnowLI, a system that follows the knowledge discovery in databases (KDD) process and aims to provide valuable knowledge from user-generated content in location-based social networks including Twitter, Foursquare, Instagram and Flickr.

This chapter describes the implementation of a web application, as well as the methods and technologies used to acquire, transform, store, analyze and visualize the data derived from the above location-based social networks.

The first section of this chapter describes the technologies used for the implementation and the system architecture. In the second section, the processes of data acquisition, cleansing and storage, analysis, and visualization are described in detail. The subsection 3.2.1 describes the four selected social networks and the steps that are required to collect public user-generated content, to transform these unstructured data into structured, and the storage in the SQL database and data warehouse. The subsection 3.2.2 describes the analytical processes performed to extract the sentiment from the records that containing text labeled as positive, neutral, or negative, as well as the entity analysis of each text to detect which entities are referred by the users' posts (e.g. locations, persons, events, consumer goods etc.) In subsection 3.2.3, the data querying and visualization processes are described.

This step has the main objective the creation of some useful visuals that will be used for decision-making purposes.

3.1 System architecture & technologies

The system's architecture is based on the traditional client-server communication and it is hosted in Google's cloud infrastructure known as Google Cloud. This cloud-hosted system, except the clients (personal computers - PCs, tablets, and smartphones) contains four components which are: *a web server*, *an SQL database*, *a SQL data warehouse*, and *a file server*.

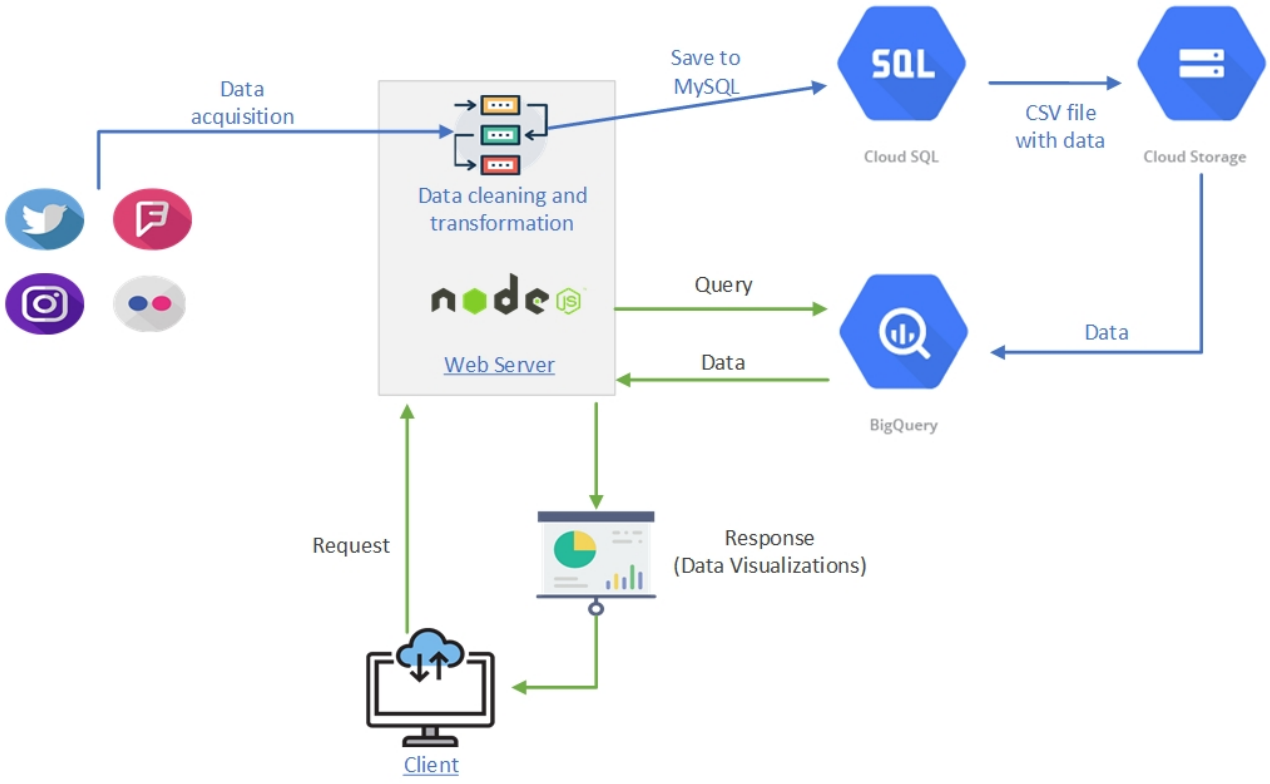


Figure 5 - System architecture

Among the used technologies are Node.js, AngularJS, Google Maps, Google BigQuery and MySQL. Both server-side and client-side of our application was written in JavaScript using Node.js and AngularJS (along with HTML and CSS) respectively. For the data visualization, the libraries D3.js (Data-Driven Documents) and Google Charts are selected, and for writing the system's code, the editor Visual Studio Code by Microsoft was used.

The web server hosts the Node.js platform, an open-source JavaScript run-time environment built on Google Chrome's V8 JavaScript engine. Node.js along with the appropriate libraries (e.g. express.js) can act as a web server as it can execute JavaScript code server-side. Like other programming

languages including Java, PHP, and Python, the Node.js run-time environment can serve documents and data to the clients, as well as it can communicate with databases for data storage and retrieval, and web APIs for data exchange.

The need for data storage is covered with using the Google Cloud SQL service in which a MySQL database instance has created to store all the data that are required by the system (e.g. users, sessions, projects, regions etc.), as well as the data that are fetched from the social networks (e.g. places, posts, images, coordinates, users profiles etc.).

As the system is continuously collecting data from the several sources and the size of the database grows, the execution time of the queries could become slower. This results in a worse user experience as the users should wait for the server to process the database queries. To address this issue, using a Data Warehouse deemed necessary. The data from the most frequently queried tables are also copied from the MySQL database to Google's data warehousing service, Google BigQuery. By querying the table in the data warehouse, great speed is gained (fast query processing and responses), since it is optimized for analytic access patterns and processes highly complex queries overall data [62]. The system also contains a file server which runs on the Google Cloud Storage service to store some necessary CSV, SQL and text files.

3.2 Data flow and processing

Twitter, Instagram, Foursquare and Flickr are some of the most popular online social networks that people used to share their experiences, store and organize important photos, and to communicate. Each minute, 470K Twitter posts (tweets) published, 49K Instagram photos uploaded[1], while Foursquare has recently overpassed the 3B monthly place visits[2] and Flickr host more than 25M photos on high traffic days[3].

We have selected these platforms as our data sources as they provide the two types of user-generated content we wanted to collect, textual posts and photos, as well as they offer the required open web APIs (Application Programming Interfaces) which allow us to programmatically read and therefore collect data to our system. The four selected social networks are classified into two categories according to the content users are generating, Media Sharing (Photos) and Posts & Reviews (Text).

A general overview of the above data flow steps is presented in the following figure (Figure 6).

[1] Domo (2018) - Data Never Sleeps Infographic, version 6.0
[2] Medium (2018) - Foursquare's Third Consecutive Year of 50% Revenue Growth or Better
[3] Expandedramblings (2018) - 17 Interesting Flickr stats and facts (May 2018)

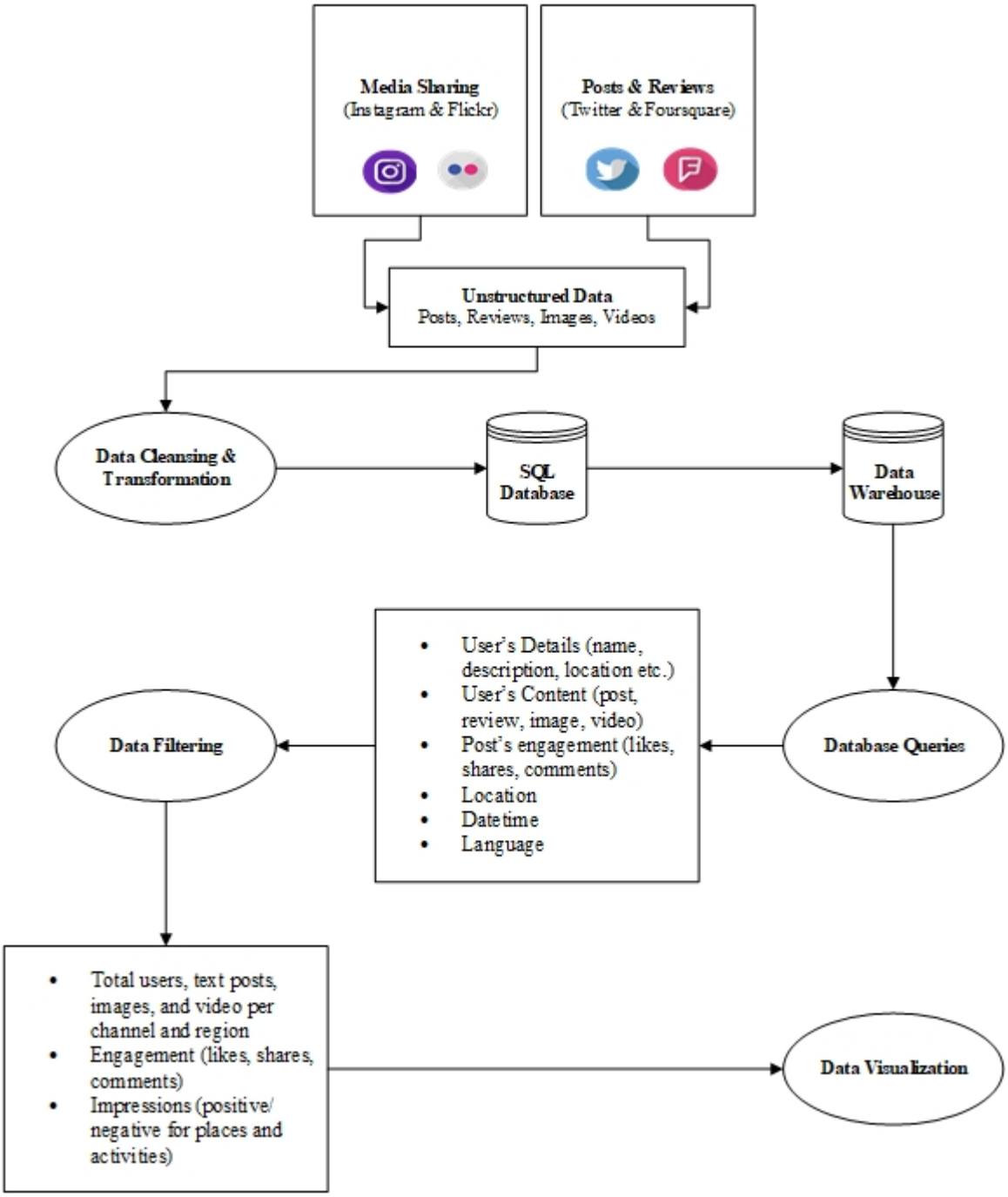


Figure 6 - Data flow Diagram

3.2.1 Data acquisition, cleansing and storage

The main process which runs continuously in the background by the web server is the data acquisition that includes the cleansing and storage of the acquired data. In this process, the web server communicates with the web APIs (Application Programming Interfaces) of each social network

every *t* seconds to fetch the newly geo-tagged user-generated content including places, textual posts and photos.

Each API has different limits for the incoming calls and thus the time *t* is different for each social network to avoid parallel requests from web server to all the social networks' APIs in the same time.

To fetch the user-generated data from social networks, the users of our web application should have created at least one project (Figure 7) which contains at least one geographic region of interest (ROI).

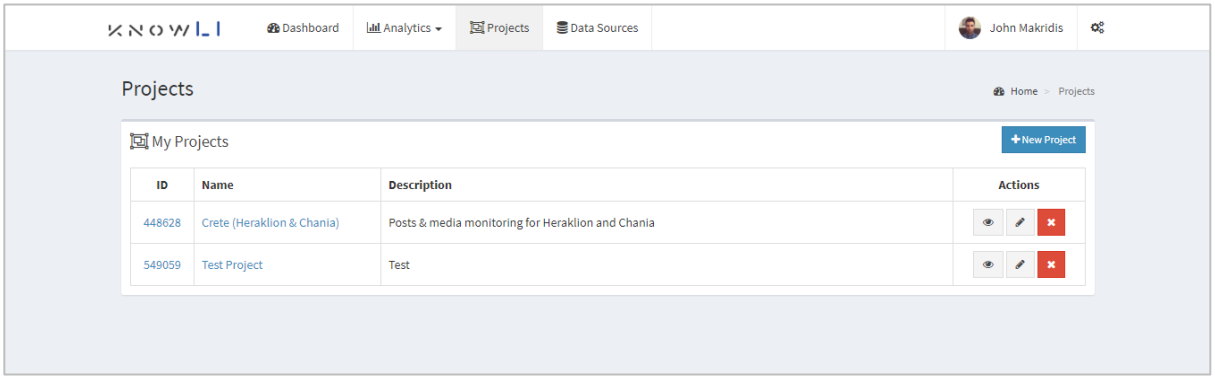


Figure 7 - Projects page

Figure 8 shows the map with two regions for the selected project as they are drawn by user. For each ROI, there are two types of requests to be executed repetitively:

1. Requests for places (venues, points of interest and businesses inside the selected region)
2. Requests for textual posts/reviews & pictures for each place

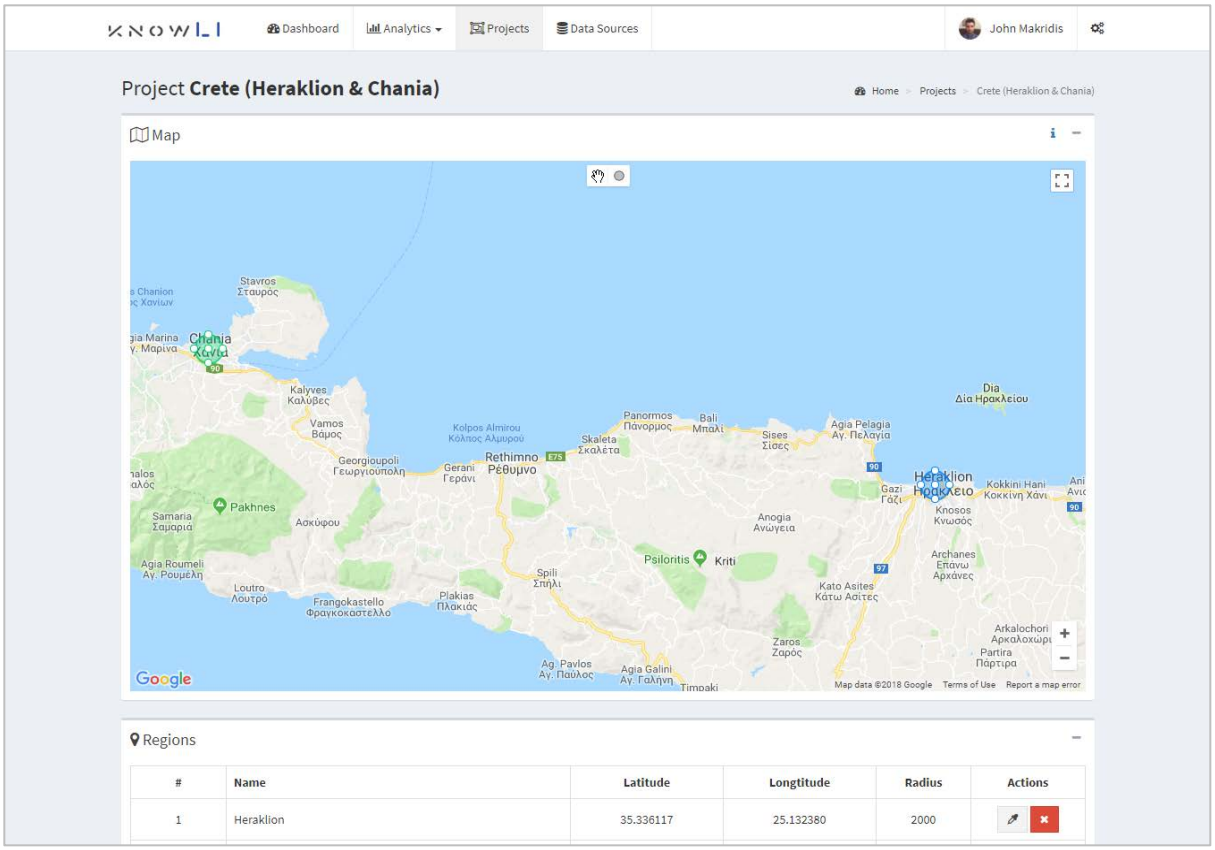


Figure 8 - A single project including its regions drawn on the map

Each of these requests is sent by the server to the social network's API through the HTTP protocol (sending an URL with the API address and multiple parameters). One of the mandatory parameters is the API key or client id generated in the social network's developers' website that is used to authenticate the requests (by whom the request was performed). The other parameters specify the data we want to receive. After the execution of the request, the social network responds with data objects in JSON (JavaScript Object Notation) format.

Figure 9 shows an HTTP GET request with the required parameters to Foursquare API, searching for places in the selected region.

https://api.foursquare.com/v2/venues/search/?intent=browse&ll=35.336117,25.132380&radius=2000&v=20180530&m=foursquare&limit=50&client_id=[your_cliend_id]&client_secret=[your_client_secret]


```
{
    "meta": {
        "code": 200,
        "requestId": "5b14e761dd57973ba199e7a7"
    },
    "response": {
        "venues": [
            {
                "id": "56ab379f498e0aacc2127cb4",
                "name": "Alpha Bank Κορνάρου",
                "location": {
                    "lat": 35.336364894269565,
                    "lng": 25.133733350804846,
                    "labeledLatLngs": [
                        {
                            "label": "display",
                            "lat": 35.336364894269565,
                            "lng": 25.133733350804846
                        }
                    ],
                    "distance": 125,
                    "cc": "GR",
                    "country": "Ελλάδα",
                    "formattedAddress": [
                        "Ελλάδα"
                    ]
                },
                "categories": [
                    {
                        "id": "4bf58dd8d48988d10a951735",
                        "name": "Bank",
                        "pluralName": "Banks",
                        "shortName": "Bank",
                        "icon": {
                            "prefix": "https://ss3.4sqi.net/img/categories_v2/shops/financial_",
                            "suffix": ".png"
                        },
                        "primary": true
                    }
                ],
                "referralId": "v-1528096707",
                "hasPerk": false
            },
            {
```


Figure 9 - Example places search on Foursquare API & the response with places array

To search for places in social networks we need either the region's coordinates or the region's name. Before the request, an SQL query must be executed in the database to get the region's latitude, longitude and name. Using region's coordinates, our search provides more accurate data as we request for places near a real point on the map. The alternative way can provide irrelevant data (noise) as there may be several other places with the same or similar name but is useful if the web API of the social network doesn't provide the capability of searching using coordinates.

After the successful communication with the social network's API and having acquired the places of the selected region, fetching users' posts inside that region or place is followed. In the case of posts

search in a region, the coordinates or the region's name should be defined as parameters. However, the posts search in fetched places, must include only the place's unique identifier (ID).

Most web APIs divide the records into different pages to prevent responses with large amount of data. Thus, for each region and place, a recursive function is called to fetch the available posts by traversing all the pages.

Figure 10 - Example request for posts from Foursquare API & response with posts array

Figure 10 shows the response with posts (reviews) from the Foursquare API for a given place (in our example Starbucks store in Heraklion, Crete).

After the data acquisition requests, for both places and posts in the selected region(s) the data cleansing is followed. This process is performed to keep the necessary (for our application) data, parse them if it needed, and transform them in a structured format to be stored in our SQL database and data warehouse.

These transformations including dates which in some cases must be converted from Unix time to ISO Date (YYYY-MM-DD) format, users' profile photos in which the URL should be created by merging different variables (as shown in Figure 3.2.1.4), posts' coordinates which should be transformed from array to two decimal numbers, and hashtags, which must be extracted from plain text.

Then, having the acquired data in a structured (column-based) format, the SQL queries for their storage are executed. Each entity in our system has its corresponding data table. The main tables in the database are (1) Projects, which stores the projects created by the users through the web application, (2) Regions, for storing the user's ROIs which are drawn in project's map, (3) Places, holds all the fetched places (venues, businesses, points of interest etc.) for each ROI, and (4) table Data which stores all the fetched posts from the social networks' web APIs.

Column Name	Description
id	Auto increment number
project_id	Project ID
project_name	Project name
project_description	Project description
u_id	User ID (to who the project belongs)

Table 1 - The columns of Projects data table

In each row of the Projects' table, there are five columns that they contain the following: unique auto increment number, user's ID, project's name and description. This table is designed and created to store different projects as users may want to divide the regions that they want to monitor in city, country or continent levels. For instance, users can create a project to draw regions that they want to

monitor in Greece and another project to monitor regions in Spain. If the number of regions is small, then they could be assigned in one project.

Having created one or more projects, the user should draw the regions of his interest inside them (Figure 11).

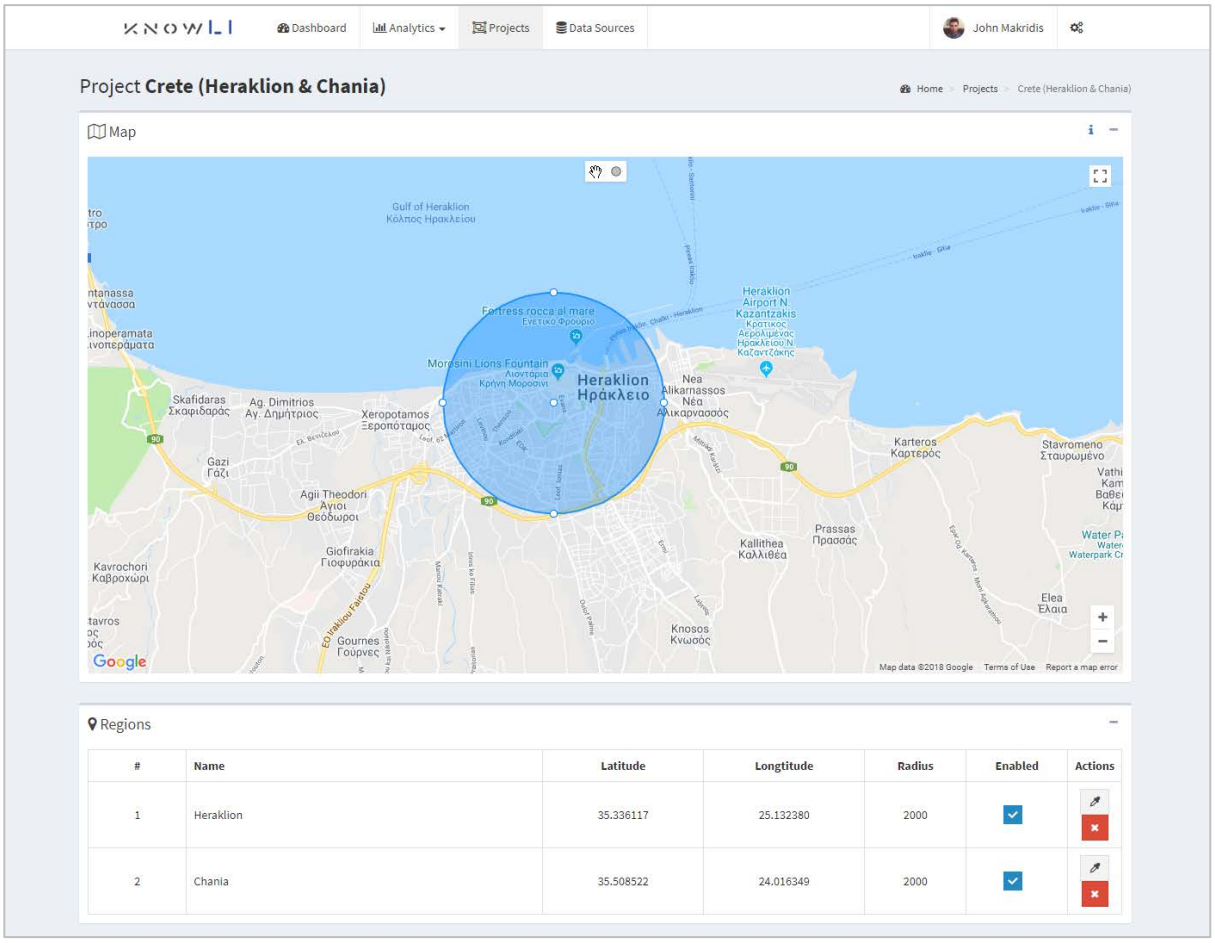


Figure 11 - A region drawn by the user in a specific project

For each region that user draws, there are some necessary data that should be stored in the database's Regions table (Table 2). These are (1) roi_id, a unique number generated by the system, (2) region's coordinates including latitude, longitude, and the radius from the center, (3) the identifiers of region's project and user, (4) a title, small description, and color, and (5) the region's status whether enabled or disabled from fetching data.

Column Name	Description

id	Auto increment number
roi_id	Region ID
roi_lat	Region latitude
roi_lng	Region longitude
roi_radius	Region radius
project_id	Project ID (in which project the region belongs)
userID	User ID (which user created the region)
title	Region title
description	Region small description
is_enabled	Region is enabled or not for data fetch
roi_color	Region color

Table 2 - The columns of Regions data table

The next step, after the creation of the ROIs, is to fetch the places inside them. As Twitter and Flickr don't provide the capability for creating pages or profiles about businesses, venues or places, we periodically executing HTTP requests to the web APIs of Foursquare and Instagram to fetch the places' data by giving as parameters either the coordinates or name of our ROIs.

The responses from the two APIs with the places' data contain several useful information for both business-level and spatial analyses. The most important information that must be stored in the Places table (Table 3) are among others, the place's category, average rating, total ratings and total check-ins.

Column Name	Description	Column Name	Description
id	Auto increment number	reviews	Number of reviews
place_name	Name of the place	categories	Place's categories
place_id	Identifier of the place	price_tier	Place's price tier
roi_id	Identifier of the region (in which region the place belongs)	price_message	Price's description (cheap, expensive etc.)
location_lat	Latitude of the place	price_currency	Price's currency (euros, dollars etc.)
location_lng	Longitude of the place	rating	Average rating of the place
distance	Distance from the center of region	ratings_total	Total ratings (count)
postal_code	Postal code of the place	source	The source of the place (which social network)
country_code	Country code of the place	lastFetch	Unix time of the last API call
country	Country of the place	lastFetchDate	ISO Date of the last API call
city	City of the place	nextFetch	Unix time of the next API call
state	State of the place	nextFetchDate	ISO Date of the next API call
address	Address of the place	place_shortName	Short name of the place (acronym)
checkins	Number of check-ins	place_url	Place's web address (URL)
users	Number of users checked-in		

Table 3 - The columns of Places data table

Having defined the regions in the web application and acquired their places from social networks, the next operation is fetching actual data derived from user-generated content in these regions and

places. The acquisition of these geo-tagged posts is the process that executed more frequently among the others in the background as (1) our visuals are depending on them, and (2) we want the system acquiring the information as soon as possible (near actual creation time from the user).

From the type of the social networks (text posts and reviews, or media sharing) and the information publicly available from them, the posts are stored in the corresponding columns of the *data* table. This table stores some information around a user's post (Table 4) which informally could be divided into 5 categories:

1. Columns for textual posts (text, hashtags, language, sentiment, entities)
2. Columns for photos and videos (title, caption, dimensions, views)
3. Columns for post's location (latitude, longitude, place's id, place's name)
4. Columns for user's data (name, profile photo, location, likes, friends)
5. Columns for general data including creation date, source, URL and post engagement (likes, shares, comments),

Column Name	Description	Column Name	Description
id	Auto increment number	photo_server	The server of the photo (Flickr parameter to fetch the photo's detail along with the photo's id, secret, and farm)
roi_id	Region identifier (in which region the post is published)	photo_farm	The farm of the photo (Flickr parameter to fetch the photo's detail along with the photo's id, server, and secret)
record_id	Unique identifier for the post (the post's ID on the specific social networks)	photo_dateCreated	Creation date of the photo
code	Post's unique code (Instagram only)	video_views	Number of the video's views
place_id	Identifier of the post's place (in which place the post is published)	is_video	Binary number which defines whether the post is video or not (Instagram only)

place_name	Name of the post's place (in which place the post is published)	video_url	Web address (URL) of the video
channel	Social network of the post	canonical_url	Unique web address of the post
post_latitude	Latitude of the post	user_id	Post's user identifier
post_longitude	Longitude of the post	user_username	User's username
text	Post's text	user_name	User's full name
language	Post's language	user_location	User's location (text)
type	Post's type (textual post, photo, or video)	user_photo	User's profile photo
source	Social network's source (i.e. Foursquare for iOS)	user_gender	User's gender
likes	Number of likes	user_url	User's profile web address
shares	Total number of shares	user_posts	Number of user’s posts
comments	Total number of comments	user_followers	Number of user’s followers
saves	Total number of saves (Foursquare only, when a user saves a review written by another user)	user_friends	Number of user’s friends
agrees	Number of agrees (Foursquare only, when a user agrees with a review written by another user)	user_likes	Number of user’s likes (by other users)
disagrees	Number of disagrees (Foursquare only, when a	user_description	User's profile description

	user disagrees with a review written by another user)		
hashtags	Post's hashtags	user_dateCreated	Creation date of the user (in social network)
photo_id	Photo's identifier	user_lang	User's language (as it is defined in the social network)
photo_url	Photo's web address (URL)	dateCreated	Unix time of the post's creation
photo_title	Title of the photo	dateCreatedUnix	ISO Date of the post's creation
photo_caption	Caption of the photo	dateFetched	Unix time of the post's fetch (when it was discovered by our system)
photo_views	Number of photo's views	dateFetchedUnix	ISO Date of the post's fetch (when it was discovered by our system)
photo_width	Width of the photo	ms_sentiment_score	Post's sentiment (decimal number from 0 to 100)
photo_height	Width of the photo	entities	Post's entities (entities/topics extracted by the post's text)
photo_secret	The secret of the photo (Flickr parameter to fetch the photo's detail along with the photo's id, server, and farm)		

Table 4 - The columns of Data (acquired posts) table

The last step in Storage process is the copy of the acquired data to a SQL-based data warehouse on Google's cloud infrastructure. The data warehouse is a crucial asset in our application as it provides the capability for supporting real-time data-driven decision making and online analytical processing (OLAP). By querying the tables in the data warehouse, great speed is gained (fast query processing and response), since it is optimized for analytic access patterns and processes highly complex queries over all data [62], [63].

To test the performance in MySQL Database Server and Google BigQuery Data Warehouse we executed a simple “SELECT” query on both systems. In our demonstration, we are querying for all the columns of the latest one hundred thousand records in the “Data” table which holds all the acquired users’ posts.

```
SELECT * FROM `data` ORDER BY `data`.`id` DESC LIMIT 100000;
```

The test resulted in having the response from data warehouse 66% faster than MySQL database (Figure 12).

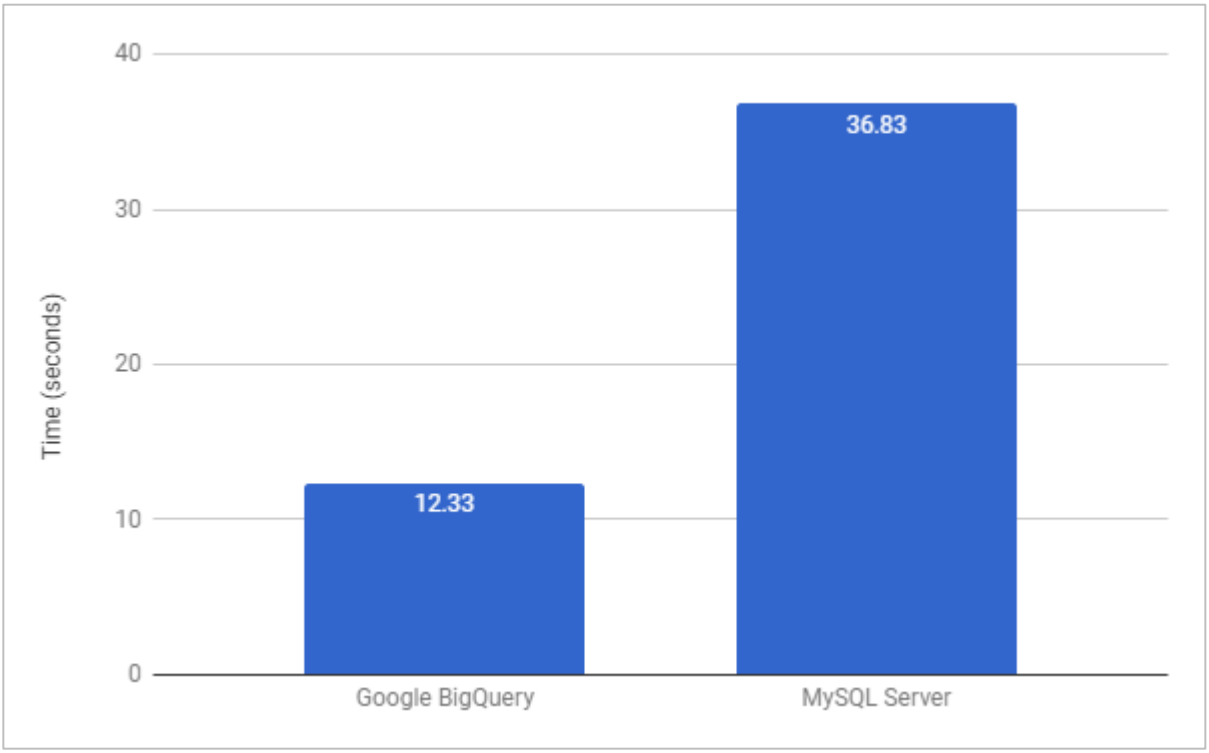


Figure 12 - Query processing time in MySQL database and Google BigQuery

3.2.2 Data analysis

The next step after obtaining, cleaning, and storing social media data is analysis. A large percentage of the acquired posts from social networks contain text. This user-generated text in most cases contains user’s views and impressions about a topic, product, place etc. that he is talking about.

One of the main goals of this thesis is to extract this information from the posts that are published inside our regions of interest and thus, to achieve this, our system uses the Google Cloud Natural Language API. This API contains several machine learning models which allow us to perform sentiment and entities analyses on users' textual posts to extract their sentiment classified as positive, negative or neutral, and entities/labels that are referred to these posts (e.g. location, person, event etc.) respectively. Both analyses using Google's API was applied in Twitter posts and Foursquare reviews as they contain more important and meaningful text than Instagram and Flickr.

3.2.2.1 Sentiment analysis

Using the Sentiment Analysis service provided by Google in their Cloud Natural Language API, we can send requests including the text of our acquired posts and get their sentiment on a numeric scale. Each text can be classified as negative (score -1.0 ~ -0.25), neutral (score -0.25 ~ 0.25), and positive (score 0.25 ~ 1.0).

In more detail, the Google's Sentiment Analyzer is firstly extracting the sentences from the given text, and then, it classifies each sentence (or text fragment/word) as negative, neutral or positive.

To achieve this, for each sentence, the algorithm defines a score vector s^m which encodes sentiment scores for every word in WordNet [64]. Given a tokenized string (text broken into pieces like words, keywords, phrases and symbols) $x = (w_1, w_2, ..., w_n)$ of words, the algorithm classifies its sentiment using the following function, where S_i is the score of each word.

$$\text{score}(x) := \sum_{i=1}^{n} s_i.$$

In some cases, when the |score(x)| is below a predefined threshold the algorithm classifies the string as neutral [65].

The following example (Figure 13), shows a sample text from an acquired post in Twitter sent to the Google's Natural Language API, and the response from the API in JSON format.

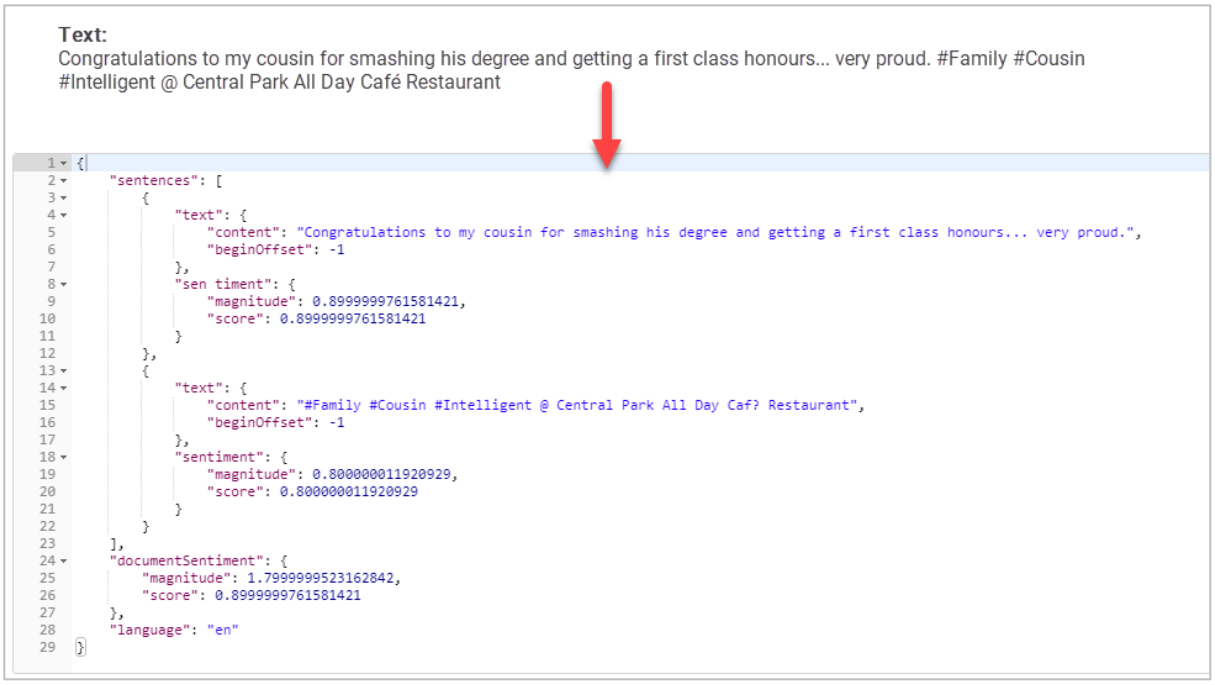


Figure 13 - Example sentiment analysis for a given text

In this example, Google's algorithm breaks our text in two sentences which are divided by a full stop. For each sentence, the algorithm computes the sentiment score, and finally, it computes the sentiment score for the whole text included in the *documentSentiment* object. Also, the algorithm returns the text's language which was detected during the process. According to the numeric range, the sentiment score belongs, our system assigns a label of positive, neutral or negative for the specific text. Then, the specific record which contains this text in our database is updated with both score and label in the appropriate columns providing a valuable information for our acquired textual posts.

3.2.2.2 Entity analysis

The Google's Entity Analysis service provided through their Natural Language API, is a procedure like Sentiment Analysis but it also involves a semantic perspective. According to the official documentation, *"Entity Analysis inspects the given text for known entities (proper nouns such as public figures, landmarks, etc.), and returns information about those entities"* [66].

Although there is no published literature by Google on how entity analysis works, by reading the official documentation and using the service in this implementation we can also understand other Google services may be involved to this process including Google Search and Google Maps.

The following figure (Figure 14) shows the response from Google's API with the detected entities for the tweet *"Lovely evening cruising on the Mediterranean with @RoyalCaribbean heading back to Venice!!"* which published inside one of our regions of interest. In this text, the entity analyzer has correctly detected three entities, including two locations, the Mediterranean Sea and Venice, and one person which is another user of Twitter with username @RoyalCaribbean. We observe each entity have a *metadata* object which contains Wikipedia URLs or other websites where this word was referred at.

Text:
Lovely evening cruising on the Mediterranean with @RoyalCaribbean heading back to Venice!!


```
{
    "entities": [
        {
            "mentions": [
                {
                    "text": {
                        "content": "Mediterranean",
                        "beginOffset": -1
                    },
                    "type": "PROPER",
                    "sentiment": null
                }
            ],
            "metadata": {
                "mid": "/m/04swx",
                "wikipedia_url": "https://en.wikipedia.org/wiki/Mediterranean_Sea"
            },
            "name": "Mediterranean",
            "type": "LOCATION",
            "salience": 0.3980683982372284,
            "sentiment": null
        },
        {
            "mentions": [
                {
                    "text": {
                        "content": "@RoyalCaribbean",
                        "beginOffset": -1
                    },
                    "type": "PROPER",
                    "sentiment": null
                }
            ],
            "metadata": {},
            "name": "@RoyalCaribbean",
            "type": "PERSON",
            "salience": 0.3166342079639435,
            "sentiment": null
        },
        {
            "mentions": [
                {
                    "text": {
                        "content": "Venice",
                        "beginOffset": -1
                    },
                    "type": "PROPER",
                    "sentiment": null
                }
            ],
            "metadata": {
                "wikipedia_url": "https://en.wikipedia.org/wiki/Venice",
                "mid": "/m/07_pf"
            },
            "name": "Venice",
            "type": "LOCATION",
            "salience": 0.2852974236011505,
            "sentiment": null
        }
    ],
    "language": "en"
}
```


Figure 14 - The entities extracted from the given text

Both types and mentions of entities are important information as they provide us the ability to understand what is mentioned inside a text (entity's type) and where this entity is mentioned.

In that context, the communication between our system and Google's Natural Language API to extract the entities from the acquired textual posts', was deemed necessary since one of our goals was to answer the questions "*What entities users are mentioning in their posts?"* and "*Which entities are mentioned most frequently?*".

3.2.3 Data querying and visualization

The last step in our system's data flow is the data querying and visualization. After having acquired, transformed, stored and analyzed the data from social networks, the next process has to do with the presentation of data properly through interactive and easy-to-understand visuals to the stakeholders. By querying the *data* table in the data warehouse, great speed is gained since it is optimized for processing highly complex queries over huge amounts of data, and thus the information is loading faster on the user's screen.

In our implementation, we categorized the generated visuals in two categories (1) posts & photos analytics, and (2) business-level analytics, as they described in the following subsections and presented in the next chapter (chapter 4).

3.2.3.1 Posts & photos analytics

The category 'posts & photos analytics' in our web application, is divided into two sub-categories, posts analytics and photos analytics, which contain all the visuals that are presenting analytics and insights from textual posts and photos respectively. Several visualizations generated to present these data according to the data types and the appropriate charts and graphs to display them in users.

Regions overall numbers

For each region drawn in our application's projects, a small box is automatically generated in the dashboard showing the overall numbers of the region in a given date range by the users. These numbers representing the total users, posts, photos, engagement (likes, shares, comments), and sentiment (classified as positive, neutral, or negative) for the specific region.

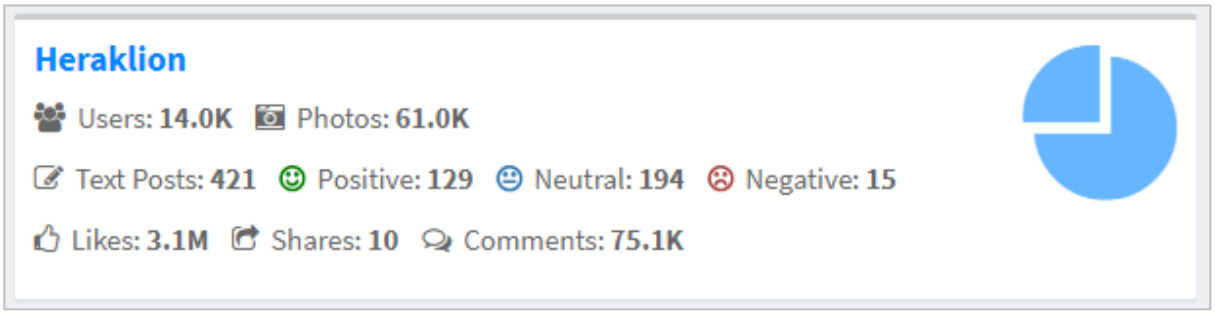


Figure 15 - A region's overall numbers

Posts' map

In the posts map, all the acquired posts are displayed as markers on the map based on their actual coordinates or the place they published (Figure 16). By clicking on a marker, a popup appears which contains the source, sentiment, place, and text or photo of the specific post. On this popup, we do not display the user's personal information like his full name, profile picture, description, profile URL, etc. respecting his privacy.

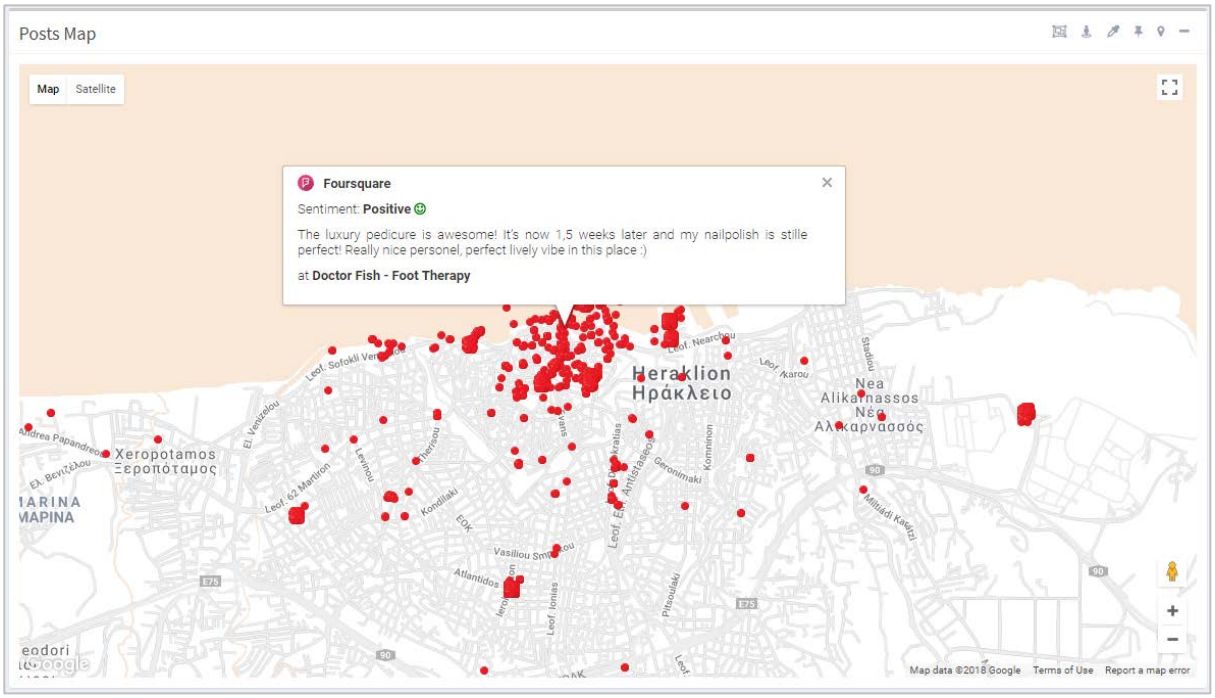


Figure 16 - Posts' map

As shown in figure below (Figure 17), the posts' display could also switch into clusters to provide us the knowledge of which subregions have the most posts or into heat map where subregions with most posts represented with warmer colors (Figure 18).

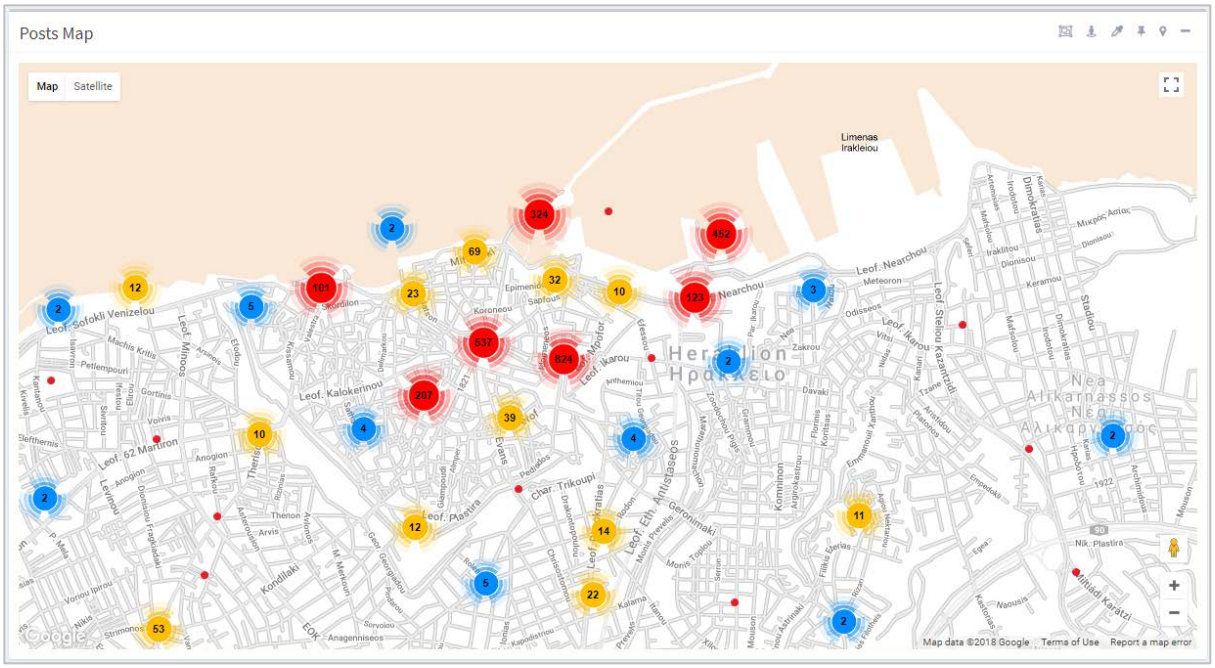


Figure 17 - Posts' map with posts' clusters

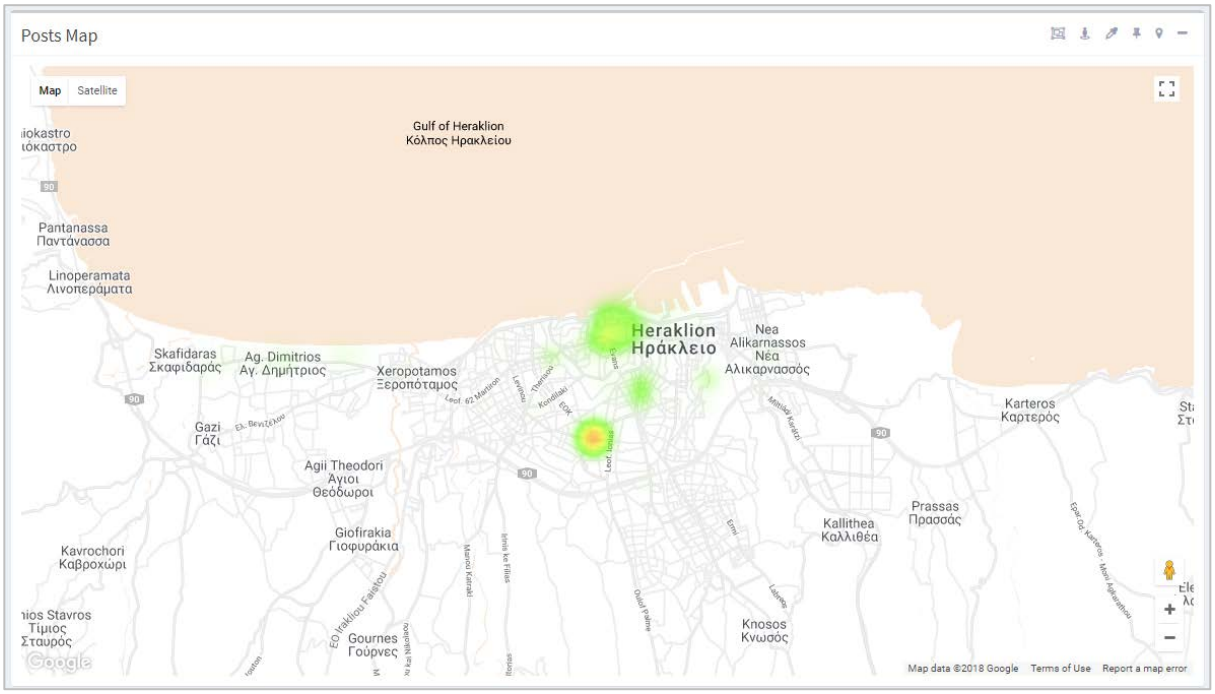


Figure 18 - Posts' heatmap

Total textual posts and sentiment per day

The chart with total textual posts and sentiment per day distributes the numbers about textual posts as they presented in the regions' overall numbers boxes in a daily view. In the horizontal axis, the

dates of our defined date-range are shown, as well as the number of posts per day classified as positive and negative on the vertical axis.

In this chart (Figure 19), we can observe which days users are more active (according to their published posts) by watching the number of posts per day, as well as we can investigate what happened in a particular day if we observe an unexpected growth of negative against positive or vice versa.

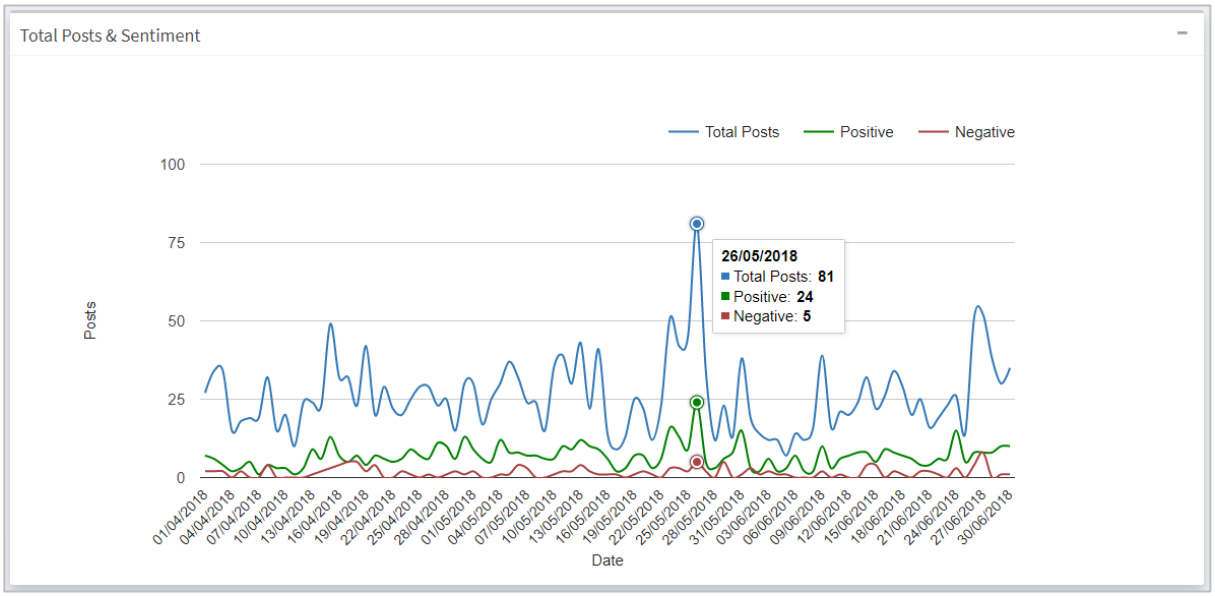


Figure 19 - Total textual posts with sentiment

Total posts per channel

The distribution of textual posts and photos per social network provide us the information about which of these platforms users prefer to share their impressions, experiences and opinions in the digital world. Since in our database and data warehouse we are storing this information (where a specific post or photo is published), we can perform a simple query to visualize this information to the stakeholders.

Posts' sentiment

As described in section 2.3.2.6, the analysis and extraction of users' sentiment from their published post, is significant information for both local authorities' and businesses' decision-making process. Our system communicates with Google's Natural Language API (see section 3.2.2.1) to extract the sentiment of the acquired textual posts from Twitter and Foursquare users and providing the

capability of giving the general picture about users' impressions in the defined regions and date ranges.

Posts' entities

The entity analysis in our application provides us the knowledge of "what the users are talking about in social networks" inside our regions of interest. Similarly to the sentiment analysis, our system uses Google's Natural Language API (as described in section 3.2.2.2) to extract the entities/topics users are talking for.

Posts' languages

In web and social media analytics, the detection and visualization of user-generated content language is a vital procedure. This information is used mostly in marketing and advertising processes since the stakeholders want to create personalized/targeted advertisements and plan their marketing strategies. For example, a tourism company wants to create a digital advertisement for a specific target group, people that are from a specific country or speaking a specific language. Without having this information about visitors' country and language from their website or social networks, the company can't implement the planned advertisement.

In that context, our web application stores the text's language as it fetched from social networks web API or detected from sentiment analysis as described in section 3.2.2.1. Having stored the posts' languages, we can execute a simple query in the data warehouse and distribute the post in languages.

Top hashtags and words

As mentioned in section 3.2.1 (data cleansing process), a custom algorithm was developed to extract the hashtags from the acquired textual posts. Having stored the posts' hashtags in a separate column, we can execute queries in the data warehouse to get all the records that include hashtags, create clusters from them, and visualize them as a words cloud.

Word cloud is a significant branch of data mining which has gained increasing attention and more opportunities in the Big Data era. It is a type of weighted list to present text data which contain words that are used with high frequency [67].

Since a large number of social networks users are adding hashtags in their posts to mention specific topics or declare their impressions, it is considered important for the stakeholders to analyze these words to extract the most-discussed topics, find content tagged with specific words, and plan their

marketing strategies and campaigns (hashtag marketing) [68], [69]. This method provides to the brands the capability of making their content more searchable, encourage users to talk about their brand, and increasing their engagement. Similarly to hashtags, the top words cloud shows the words were included most in acquired posts' text. By observing the words cloud, we can identify what users are talking about (extract topics, locations, events etc.).

Social engagement

Social engagement is a metric which sums up the number of likes, shares and comments on posts published in social networks for several topics, regions and periods. The analysis of social engagement can be used for the evaluation of the attractiveness of tourism experiences associated with an event and can be a significant factor for more personalized offers focused on customer satisfaction. In addition, the social engagement can also clarify us the special happenings and days, as well as display users with the most followers (influencers) in social networks.

Influencer marketing or e-word of mouth marketing can contribute to enhancing destination attractiveness or destination branding since influencers can spread messages affecting communities in the digital world.

Posts per day of week and hour

Having the actual dates of the acquired posts, we can create visualizations to distribute this information in bar charts with days of week and hours. These charts provide us the knowledge of which days and hours users are more active, generating content, interacting and communicating in the social networks.

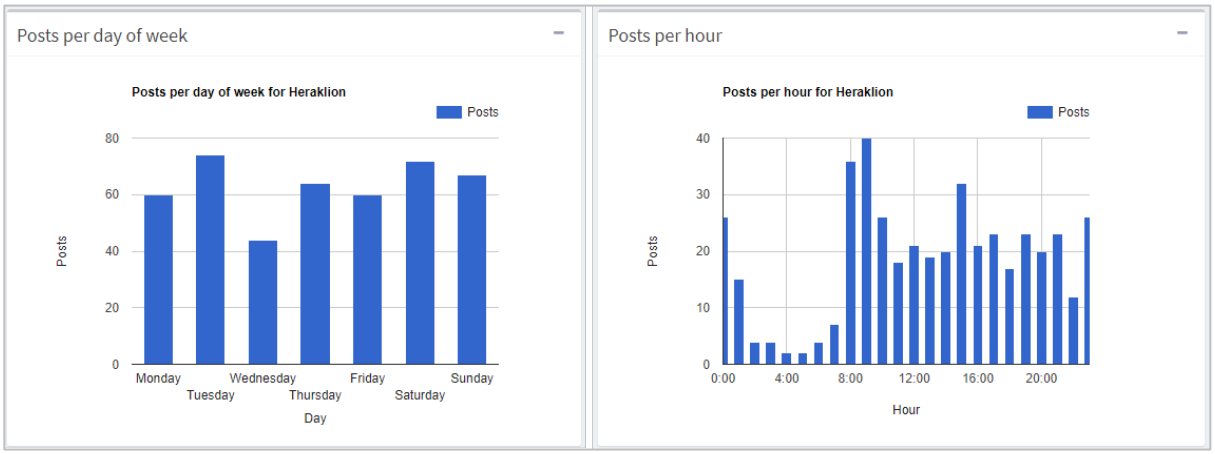


Figure 20 - Posts per day of week and hour

3.2.3.2 Business-level analytics

The visuals generated in the business-level analytics, are presenting information about the acquired places' pages in Foursquare. These pages are mostly created and managed by businesses around the world to present their services, prices, contact details, and interact with their customers.

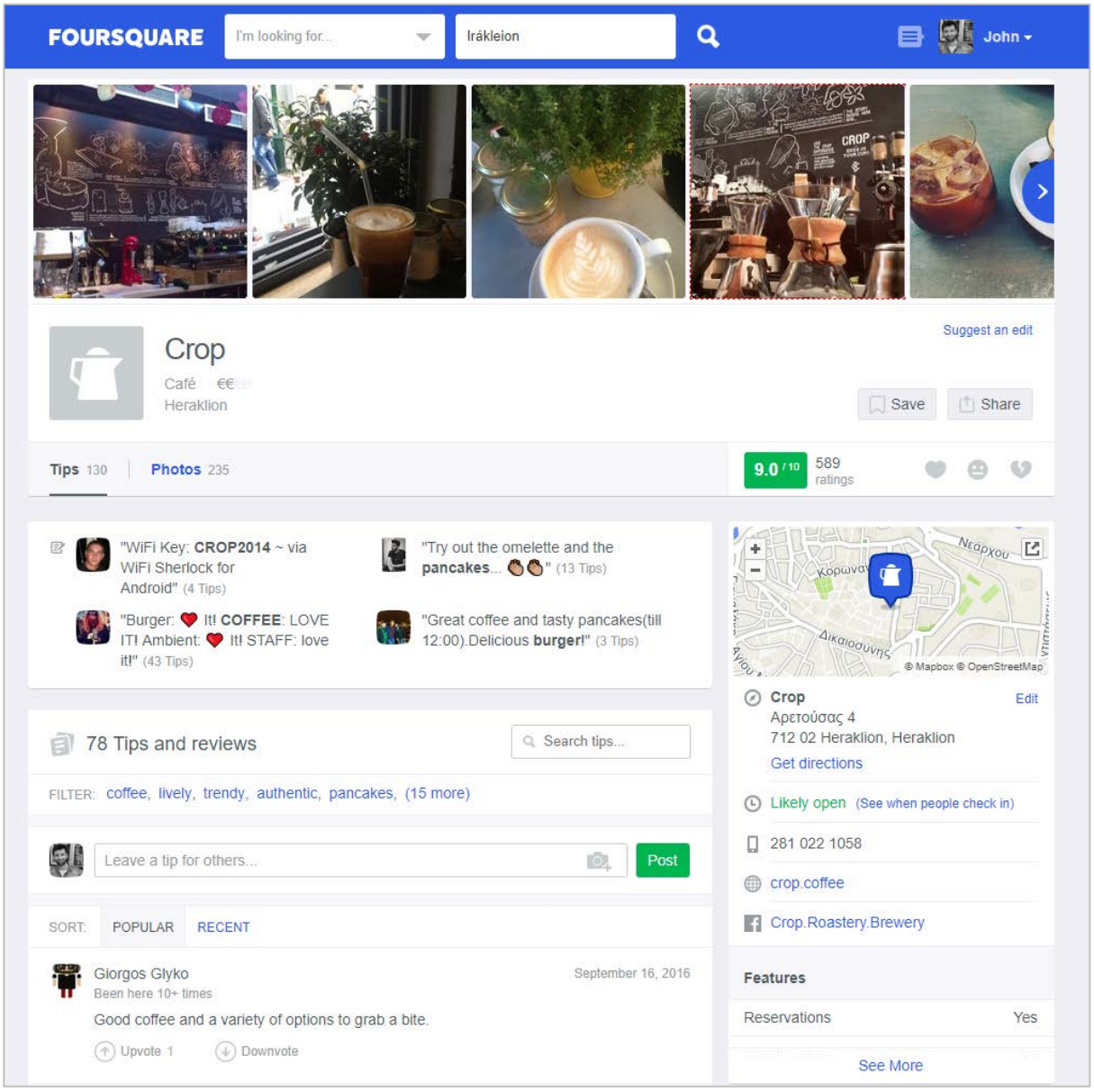


Figure 21 - A place's page on Foursquare

The places' data we have collected from Foursquare web API, among others consists of place's name, category, address, city, coordinates, total users, check-ins, average rating and total ratings. Having these data, for each project, we are generating two data visualizations to represent valuable

information about the acquired places: (1) a places' map extended with a data table, and (2) a pie chart which displays the top businesses types for each region of interest.

Places' map

The places' map displays all the acquired Foursquare places as markers from the place's actual coordinates. By clicking on a place's marker, a popup appears which contains some general information about this place including its name, category, address, source (platform), and a data table which contains its average rating, total ratings, check-ins and reviews/posts.

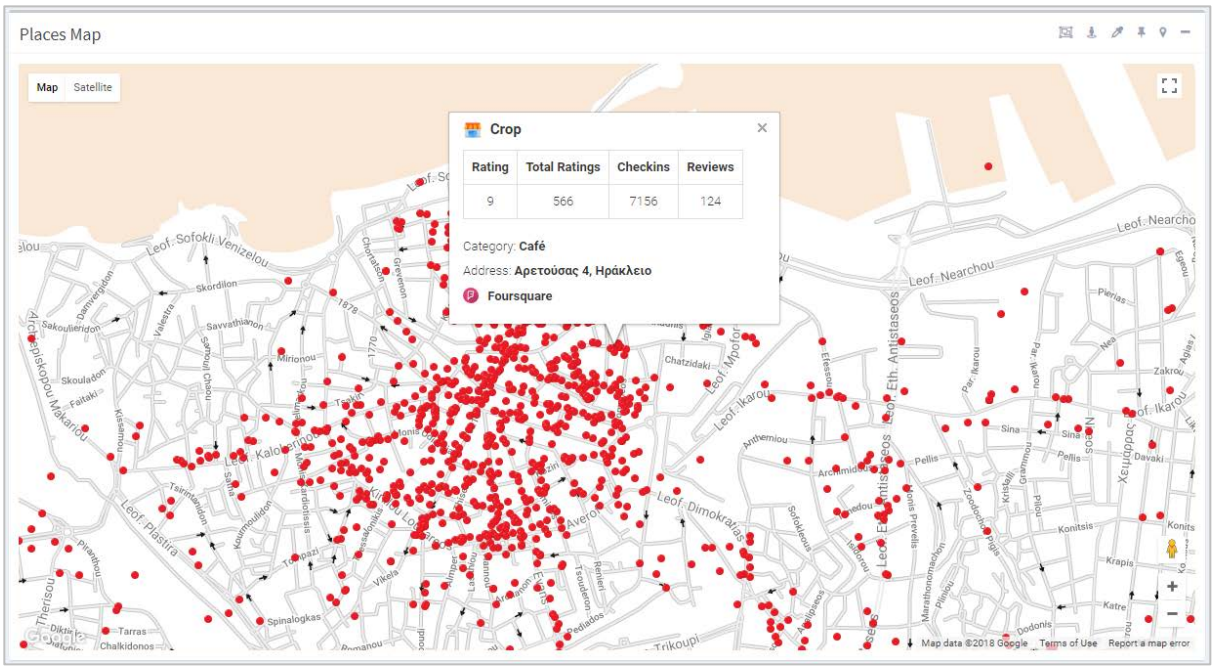


Figure 22 - Places' map

By observing the places' map, the stakeholders like local authorities can investigate what businesses exist near a subregion, road, plaza etc. and see their details, as well as the businesses can more easily find and monitor their competitors.

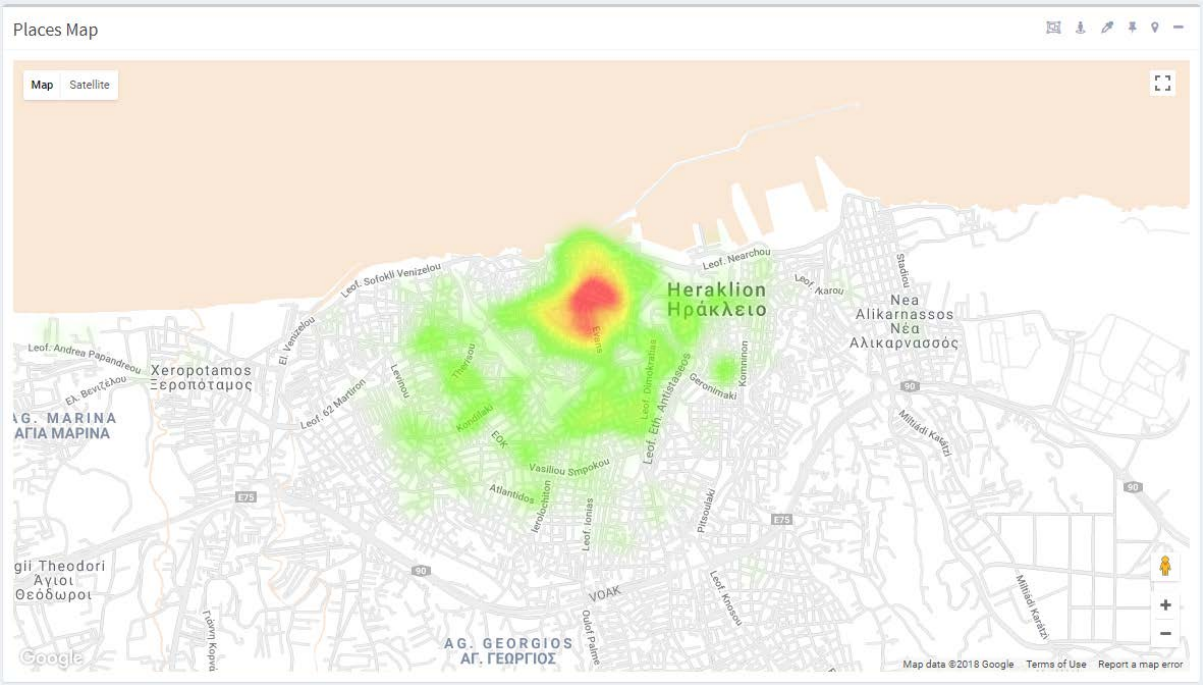


Figure 23 - Places' heatmap

In addition, the places presented in the map, are also displayed in a data table which can be ordered by rating, total ratings, check-ins, reviews, and region. Also, the user can search for businesses by writing specific terms in the search field, for example “Restaurants Heraklion” to see the restaurants in the region of Heraklion.

Regions' Places

Show 10 entries Search:

#	Place Name	Rating	Total Ratings	Checkins	Reviews	Categories	Region
1489	Chania Old Port (Παλιό Λιμάνι Χανίων)	9.6	1256	25404	125	Harbor / Marina	Chania
14	Crop	9	566	7156	124	Café	Heraklion
86	Central Park	8	426	9722	97	Café	Heraklion
38	Think Tank	8.7	342	6806	39	Wine Bar	Heraklion
73	Χάλαβρο Open Bar	8.2	328	4095	57	Bar	Heraklion
61	Liontaria Square (Πλατεία Λιονταριών)	8.4	327	14312	52	Plaza	Heraklion
15	Heraklion Archaeological Museum (Αρχαιολογικό Μουσείο Ηρακλείου)	9	302	2562	56	History Museum	Heraklion
1516	Ζάχαρη & Αλάτι	9.1	263	2842	103	Food	Chania
10	Ευκάλυπτος	9.2	252	3389	43	Bar	Heraklion
9	Hacienda	9.3	242	2267	45	Coffee Shop	Heraklion

Showing 1 to 10 of 2,412 entries Previous 1 2 3 4 5 … 242 Next

Figure 24 - Places' data table

Businesses types

The second visualization generated in business-level analytics page, is a pie chart per region of interest that indicates the number of businesses grouped by their type (Figure 25). Having this information (about the number of businesses per type), businesses can answer to questions such as *"is this region having businesses of my industry?"* and *"where to create my new store?"*.

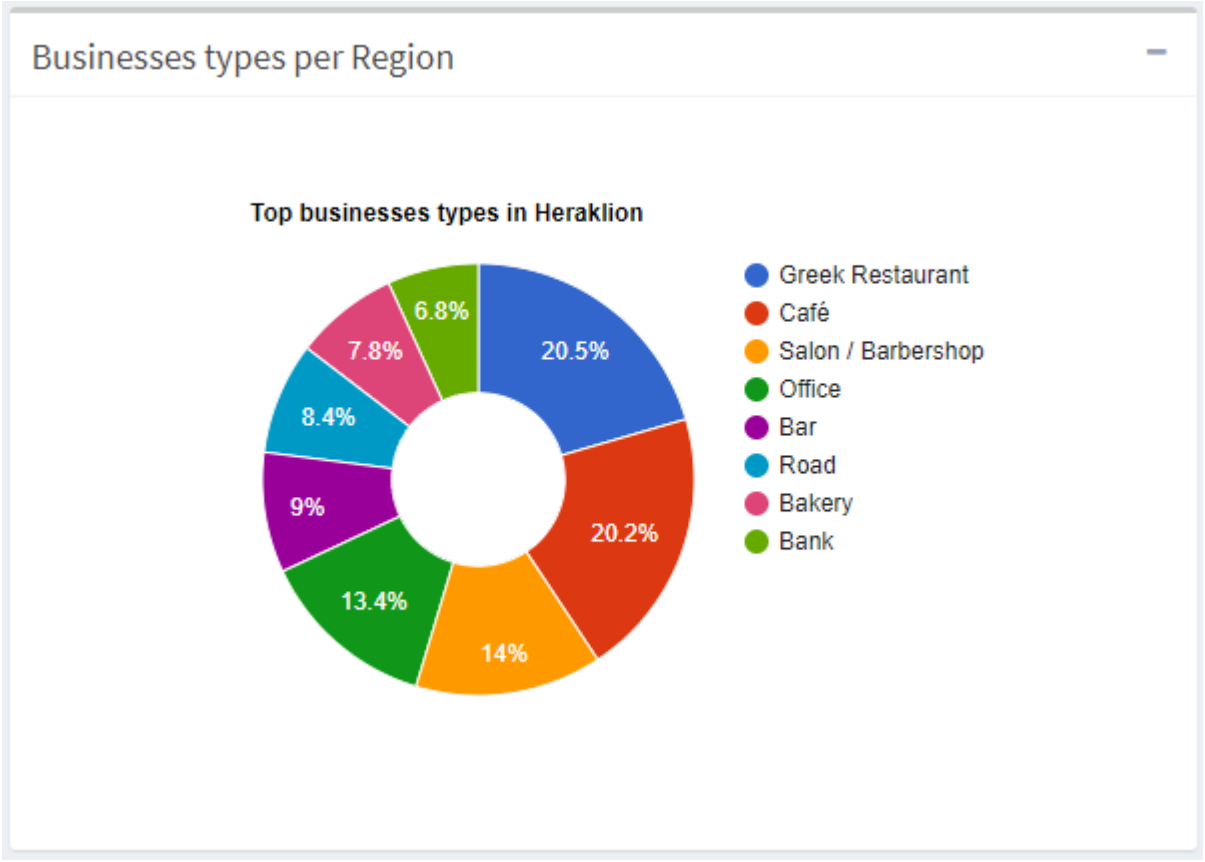


Figure 25 - Businesses types per region

Chapter 4 - THE CASE OF CRETE

Using location-based social network data is extremely beneficial for many parties including citizens, governments and it can be life-critical. Web 2.0 applications and tools transforming the role and behavior of travelers, make tourism stakeholders reorganize their operations and business models based on these data [70]. In that context, having a stable version of our system, we created a project inside our web application and started monitoring our regions of interest.

This thesis among other issues examines questions related to users' preferences like which is the most popular destination and which destination is more preferable [57] for the two largest cities of Crete region (Heraklion and Chania).

The translation of enormous amounts of data into valuable information for decision making is the first challenge and putting this information into a comprehensible form so that managers can process the given data is another challenge for them to conquer [2]. Using data visualizations, the extraction of valuable knowledge for stakeholders can be achieved [71]. The data obtained from the LBSN used in this study for three months (April to June 2018) are presented.

During the three months, 28K users posted 2.3K textual posts (Twitter posts & Foursquare reviews) and 121K photos/videos (on Instagram and Flickr). Using Google's Natural Language API, we analyzed these 2.3K textual posts to extract their sentiment. This resulted in 624 positive, 140 negative, and 1081 neutral posts. The sentiment of the rest posts was not detected since Google's sentiment analysis supports ten languages at this time.

Regarding social engagement, both textual posts and photos/videos reached 5.3M likes, 1.3M comments and 487 shares. Figure 26 displays the above numbers separated for each region.

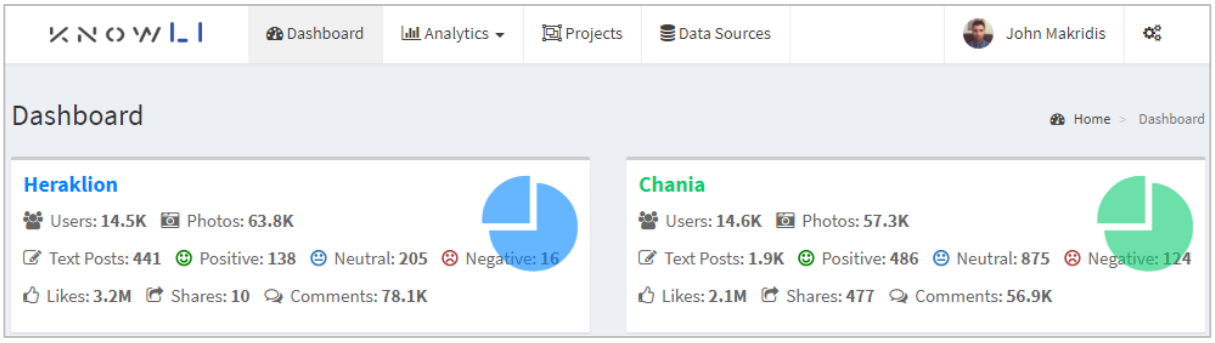


Figure 26 - Total users, posts, sentiment, photos and engagement per city

The sentiment analysis provides great insights into the sentiment of users in the region [56], while the negative posts can provide discovery of existing problems and directions for improvement. In the following data visualization, (Figure 27), the trend of the total posts of the users in the examined region (both cities) is presented.

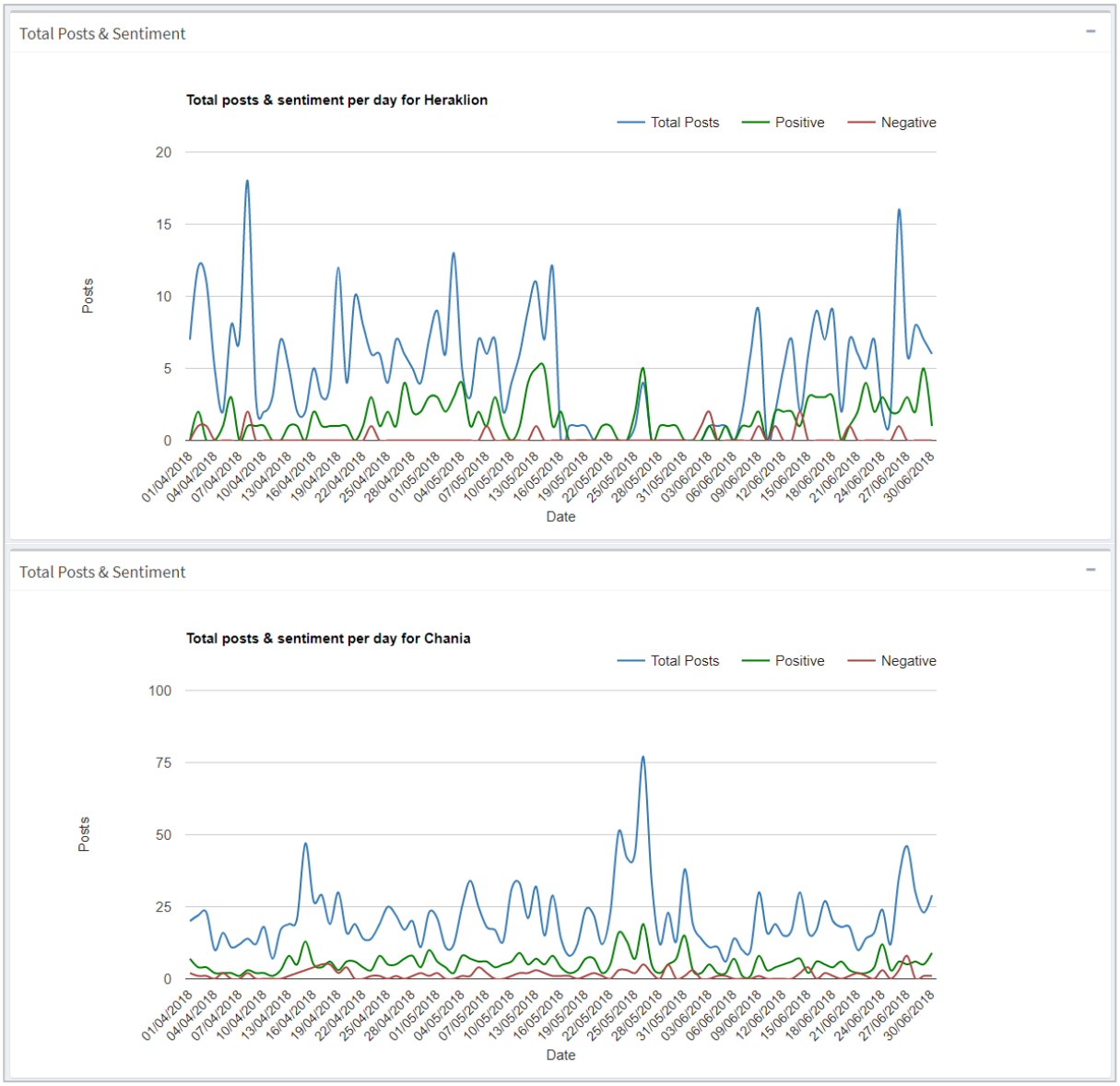


Figure 27 - Total posts and sentiment per day for Heraklion (top) and Chania (bottom)

In addition, the sentiment analysis of the text provides the feeling of the visitors derived from their posts on location-based social networks [55]. Therefore, the positive sentiment dominates during almost all the time period examined, while the negative posts are in lower levels. Through the trend

lines of the above figure, the trend of the visitors' posts is presented providing insights about their movements during specific days.

By displaying social networks' users' geo-tagged posts and photos as markers on the map, and by creating clusters of them, we can identify which are the most visited areas of a city.

In the following maps (Figure 28), the most popular places of Heraklion and Chania are presented with red markers. At first, the most popular places in Heraklion city are city's center, Koule Fortress, Natural History Museum of Crete, Heraklion Archaeological Museum, Port, and the local market of Heraklion (shopping streets). On the other side, the most popular places in Chania include the city's center, Old Venetian Harbor, Lighthouse and the local market of Chania.

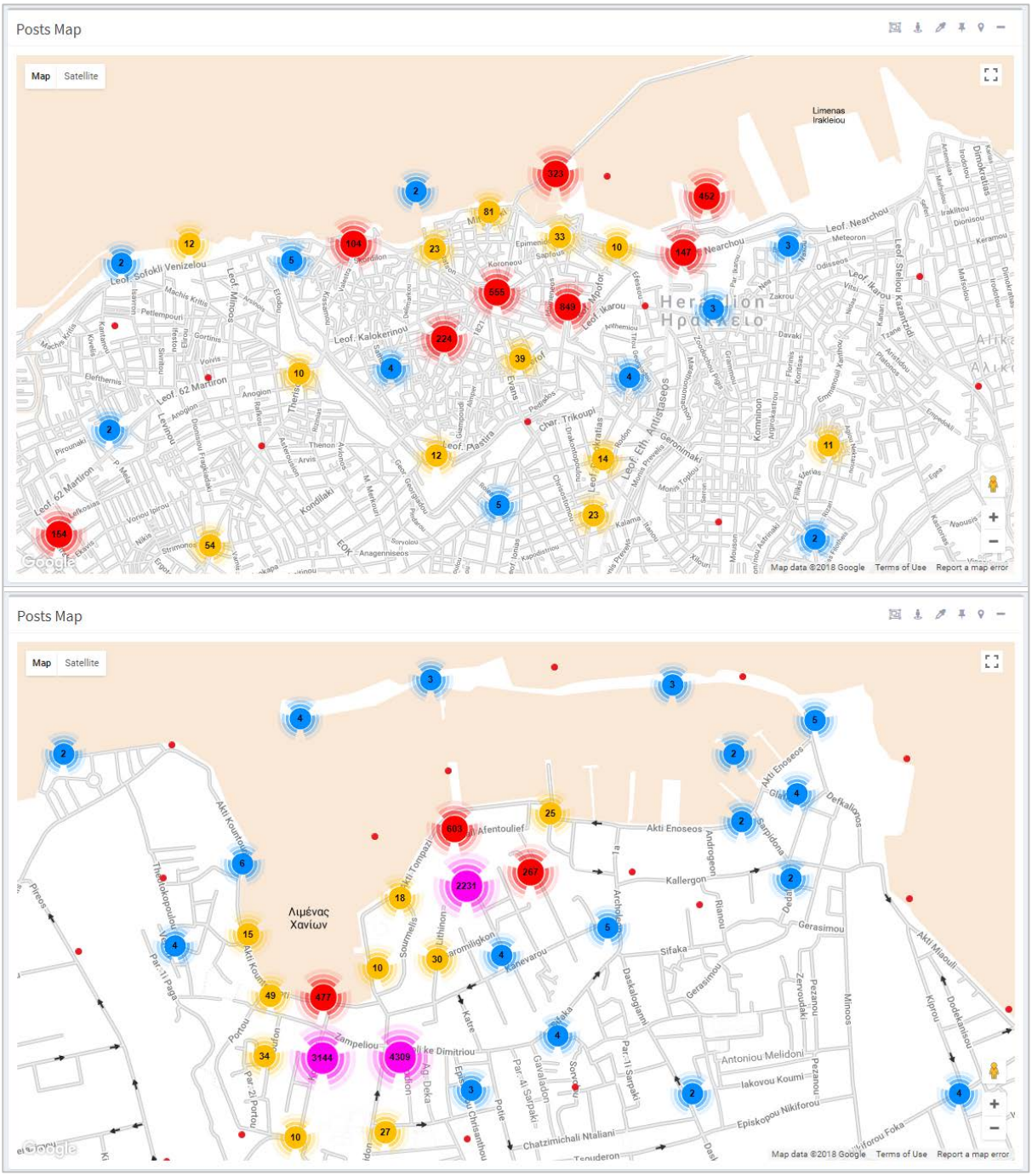


Figure 28 - Map with posts' clusters for Heraklion (top) and Chania (bottom)

Through the places' data visualization as markers on the map from their actual coordinates, we can identify which areas of two cities have the most places (places' pages created in Foursquare). By clustering these markers (Figure 29), we can observe the most places are in the center of the two

cities (red circles) or areas near it. The center of Heraklion has approximately 800 places, while the center of Chania has 300.

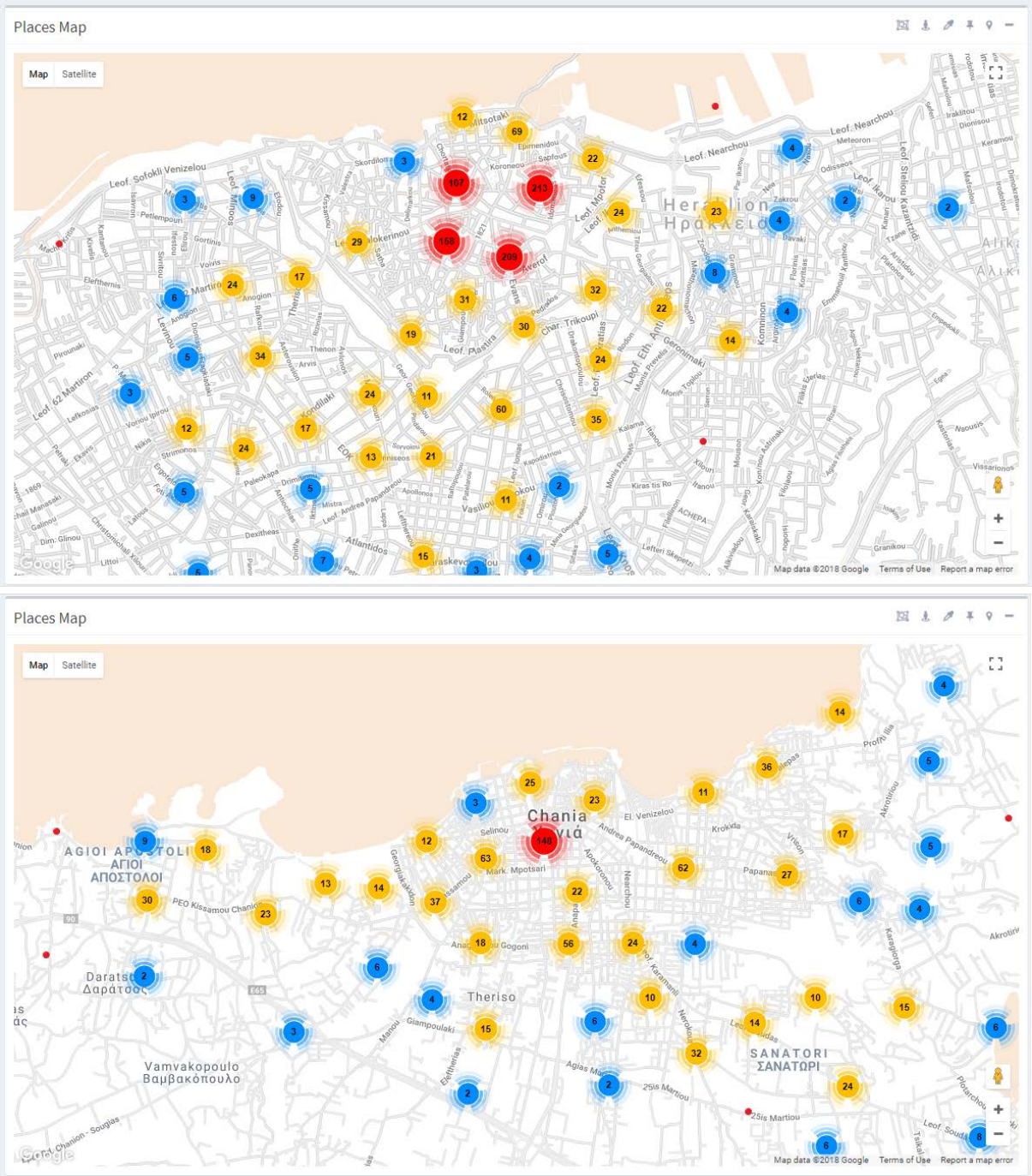


Figure 29 - Map with places' clusters for Heraklion (top) and Chania (bottom)

The above places (businesses, sights, points of interest, roads etc.) are also distributed into categories. The following pie charts (Figure 30) shows the top business types per city. In Heraklion, 20.5% of

places are Greek restaurants (66 entries), 20.2% cafe (65 entries) and 14% salons/barber shops (45 entries), while in Chania the top three categories are Hotels with 15.4% (27 entries), automotive 14.3% (25 entries), and cafe 13.1% (23 entries).

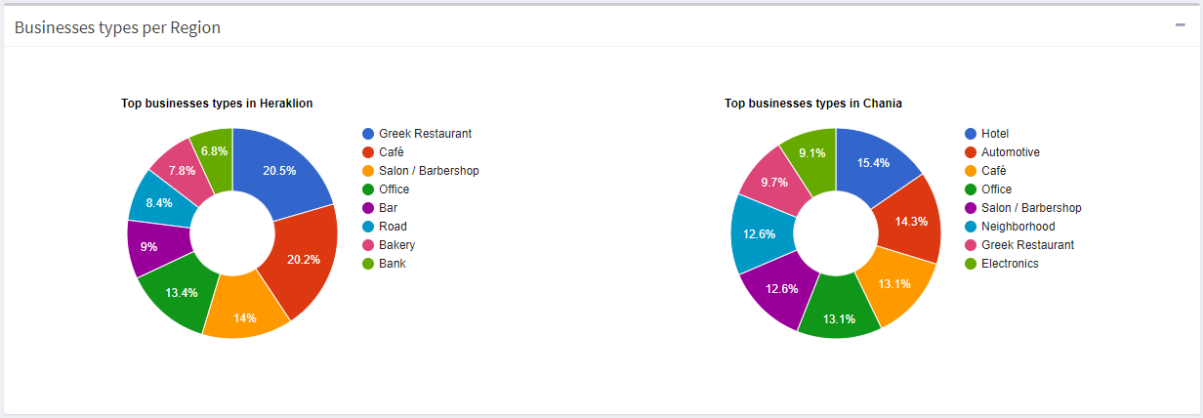


Figure 30 - Top business types for Heraklion (left) and Chania (right)

As shown in Figure 31 a line chart is generated in our dashboard in which the highest points regarding social engagement (sum of likes, comments and shares of posts) can be observed for each region of interest and selected date range. By investigating in more detail the posts that are published on the days with the highest social engagement, we can identify (1) why users interacted with each other, (2) which are the most popular, important or most-discussed topics, and (3) who are the most influencing people for each region.

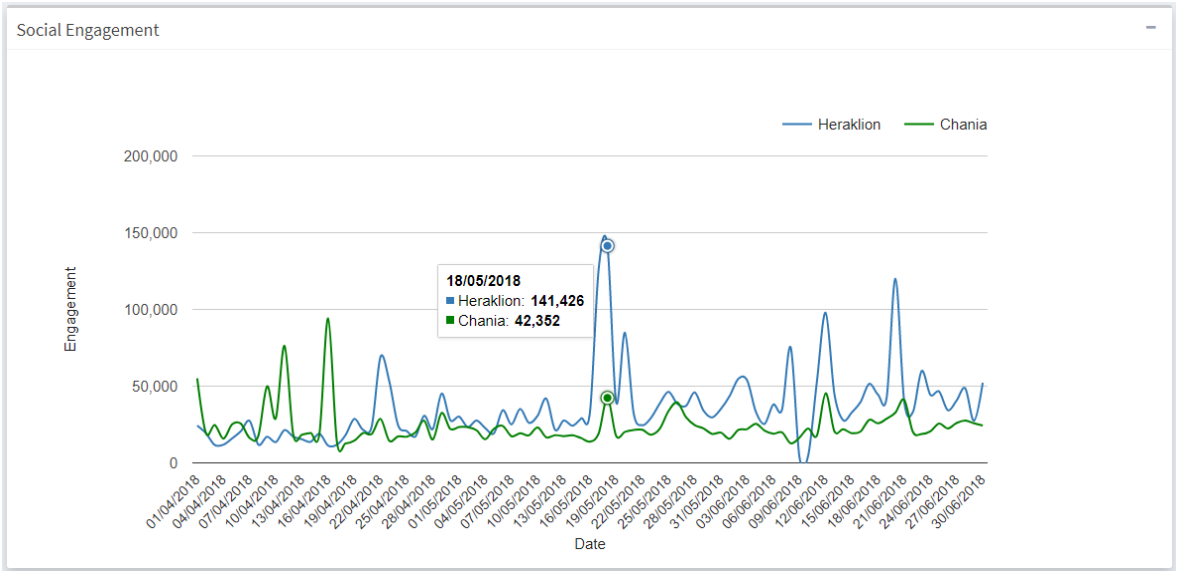


Figure 31 - Social engagement line chart for both cities

In the above social engagement line chart, we can observe there are three high points in terms of social engagement for Heraklion and one for Chania. On 17th and 18th of May 2018, a London-based make-up artist with 1.1M followers uploaded from Heraklion two photos on Instagram reaching 107K and 98K likes respectively, while on 20th of June, a Czech-based blogger-youtuber with 419K followers on Instagram, posted a photo of a beach in Heraklion during her vacations.

On the other hand, an Indian singer with 11M followers uploaded on Instagram a teaser of her new video clip filmed in Chania reaching the same day 217K video views and 77K likes. It can be observed that an influencer's posts could be a significant factor for a destination's brand awareness since influencer marketing is considered a viral marketing technique contributing in building brand awareness [72].

The following bar charts (Figure 32) shows the total posts and photos per region of interest and channel (source) for the given date range. By observing these two charts and comparing the posts and photos, we can identify in which social network users are publishing their textual posts and photos respectively. This could be an important metric in the planning phase of a social media marketing strategy since it gives to the stakeholders the clarification of which social network they should use to publish their content to increase their users' engagement.

Twitter was most widely used than Foursquare and Instagram against Flickr for photos respectively. Therefore, it seems that most users for both cities used Twitter and Instagram networks for sharing their experiences and communicating in social media. As we can see, Chania holds the 80% of the total textual posts while in Heraklion uploaded the 53% of the total photos.

Figure 32 - Total textual posts and photos per channel and region

Another useful data visualization is the distribution of the acquired data in days of the week based on the date posts are created by the users. This information helps businesses identifying the popular days of the week for publishing content in their digital marketing channels.

Figure 33 shows two histograms with the total posts and photos per day of the week for the two cities. During the period examined, Tuesday was the day with the most Twitter posts and Foursquare reviews in Heraklion, in contrast with Chania where users were more active on Saturday. However, for uploading their photos, users were online mostly on Tuesday, Wednesday, and Thursday for Heraklion, while in Chania the photos and videos were uploaded mostly on Sundays.

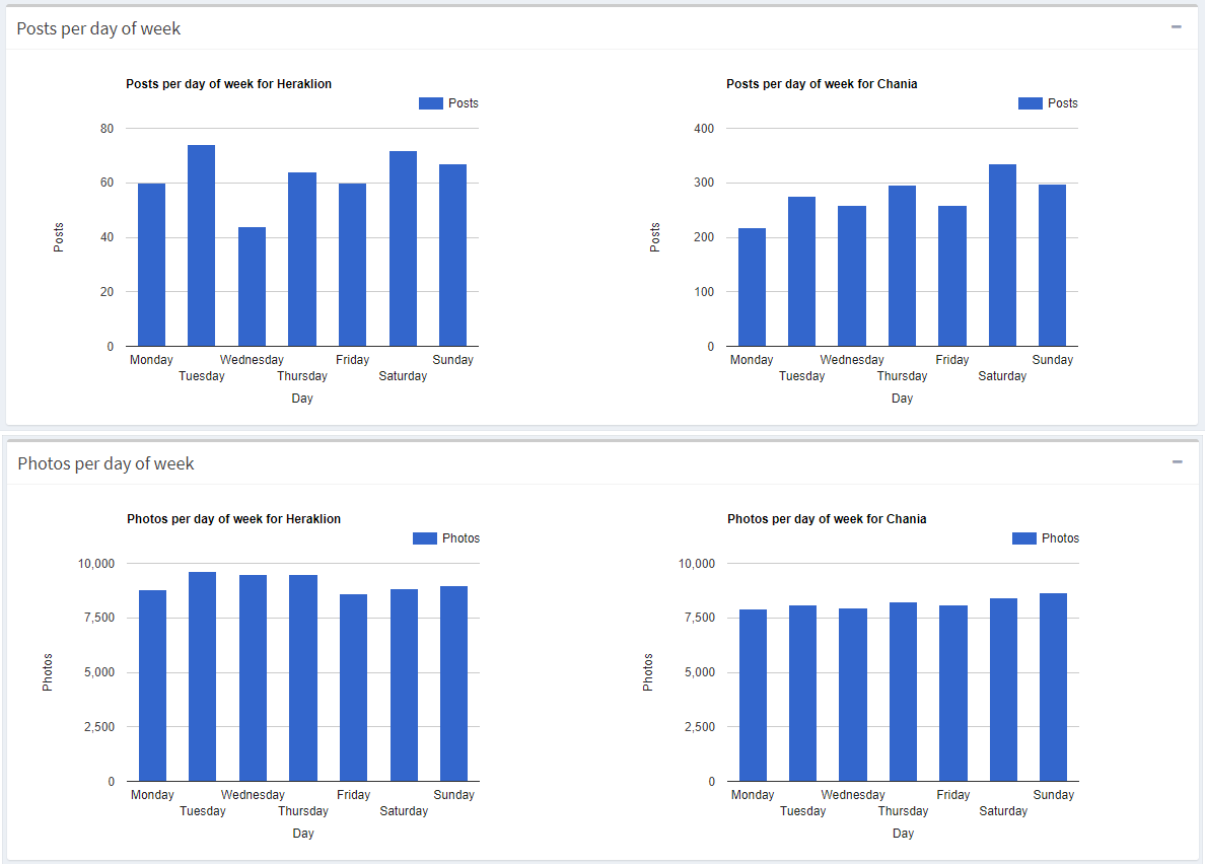


Figure 33 - Textual posts and photos per day of week for Heraklion (left) and Chania (right)

Another useful information also extracted from posts' creation date and can be combined with the above graph of top days of the week, is the hours users are more active in social networks. For instance, by combining this information, advertising companies can create more targeted advertisements as they have the knowledge of what days and hours users are online, increasing in that way the possibility of users will watch these advertisements.

Figure 34 shows the hours users generated the most content in the two cities for the selected date range. In Heraklion, the most posts were created in morning hours 08:00 to 10:00, while photos are uploaded mostly in evening hours from 19:00 to 22:00. In contrast, both textual posts and photos in Chania were published mostly in evening hours between 18:00 and 22:00.

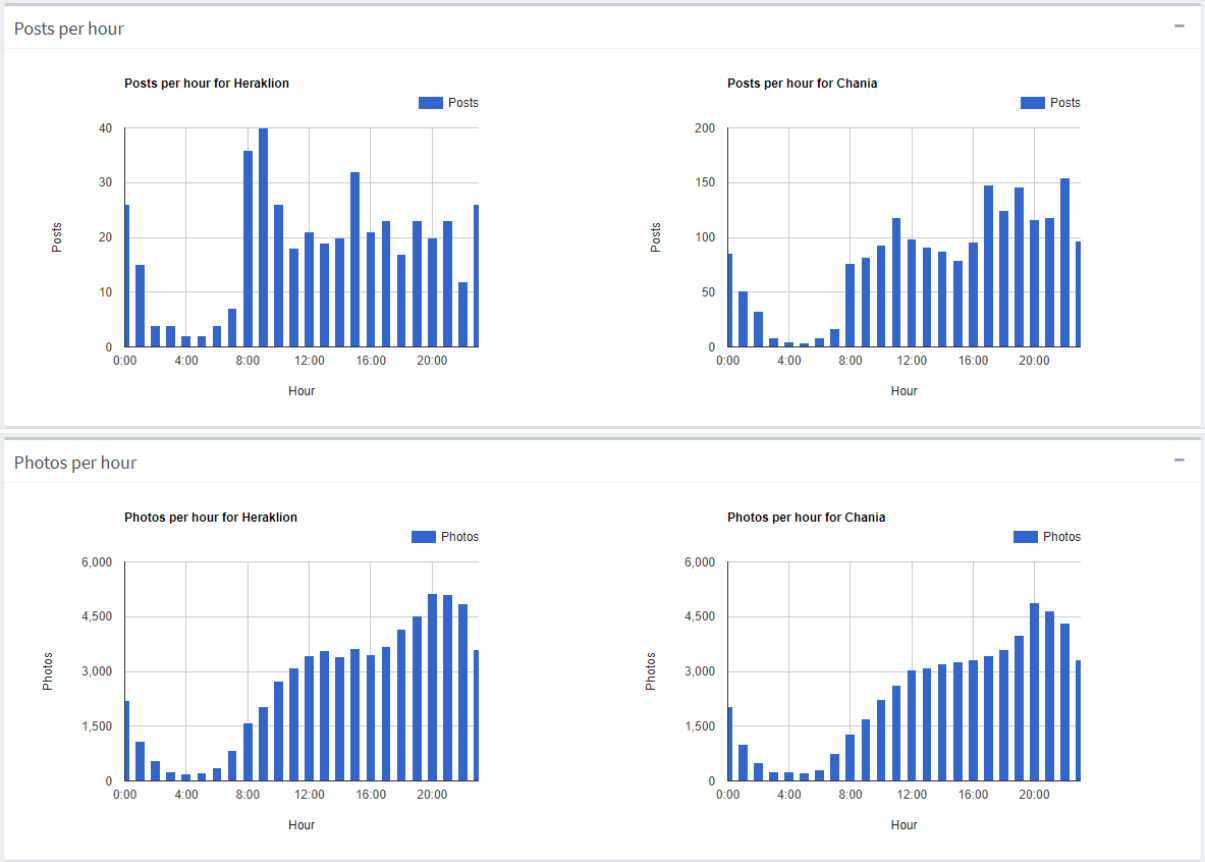


Figure 34 - Textual posts and photos per hour for Heraklion (left) and Chania (right)

As described above, the analysis of post's text can provide significant insights for businesses and local authorities. The following three data visualizations are showing the outcomes from the analysis of post's sentiment and entities, as well as the languages detected for each text.

These visualizations were generated to clarify us about (1) whether the users' impressions are either positive or negative, (2) what the users are talking about in social networks, e.g. which entities or topics, and (3) from where these users are from according to their language.

Figure 35 shows the classification of posts' sentiment for the two cities. We observe that negative posts are in lower levels than positives for both cities. By identifying the themes of the positive and negative posts, we can provide to the stakeholders useful insights that allow them to strengthen their services and build brand awareness, as well as to address and prevent issues.

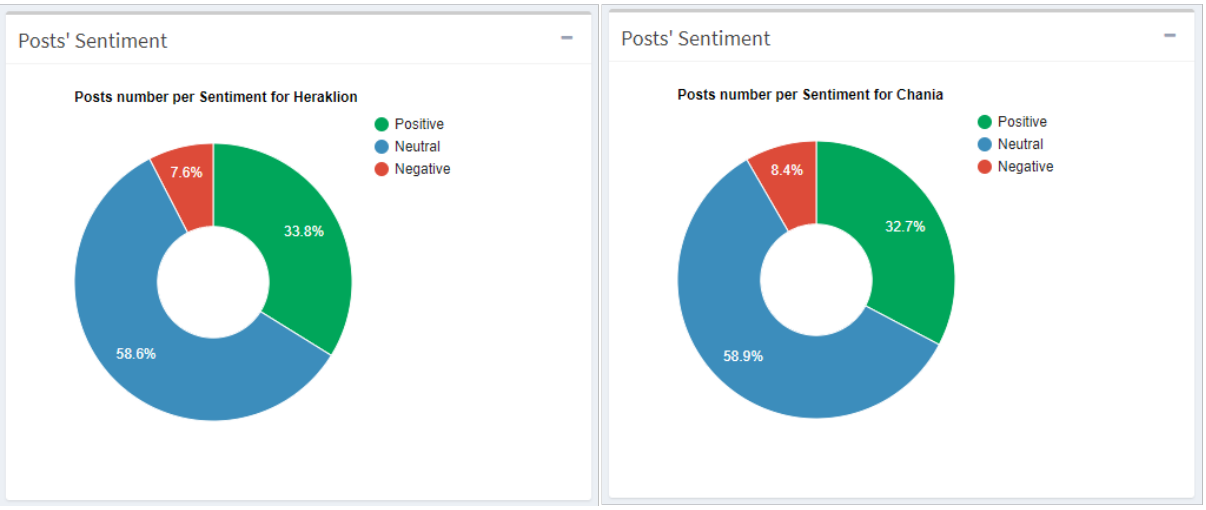


Figure 35 - Pie chart with textual posts' sentiment for Heraklion (left) and Chania (right)

By investigating in more detail the classification of sentiment from pie chart with executing database queries, we can identify the texts that are contained in each class and analyze them further. Our findings provided valuable knowledge about the visitors' impressions in the examined cities (Table 5).

	Heraklion	Chania
Most positive reviews	Places' beauty, quality of services, food tastefulness	Places' beauty, customer service
Most negative reviews	Customer service, outdoor activities	Quality of services and products

Table 5 - Visitors' most positive and negative impressions

The most positive reviews for Heraklion was about places' beauty, quality of services and food tastefulness, while the most negative were about customer service (rude employees) and poor quality in outdoor activities. On the other hand, the most positive reviews for Chania was about places' beauty and customer service while the most negative was about the quality of hotel services and food.

Apart from the users' impressions, a pie chart was generated for each city to show us the entities that are mentioned most times in their posts. By observing the entities' pie chart (Figure 36), the stakeholders can take advantage of this information to make strategic decisions. For example, local authorities can monitor their regions of interest to identify whether visitors or local people are talking about a scheduled event or not and provide to them the capability of promoting it better in the social networks.

We observe users have mentioned the same entities in almost the same levels between the two cities. In most cases, users have mentioned one or more locations in their posts with 41.6% for Heraklion and 40.5% for Chania, while the *person* entity followed with 35.2% and 36.6% respectively.

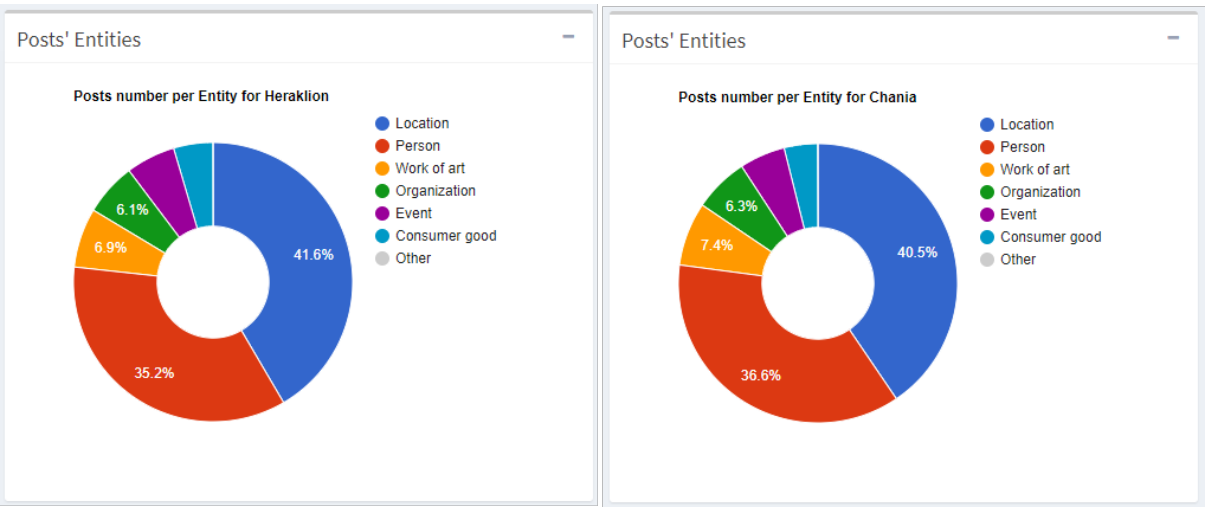


Figure 36 - Pie chart with textual posts' entities for Heraklion (left) and Chania (right)

Having the impressions and most mentioned entities from users' posts, the next clarification that is provided from users' texts is user's origin. Apart from English, a language spoken by people in many countries, we can approximately classify the users' origin by the detecting their post's language.

Figure 37 shows a pie chart for each region that visualizes the top languages from users' posts. We observe that English and Greek are the languages with the most posts. If we exclude these two languages, we can then identify visitors' origin since in the next positions the Spanish, Russian, French, Turkish, Estonian and other languages are followed.

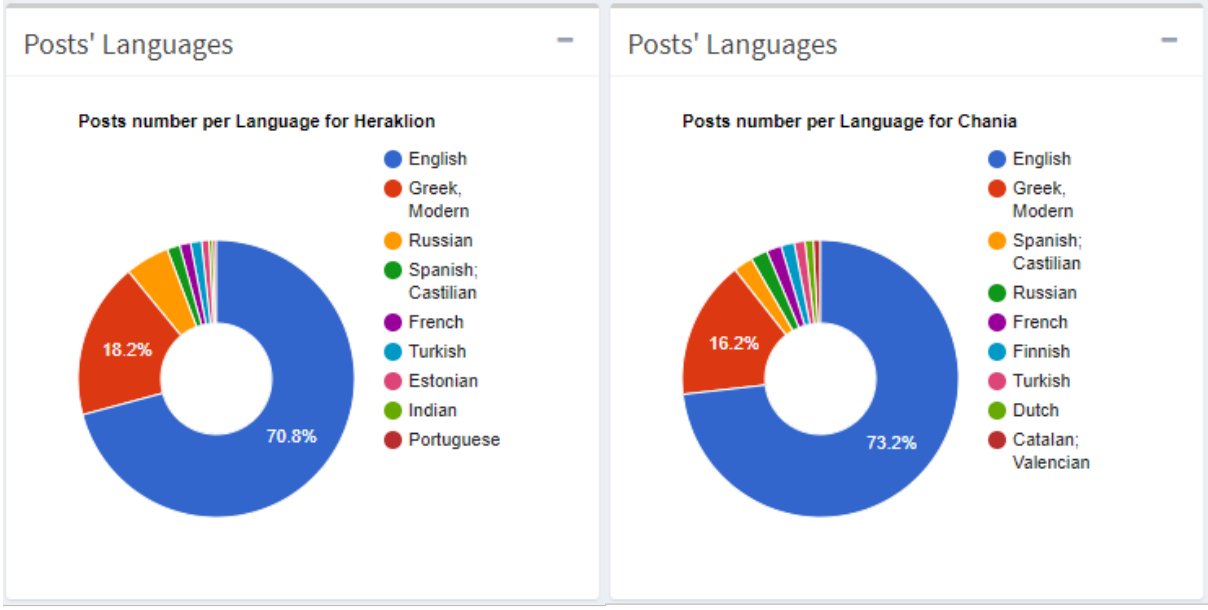
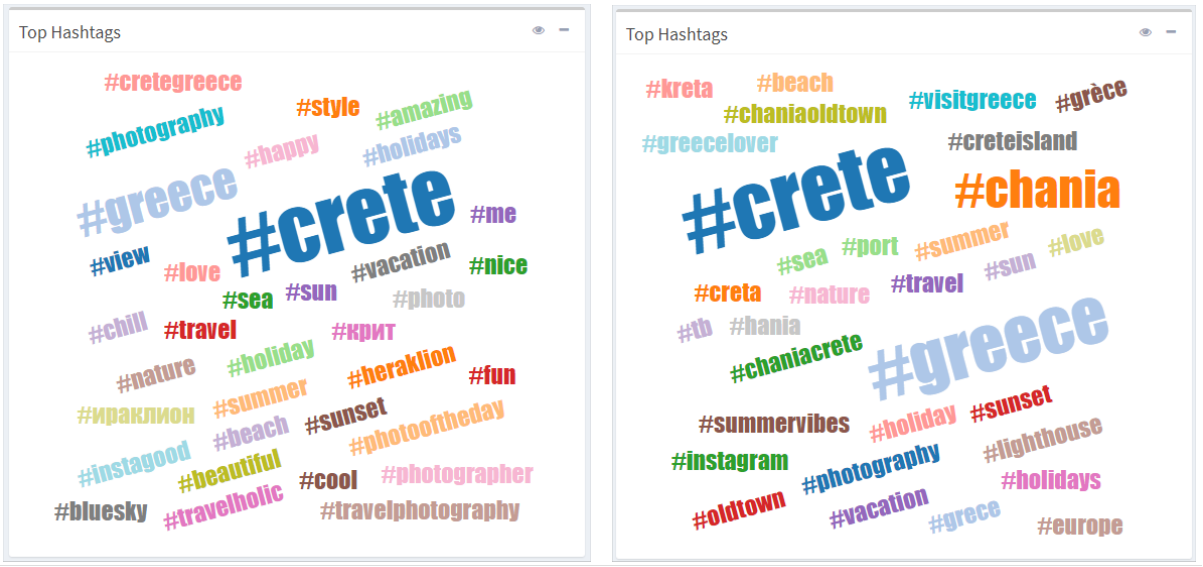


Figure 37 - Posts' languages for Heraklion (left) and Chania (right)

Hashtags have become a way for people and brands to create discussions, interact with each other, promote their services and products to massive amounts of social networks users, and create successful hashtag campaigns.

Figure 38 shows a word cloud with the top hashtags extracted from users' posts published in social networks and fetched in our database. The hashtags that are visualized with larger font size indicate they have a higher frequency among the others. In addition, by clicking on the eye icon of the top hashtags box, a popup appears that displays a table with the hashtags ordered by their frequency.

We observe that the top two hashtags for both cities are #crete and #greece. For Heraklion, the next most-used hashtags were #heraklion, #summer, #sea, and #holidays, in contrast with Chania where the most-used hashtags were #chania, #summer, #chaniacrete, and #sea.



Hashtags

Hashtag	Count
#crete	47
#greece	39
#heraklion	17
#summer	14
#sea	13
#holiday	12
#travel	12
#love	11
#sun	9
#beach	9

Hashtags

Hashtag	Count
#crete	44
#greece	42
#chania	37
#summer	14
#chaniacrete	12
#sea	11
#sunset	9
#holiday	8
#travel	8
#sun	7

Figure 38 - Word cloud and table with the top hashtags for Heraklion (left) and Chania (right)

The same method with top hashtags cloud was applied to fetched posts' text. This procedure allows us to discover which are the most-used words in users' text without reading each post separately. The combination of words cloud with the entities analysis presented above, can provide us the knowledge of what users are mentioned in their posts (which entities), as well as the content (words) of these mentions.

Figure 39 shows a word cloud with the most-used words in users' posts for each city. In addition, the words cloud can be displayed as a table showing these words ordered by their frequency inside the texts.

We can observe that some words are appearing in users' text for both cities. The most-used words including Crete, Greece, Chania, and Heraklion. Since our examined period (April 2018 to June

2018) was inside the tourist season, we notice that the rest of most-used words was related to this, since visitors mentioned words like travel, summer, holiday and beach in their posts.

Words

Word	Count
crete	133
greece	111
summer	45
heraklion	40
love	30
travel	29
holiday	27
beach	25
beautiful	23
kreta	21

Words

Word	Count
chania	147
crete	140
greece	124
summer	44
holiday	32
love	28
sunset	27
travel	25
grece	23
with	21

Figure 39 - Word cloud and table with the top used words for Heraklion (left) and Chania (right)

Chapter 5 - CONCLUSION & DISCUSSION

Summary

It is widely accepted that big data analytics entered and established its existence across organizations and became one of the fastest-growing segments of business intelligence. Modern organizations migrate from human-centered descriptive and diagnostic analytics towards the advanced, machine-centered predictive and prescriptive approaches. Big data analytics can transform the way organizations operate and increase the business adaptivity to the rapidly changing environment. Using data processing tools and advanced analytics techniques, the organizations can predict and observe future trends, build econometric models and identify customer needs. In addition, this approach enables organizations to conduct the necessary changes to keep away from crises, shortens the time of decision-making process and develop fully automated actionable systems.

Big data analytics supported with machine learning is used in a wide range of applications across multiple domains and industries including healthcare, insurance, marketing, and location intelligence. Sales and marketing departments can benefit by using clustering techniques in market segmentation and optimization of marketing campaigns.

In summary, the purpose of this dissertation was to design and develop a web application for discovering knowledge through actual user-generated content on location-based social networks and emphasizing the significant impact of “where”. Through the acquisition of geospatial data from four popular LBSNs including Twitter, Foursquare, Instagram and Flickr our research questions were answered. Our objective was to answer what are the visitors’ behavior, impressions, and preferences for tourist destinations, and what decisions local authorities and businesses can take to make a more efficient promotion of these tourist destinations, to improve the existing facilities and activities, and to create new experiences for attracting the interest of more potential visitors. In order to answer these questions and find out what insights can be extracted from the analysis of LBSN data, we used the case study of two cities - Heraklion and Chania - from Crete (Greece).

Through the representation of users’ posts as markers on a map based on post’s coordinates, we identified the most visited places in both cities. These location-tagged posts were distributed and visualized in four bar charts representing the days and hours with most posts (when the users were more active), and the social networks with the most textual posts and photos (in which social network these posts published - the most popular social network) for each city.

The textual posts from Twitter and reviews from Foursquare were analyzed using Google's natural language processing API to extract their sentiment, entities and language. We decided to apply this analysis only on Twitter and Foursquare as their texts are more meaningful than Instagram and Flickr which are platforms for sharing mostly media such as photos and videos. The results from sentiment and entity analysis were visualized with three pie charts to present the users' impressions in two cities (either positive, negative or neutral), the entities that were mentioned most in their posts (e.g. locations, persons, events, organizations etc.) and the texts' languages to identify the languages with most posts and approximately identify the users' origin.

Since hashtags are widely used for tagging text or media with a topic, place etc., we extracted the hashtags from all posts (both textual and media) by writing a custom algorithm to discover the themes that mostly used in users' posts. Similarly to hashtags, we wanted to represent the most frequently used words in users' posts to understand in more depth what the users are talking about, thus both of this information were visualized with two words clouds respectively.

Influencer marketing or e-word of mouth marketing can contribute to enhancing destination attractiveness or destination branding since influencers can spread messages affecting communities in the digital world. To identify the most influencing users in two cities, we generated a line chart that represents the social engagement per day for each region of interest. The analysis of social engagement can be used for the evaluation of the attractiveness of tourism experiences associated with an event and can be a significant factor for more personalized offers focused on customer satisfaction. Finally, to identify the areas with the most places (businesses, venues, points of interests etc.) we visualized them on a map using their actual coordinates. For each of these places, several important information is displayed such as the average rating, number of total ratings, and the number of check-ins and users' reviews. In addition, a pie chart was generated for each city that represents the distribution of places into business categories. This gave us the knowledge of which categories have the most places and how many places exist on each category.

Our findings for the examined two cities - Heraklion and Chania – for the period of April 1st, 2018 to June 31st, 2018 are presented in the following table.

	Heraklion	Chania
Total users	14.5K	14.6K
Total posts	441 textual posts & 63.8K photos	1.9K textual posts & 57.3K photos
Social engagement	3.2M likes + 10 shares + 78.1K comments	2.1M likes + 477 shares + 56.9K comments
Popular places	City's center, Koule Fortress, Natural History Museum of Crete, Heraklion Archaeological Museum, Port, local market of Heraklion	City's center, Old Venetian Harbour, Lighthouse, local market of Chania
Textual posts' sentiment	Positive 33.8%, Neutral: 58.6%, Negative: 7.6%	Positive 32.7%, Neutral: 58.9%, Negative: 8.4%
Most positive themes	Places' beauty, quality of services, food tastefulness	Places' beauty, customer service
Most negative themes	Customer service, outdoor activities	Quality of services and products

Most mentioned entities	location 41.6%, person 35.2%	location 40.5%, person 36.6%
Top languages	English, Greek, Russian, Spanish, French, Turkish	English, Greek, Spanish, Russian, French, Finish
Top hashtags	#crete, #greece, #heraklion, #summer, #sea	#crete, #greece, #chania, #summer, #chaniacrete
Most-used words	Crete, Greece, summer, Heraklion, love	Chania, Crete, Greece, summer, holiday
Days with most posts	**Texting:** Tuesday, **Media sharing**: Tuesday, Wednesday, Thursday	**Texting:** Saturday, **Media sharing**: Sunday
Hours with most posts	**Texting:** morning hours from 08:00 to 10:00, **Media sharing:** evening hours from 19:00 to 22:00	**Texting:** evening hours from 18:00 to 22:00, **Media sharing:** evening hours from 18:00 to 22:00
Top businesses types	Greek restaurants 20.5%, Cafe 20.2%, Salons/barber shops 14%	Hotels 15.4%, Automotive 14.3%, Cafe 13.1%
Top social networks	**Twitter** for texting and **Instagram** for media sharing	**Twitter** for texting and **Instagram** for media sharing

Table 6 - Case study findings

Achievements and future work

By implementing the KnowLI platform, in this thesis, we achieved the knowledge discovery from actual user-generated content in location-based social networks. Through descriptive analytics and data visualizations all the information around persons and their posts for both cities were identified, including what a person was talked about, what hour and day this was published, in which social network, where it was published (in which location) and by whom.

It would be desirable if the proposed framework would be extended by applying predictive and prescriptive analytics on spatial data in order to predict the movements of visitors and local people inside the defined regions of interest, and provide alerts or reports for what decisions the stakeholders should take to improve the existing infrastructures and activities, create more efficient promotions of the tourist destinations, attract more visitors, and increase the satisfaction of the local people.

Bibliography

[1] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013.

[2] H. Chen, R. H. L. Chiang, and V. C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact,” *MIS Quarterly: Management Information Systems*, vol. 36, no. 4, pp. 1165–1188, 2012.

[3] K. Vassakis, E. Petrakis, and I. Kopanakis, “Big Data Analytics: Applications, Prospects and Challenges,” in *Lecture Notes on Data Engineering and Communications Technologies*, 2017, pp. 3–20.

[4] M. M. Rathore, M. Mazhar Rathore, A. Paul, A. Ahmad, M. Imran, and M. Guizani, “Big data analytics of geosocial media for planning and real-time decisions,” in *2017 IEEE International Conference on Communications (ICC)*, 2017.

[5] Y. Zheng, “Location-Based Social Networks: Users,” in *Computing with Spatial Trajectories*, 2011, pp. 243–276.

[6] M. J. Chorley, R. M. Whitaker, and S. M. Allen, “Personality and location-based social networks,” *Comput. Human Behav.*, vol. 46, pp. 45–56, 2015.

[7] “Most famous social network sites worldwide as of April 2018, ranked by number of active users (in millions),” *Statista.com*, Apr-2018. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed: 15-May-2018].

[8] S. Akter, M. Bhattacharyya, S. F. Wamba, and S. Aditya, “How does Social Media Analytics Create Value?,” *Journal of Organizational and End User Computing*, vol. 28, no. 3, pp. 1–9, 2016.

[9] D. Reinsel, J. Gantz, and J. Rydning, “Data Age 2025: The Evolution of Data to Life-Critical,” International Data Corporation, Apr. 2017.

[10] M. Cox and D. Ellsworth, “Application-controlled demand paging for out-of-core visualization,” in *Proceedings. Visualization ’97 (Cat. No. 97CB36155)*.

[11] P. D. Vecchio, P. Del Vecchio, G. Mele, V. Ndou, and G. Secundo, “Creating value from Social Big Data: Implications for Smart Tourism Destinations,” *Inf. Process. Manag.*, 2017.

[12] P. D. Vecchio, P. Del Vecchio, G. Mele, V. Ndou, and G. Secundo, “Creating value from Social Big Data: Implications for Smart Tourism Destinations,” *Inf. Process. Manag.*, 2017.

[13] W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *The Journal of Strategic Information Systems*, vol. 26, no. 3, pp. 191–209, 2017.

[14] W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *The Journal of Strategic Information Systems*, vol. 26, no. 3, pp. 191–209, 2017.

[15] Sruthika S, S. Sruthika, and N. Tajunisha, "A study on evolution of data analytics to big data analytics and its research scope," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015.

[16] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, 2015.

[17] R. Markle and A. Fiebelman, Eds., *Building an Effective Data-Driven Business*, vol. 40, no. 3. 2015.

[18] A. McAfee and E. Brynjolfsson, Eds., *Big Data: The Management Revolution*, vol. 90, no. 10. Harvard Business Review, 2012.

[19] S. LaValle, E. Lesser, R. Shockley, M. Hopkins, and N. Kruschwitz, Eds., *Big Data, Analytics and the Path From Insights to Value*, vol. 52, no. 2. MIT Sloan Management Review, 2011.

[20] J. Manyika *et al.*, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey, 2011.

[21] T. A. Runkler, *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. Springer, 2016.

[22] N. Elgendy and A. Elragal, "Big Data Analytics in Support of the Decision Making Process," *Procedia Comput. Sci.*, vol. 100, pp. 1071–1084, 2016.

[23] Gartner, "Advanced Analytics," *Gartner IT Glossary*, 2017. [Online]. Available: https://www.gartner.com/it-glossary/advanced-analytics/. [Accessed: 13-Dec-2017].

[24] S. H. Kaisler, J. Albert Espinosa, F. Armour, and W. H. Money, "Advanced Analytics -- Issues and Challenges in a Global Environment," in *2014 47th Hawaii International Conference on System Sciences*, 2014.

[25] Gartner, "Gartner Says Advanced Analytics Is a Top Business Priority," *Gartner*, 21-Oct-2014. [Online]. Available: https://www.gartner.com/newsroom/id/2881218. [Accessed: 07-Feb-2018].

[26] IBM, "Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics," IBM, Oct. 2013.

[27] P. Vashisht and V. Gupta, "Big data analytics techniques: A survey," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2015.

[28] B. Neese, "Business Solutions & Big Data, Part 1: Descriptive Analytics," *Southeastern University*, 15-Feb-2016. [Online]. Available: https://online.seu.edu/descriptive-analytics/. [Accessed: 02-Jun-2018].

[29] P. Hassani, "An Insight into 26 Big Data Analytic Techniques," *Systweak Softwares*, 29-Nov-2016. [Online]. Available: https://blogs.systweak.com/2016/11/an-insight-into-26-big-data-analytic-techniques-part-1/. [Accessed: 10-Jun-2018].

[30] J. Cranshaw, J. I. Hong, and N. Sadeh, "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City," *International AAAI Conference on Weblogs and Social Media The*, pp. 58–65, 2012.

[31] Q. Wei-ning and Z. Ao-ying, "Analyzing popular clustering algorithms from different viewpoints," *J. Softw. Maint. Evol.: Res. Pract.*, pp. 1382–1394, 2002.

[32] A. M. Abirami and A. Askarunisa, "Sentiment analysis model to emphasize the impact of online reviews in healthcare industry," *Online Information Review*, vol. 41, no. 4, pp. 471–486, 2017.

[33] L. Zhang and B. Liu, "Sentiment Analysis and Opinion Mining," in *Encyclopedia of Machine Learning and Data Mining*, 2016, pp. 1–10.

[34] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.

[35] A. Sapountzi and K. E. Psannis, "Social networking data analysis tools & challenges," *Future Gener. Comput. Syst.*, vol. 86, pp. 893–913, 2018.

[36] X. Wang, X. Zhou, and S. Lu, "Spatiotemporal Data Modeling and Management: A Survey," in *Proceedings 36th International Conference on Technology of Object-Oriented Languages and Systems. TOOLS-Asia 2000*, Xi'an, China.

[37] R. Haining, "Spatial Autocorrelation," in *International Encyclopedia of the Social & Behavioral Sciences*, 2015, pp. 105–110.

[38] N. C. Mithun, T. Howlader, and S. M. Mahbubur Rahman, "Video-based tracking of vehicles using multiple time-spatial images," *Expert Syst. Appl.*, vol. 62, pp. 17–31, 2016.

[39] D. Läpple, A. Renwick, J. Cullinan, and F. Thorne, "What drives innovation in the agricultural sector? A spatial analysis of knowledge spillovers," *Land use policy*, vol. 56, pp. 238–250, 2016.

[40] H. Tyralis, N. Mamassis, and Y. N. Photis, "Spatial analysis of the electrical energy demand in Greece," *Energy Policy*, vol. 102, pp. 340–352, 2017.

[41] W. Tu, R. Cao, Y. Yue, B. Zhou, Q. Li, and Q. Li, "Spatial variations in urban public ridership derived from GPS trajectories and smart card data," *J. Transp. Geogr.*, vol. 69, pp. 45–57, 2018.

[42] S. Giebultowicz, M. Ali, M. Yunus, and M. Emch, "A comparison of spatial and social clustering of cholera in Matlab, Bangladesh," *Health Place*, vol. 17, no. 2, pp. 490–497, Mar. 2011.

[43] Z. Wang, X. Ye, J. Lee, X. Chang, H. Liu, and Q. Li, "A spatial econometric modeling of online social interactions using microblogs," *Comput. Environ. Urban Syst.*, vol. 70, pp. 53–58, 2018.

[44] A. Comber, C. Brunsdon, and M. Batty, "Geographic Analysis of Social Network Data," in *Proceedings of the 15th AGILE International Conference on Geographic Information Science*, Avignon, France.

[45] L. Ryan, *The Visual Imperative: Creating a Visual Culture of Data Discovery*. Morgan Kaufmann, 2016.

[46] M. H. W. Rosli and A. Cabrera, "Gestalt principles in multimodal data representation," *IEEE Comput. Graph. Appl.*, vol. 35, no. 2, pp. 80–87, Mar. 2015.

[47] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis: A Probabilistic Approach*. Springer Science & Business Media, 2007.

[48] S. Wexler, J. Shaffer, and A. Cotgreave, *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. John Wiley & Sons, 2017.

[49] J. Garae, R. K. L. Ko, and S. Chaisiri, "UVisP: User-centric Visualization of Data Provenance with Gestalt Principles," in *2016 IEEE Trustcom/BigDataSE/ISPA*, 2016.

[50] C. N. Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons, 2015.

[51] M. Krstajic and D. A. Keim, "Visualization of streaming data: Observing change and context in information visualization techniques," in *2013 IEEE International Conference on Big Data*, 2013.

[52] D. Clark, "Data Visualization Is The Future - Here's Why," *Forbes*, 10-Mar-2014. [Online]. Available: https://www.forbes.com/sites/dorieclark/2014/03/10/data-visualization-is-the-future-heres-why/#ed9490818840. [Accessed: 03-Feb-2018].

[53] Gartner, "Social Analytics," *Gartner IT Glossary*, 2013. [Online]. Available: https://www.gartner.com/it-glossary/social-analytics. [Accessed: 21-Jan-2018].

[54] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[55] S. B. Park, J. Jang, and C. M. Ok, "Analyzing Twitter to explore perceptions of Asian restaurants," *Journal of Hospitality and Tourism Technology*, vol. 7, no. 4, pp. 405–422, 2016.

[56] Y.-C. Chang, C.-H. Ku, and C.-H. Chen, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," *Int. J. Inf. Manage.*, 2017.

[57] R. Floris and M. Campagna, "Social Media Data in Tourism Planning: Analysing Tourists' Satisfaction in Space and Time," in *REAL CORP 2014 - PLAN IT SMART! Clever Solutions for Smart Cities.*, Vienna, Austria, 2014, pp. 997–1003.

[58] S. Shekhar and H. Xiong, *Encyclopedia of GIS*. Springer Science & Business Media, 2007.

[59] S. Milton, "Location Intelligence - The Future Looks Bright," *Forbes*, 11-Oct-2011. [Online]. Available: https://www.forbes.com/sites/stevemilton/2011/10/11/location-intelligence-the-future-looks-bright. [Accessed: 18-Jan-2018].

[60] K. Hahn, "3 Ways Location Intelligence Is Already Part of Your Life," *Ironsidegroup*, 17-Sep-2015. [Online]. Available: https://www.ironsidegroup.com/2015/09/17/3-ways-location-intelligence-is-already-part-of-your-life/. [Accessed: 04-Dec-2017].

[61] P. Franchet, "The 3 cores of Location Intelligence," *Galigeo*, 12-Mar-2017. [Online]. Available: https://saplumira.com/app/uploads/2015/11/White_Paper_Location_Intelligence_SAP.pdf. [Accessed: 02-May-2018].

[62] T. Bouadi, M.-O. Cordier, P. Moreau, R. Quiniou, J. Salmon-Monviola, and C. Gascuel-Odoux, "A data warehouse to explore multidimensional simulated data from a spatially distributed agro-hydrological model to improve catchment nitrogen management," *Environmental Modelling & Software*, vol. 97, pp. 229–242, 2017.

[63] M. Sethi, "Data Warehousing And OLAP Technology," *Int. J. Eng. Res. Appl.*, vol. 2, no. 2, pp. 955–960, Mar. 2012.

[64] Princeton University. Cognitive Science Laboratory, *WordNet: A Lexical Database for English*. .

[65] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, "Building a Sentiment Summarizer for Local Service Reviews," in *Proceedings of the WWW2008 Workshop: NLP in the Information Explosion Era (NLPIX 2008)*, Beijing, China.

[66] Google, “Analyzing Entities,” *Google Cloud Documentation*. [Online]. Available: https://cloud.google.com/natural-language/docs/analyzing-entities. [Accessed: 01-Jun-2018].

[67] Y. Jin, “Development of Word Cloud Generator Software Based on Python,” *Procedia Engineering*, vol. 174, pp. 788–792, 2017.

[68] B. (kevin) Chae, “Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research,” *Int. J. Prod. Econ.*, vol. 165, pp. 247–259, 2015.

[69] T. A. Small, “What the hashtag? A content analysis of Canadian politics on Twitter,” *Inf. Commun. Soc.*, vol. 14, no. 6, pp. 872–895, May 2011.

[70] E. Christou, *Social Media in Travel, Tourism and Hospitality: Theory, Practice and Cases*. Routledge, 2016.

[71] A. Huang, L. Gallegos, and K. Lerman, “Travel analytics: Understanding how destination choice and business clusters are connected based on social media data,” *Transp. Res. Part C: Emerg. Technol.*, vol. 77, pp. 245–256, 2017.

[72] R. Ferguson, “Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing,” *Journal of Consumer Marketing*, vol. 25, no. 3, pp. 179–182, 2008.