Chapter 1

Introduction

In this chapter we settle the main drivers, the state of the art and the general structure of this research. To begin with, we expose the context of the research in **?? Background** section. After that, in **?? Motivation** section, we explain the main reasons which move us to begin with this research line. The goals are described in **?? Objectives** section. We then present a literature review in **?? State of art** section. Finally, in **?? Project structure** section, we describes how the project is organised.

1.1 Background

Through last decades, tourism has become a popular global leisure activity. World Tourism Organization (UNWTO) has estimated that 1.2 billions of people travelled abroad in 2015. This growth is being influenced by several factors such as an increase of amount of free time, paid holidays, reduction of retirement age, increase of families income or rapid improvement of transportation systems, among others.

Due to these facts, tourism industry is the most important economic driver of many economies. Regarding to the World Travel & Tourism Council (WTTC) reports, [?], it is generating 9.8% of the wider GDP and supporting 248 millions of jobs. In particular, tourism in Spain is becoming the major contributor to its economy, representing a 16% of the total GDP. Last year 68.1 millions of tourists[1] visited Spain, marking the third consecutive year of record beating numbers. This upwards trend may be driven by the wave of terrorism that is hitting some competitor countries like Egypt, Tunisia or Turkey (which has been reflected in the world press as it is observed in Figure **??**).

Museums and architectural monuments are important touristic attractions. Spain has over 1,500 museums[2] and 13,000 protected monuments[3]. These cultural instituitions have endured economic crisis effects. Spanish government has cut

[1] Source: http://www.ine.es/inebmenu/mnu_hosteleria.htm

[2] Source:http://directoriomuseos.mcu.es/

[3] Source:http://www.mecd.gob.es/portada-mecd/

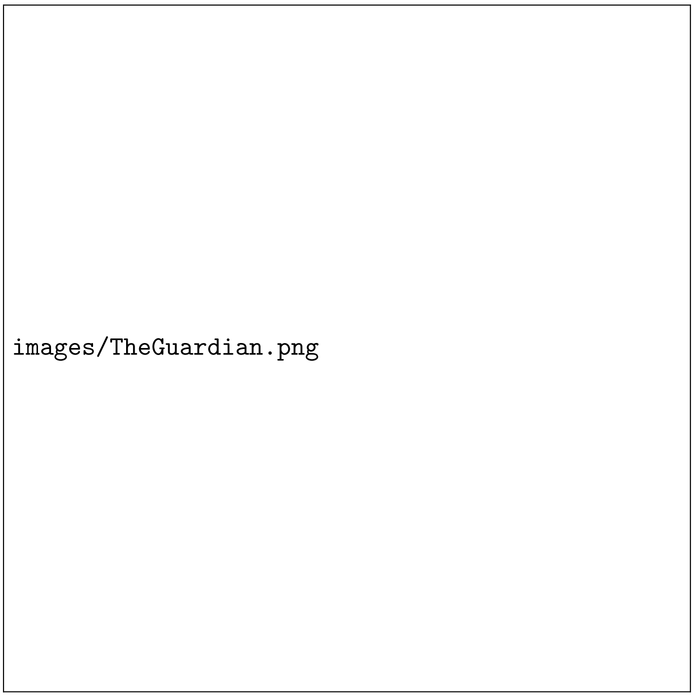

Figure 1.1: The Guardian, 3rd January 2016. Source: https://www.theguardian.com/travel/2016/jan/03/tourists-spain-avoid-terror-threat-egypt-tunisia

back a huge amount of culture budget which has affected museums and monuments operations[4]. Therefore, these instituitions have to develop new strategies so as to attract more visitors, even more with the rising tide of tourism in Spain.

1.2 Motivation

The rapid development of Information and Communication Technology (ICT) has led to the Digital Age where about 40%[5] of world population has an internet connection. This revolution has come up with the Web 2.0 concept. According to Wikipedia, Web 2.0 is a web application where the user is able to interactively share information. It is an evolution of old-fashioned Web 1.0 which was essentially static screenfuls. In this new web design, the user is invited to share its information and contribute to the web content.

Moreover, million of data is being generated daily. Since human existence beginning to 2003, it is estimated that 5 millions of terabytes of data were generated. Incredibly, over 8,000 millions of terabytes were produced just in 2015[6].

[4] Source:http://www.lavanguardia.com/local/madrid/20120716/54325908919/madrid-cierra-museo-ciudad-poder-pagar-deudas-gallardon.html

[5] Source:http://www.internetlivestats.com/internet-users/

[6] Source: https://www.youtube.com/watch?v=wWcgYZWCAXg

This fact has led into the big data concept. The big data has encompassed this fact. Traditionally, big data was defined as the three data growth challenges: volume (huge amount of data), velocity (speed of processing this data) and variety (different sources and types of data). Actually, big data means all related to data science: it is all related about extracting insights and knowleadge from data applying statistics, data mining and descriptive and predictive induction analysis tools.

Owing to this, more and more companies are aware of the essentiality of these methods. They help with making better decisions or understanding customers behaviour which has a direct effect on revenues. In this way, sentiment analysis has experienced an important growth as a research area. As it is explained in this project, sentiment analysis basically develops techniques to detect automatically positive and negative opinions which contributes to know users or customers thoughts.

Despite of the straggling situation that museums and cultural institution are going through, the application of analytics tools into this organisations is very tiny.

1.3 Objectives

This project is aimed at an alternative to surveys which present known inconvenients. It has been demonstrated that surveys information can be biased and not neutral. Moreover, some surveys ask to many questions which involves an investment of time respondent. Due to this fact, sentiment analysis has spring up as an alternative. The main idea is to apply text mining in users opinions, either in webpages or social media, so as to extract valuable information.

This master thesis is developed as a first approach to sentiment analysis into touristic attractions domain, scrapping reviews from Web 2.0. Therefore, we develop a methodology to analyse web opinions from the most visited monument in Spain, the Alhambra. Doing so, we will be able to know facts about the visit that people like and dislike.

More precisely, the first objective is to analyse reviews and build an exploratory and statistical report of this data. The second goal is to study the correlation between the user and the machine sentiment. After that, we propose to develop a predictive induction study, building classification models in order to predict automatically sentiments depending on the target variable. Finally, we carry out a descriptive induction study into negative reviews so as to discover interesting patterns.

1.4 State of the Art

Since 2000, sentiment analysis has experienced an important growth in research. The first time that *sentiment analysis* and *opinion mining* concept appeared was in [?] and [?], respectively. However, we consider [?] and [?] as the Bibles of this branch. This two references describe different machine learning and data mining algorithms applied to opinions. In [?], we find a detailed sentiment analysis survey up to 2014. In this paper, the authors present a summary table of fifty-four articles with all relative information (sentiment analysis task, domain, algorithm used, polarity, data scope, data set and language).

Over this project, we have applied a core natural language toolkit (`CoreNLP`) developed by the Natural Language Processing Group, in Stanford University. The researchers described in [?] the overall system. Concretaly, the development of the sentiment analysis algorithm is explained in [?].

Sentiment analysis has a large ream of applications. This fact has been demonstrated in the literature over the years. In [?] paper, the authors use movie reviews in order to apply machine learning methods for sentiment classification. In [?], the authors analysed reviews from banks, automobiles, travel destinations and movies. As another example, authors in [?] apply other techniques on electronic devices reviews and restaurant reviews, scrapping Amazon and Yelp webs. In this research, the authors propose two different methodologies in order to discover which aspects are evaluated in sentences: *Sentence-LDA* (SLDA) and *Aspect and Sentiment Unification Model* (ASUM). Awaring of the growth of tourism as an e-commerce bussines, [?] study machine learning techniques on hotel review from TripAdvisor. Even more, authors in [?] develop BESAHOT system for hotel managers which analyses customer opinions from several sources. In [?], the authors carry out a text mining study for extracting business values on pizza industry social media. Finally, in [?], authors develop a method to extract aspects from hotels and restaurants reviews as well as classify its sentiment using TripAdvisor as a source.

1.5 Project Structure

This project is structured as follows. The first part is an introduction chapter to sentiment analysis. This chapter tries to introduce the lecturer the concept of opinion mining and the process of sentiment analysis problem. After that, in the following chapter, we explain where we get data. Additionally, we develop a description of our data set. Then, in next chapter, we evaluate the correlation between expert and machine sentiment target. The experiments description and results discussion about classification models are presented in this same part. In the following chapter, we obtain an overview of the relation between negative opinions and a set of features through subgroup discovery. Finally, we present our conclusions and suggest future research.

The following framework, Figure **??**, gives a visual representation of our project structure:

images/esquema.png

Figure 1.2: Project structure.