MASTER'S DEGREE THESIS

Master in Data Science and Computer Engineering

Title: *Sentiment Analysis For Touristic Attractions: A Case Study On The Alhambra*.

Author: Ana Valdivia García

Advisor: Salvador García López

Department: Computer Science and Artificial Intelligence

University: University of Granada

Delivery date: 12/09/2016

Academic year: 2015/2016

ugr | Universidad de Granada

University of Granada

Mater's degree thesis

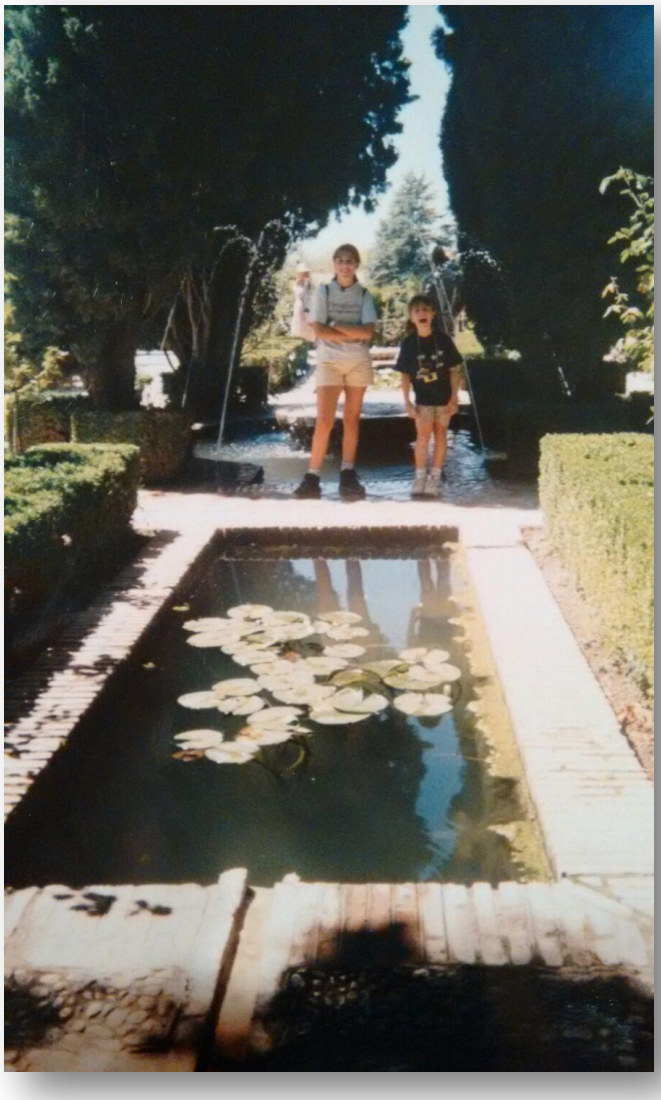
Sentiment Analysis For Touristic Attractions: A Case Study On The Alhambra

Ana Valdivia García

Advisor: Salvador García López

Computer Science and Artificial Intelligence

To my sister and my to be born nephew, Elena and Lluc.



My first, but not last, time in the Alhambra.

The student, **Ana Valdivia García, with ID 43564839-X**, guarantees, by signing this Master's Thesis, that this research work has been done respecting the rights of other authors to be cited. In this way, she declares her authorship and assumes the originality of this work.

In recognition whereof, she signs this document in **Granada on 5th of September 2016**:

Abstract

The development of Web 2.0 has led to an important amount of content in webpage. Users are free to express their opinions about products, places and events. This project research is aimed at introducing sentiment analysis into touristic attractions. To begin with, we scrap TripAdvisor reviews from the most touristic attraction in Spain, the Alhambra. We then create two sentiment labels: the expert sentiment which is the rate of the reviewer; and the machine sentiment which is extracted from a Natural Language Processing toolkit developed in Stanford University. After that, we build classification models so as to predict polarity sentiments. Finally, we develop a subgroup discovery method so as to extract valuable information about negative reviews.

Key words: opinion mining, sentiment analysis, tourism, natural language processing, subgroup discovery

Contents

Chapter 1

Introduction

In this chapter we settle the main drivers of this research. We introduce the context of the project as well as the state of the art and the general structure.

To begin with, we describe the framework in 1.1 **Background** section. After that, in 1.2 **Motivation** section, we explain the main reasons which move us to begin with this work. The goals are described in 1.3 **Objectives** section. We then present a literature review in 1.4 **State of the Art** section. Finally, in 1.5 **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, [WTTC, 2016], 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 1.1).

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

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

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 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, millions 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].

This fact has led into the big data concept. 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).

[2]Source:`http://directoriomuseos.mcu.es/`
[3]Source:`http://www.mecd.gob.es/portada-mecd/`
[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`

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 should be a must.

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 [Yi et al., 2003] and [Dave et al., 2003], respectively. However, we consider [Liu, 2012] and [Pang and Lee, 2008] as the Bibles of this branch. These two references describe different machine learning and data mining algorithms applied to opinions. In [Medhat et al., 2014], 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 of Stanford University. The researchers described in [Manning et al., 2014] the overall system. The development of the sentiment analysis algorithm is concretely explained in [Socher et al., 2013].

Sentiment analysis has a large ream of applications. This fact has been demonstrated in the literature over the years. In [Pang et al., 2002] paper, the authors use movie reviews in order to apply machine learning methods for sentiment classification. In [Turney, 2002], the authors analysed reviews from banks, automobiles, travel destinations and movies. As another example, authors in [Jo and Oh, 2011] apply other techniques on electronic devices and restaurant reviews, scrapping Amazon and Yelp webs respectively. 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, the authors in paper [Elango and Narayanan,] study machine learning techniques on hotel review from TripAdvisor. Even more, authors in [Kasper and Vela, 2011] develop BESAHOT system for hotel managers which analyses customer opinions from several sources. In [He et al., 2013], the authors carry out a text mining study for extracting business values on pizza industry social media. Finally, in [Marrese-Taylor et al., 2013], authors develop a method to extract aspects from hotels and restaurants reviews as well as classify its sentiment using TripAdvisor as a source.

Related to subgroup discovery techniques, the authors in [Herrera et al., 2011] write a survey of this branch. In this paper, they described the main concepts as well as subgroup algorithms and measures developed so far. In the article [Atzmueller and Puppe, 2006], the authors propose an efficient algorithm, SD-Map, to discover interesting patterns in data. Martin Atzmueller developed an R package implementing this method.

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 to the concept of opinion mining and the process of sentiment analysis problem. After

that, in the following chapter, we describe TripAdvisor as the source and the Alhambra as the touristic attraction to base our study. Moreover, we explain where and how 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 there too. In the following chapter, we develop a first approach between sentiment analysis and subgroup discovery research branches. We develop a methodology to obtain interesting patterns in negative opinions. Lastly, we present our conclusions and suggest future research in the final chapter.

The following framework, Figure 1.2, gives a visual representation of our project structure:

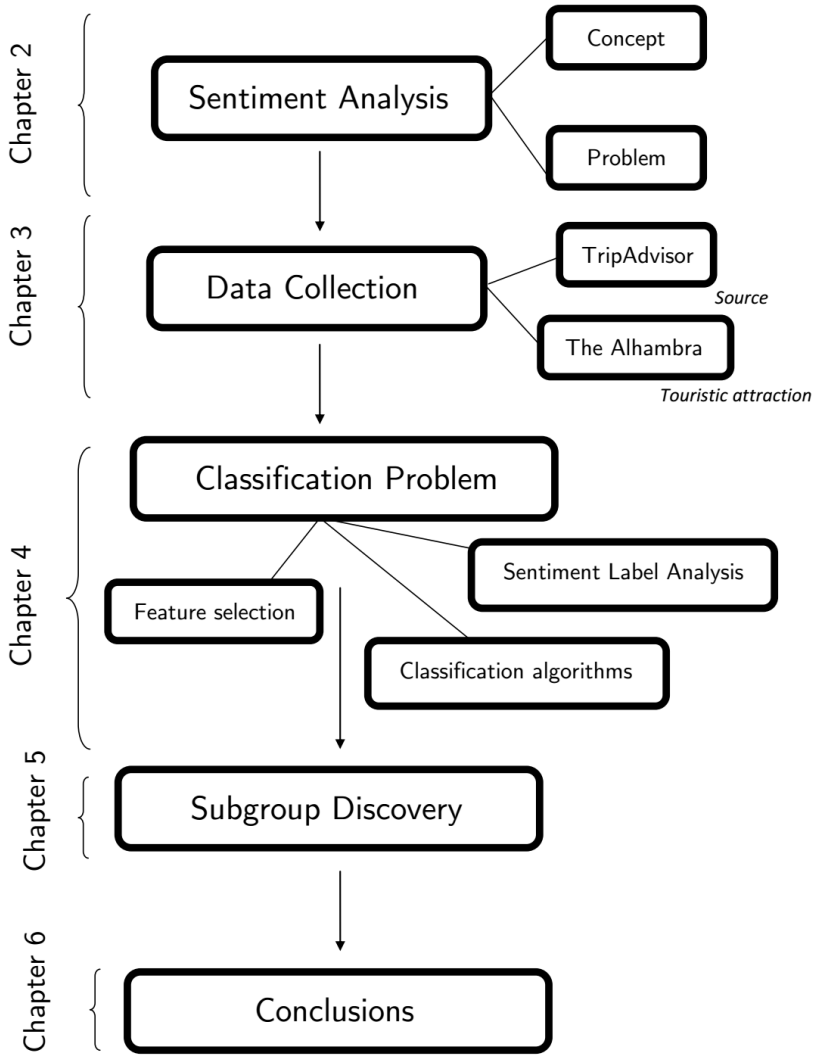


Figure 1.2: Project structure.

Chapter 2

Sentiment Analysis

In this chapter, we will introduce the concepts related to sentiment analysis. To begin with, in 2.1 **Concept of Sentiment Analysis** section, we describe the main idea of text mining. After that, we give a precise definition of opinion. Secondly, in 2.2 **The Sentiment Analysis Problem**, we explain what does sentiment analysis problem consist of. We summarise the structure of the problem and explain some of its concerns. We want to lay down this chapter as the theoretical part of our project in order to develop our experimental analysis.

2.1 Concept of Sentiment Analysis

This section gives to the lecturer a main idea of what sentiment analysis is. To begin with, we introduce text mining concept and how is related with sentiment analysis. After that, we define opinion describing its main features.

2.1.1 An introduction to text mining

As it is defined in [Tan et al., 1999], *text mining* refers to the process of extracting useful information from text. It is also known as text data mining or text analytics.

The extraction of interesting patterns in text mining has been considered a arduous task, more than data mining. The fact is that in text mining, the researcher has to deal with unstructured and fuzzy information. Therefore, the challenge resides in transforming unstructured or weakly structured to structured data.

Several areas are related to text mining. The first one is Natural Language Processing, NLP, which refers to the ability of systems to process human language. As a second example, Information Retrieval, IR, is another research branch

which objective is to find documents that may contain answers to questions. Finally, Information Extraction, IE, which goal is to extract specific information from documents.

The main disadvantage of text mining is that text is unstructured. Because of that, it is required to transform this data into structured data. The main technique is to extract features adequately which represent the text in order to apply machine learning techniques. Because of that, text preprocessing is the key step (see Figure 2.1) in this field.

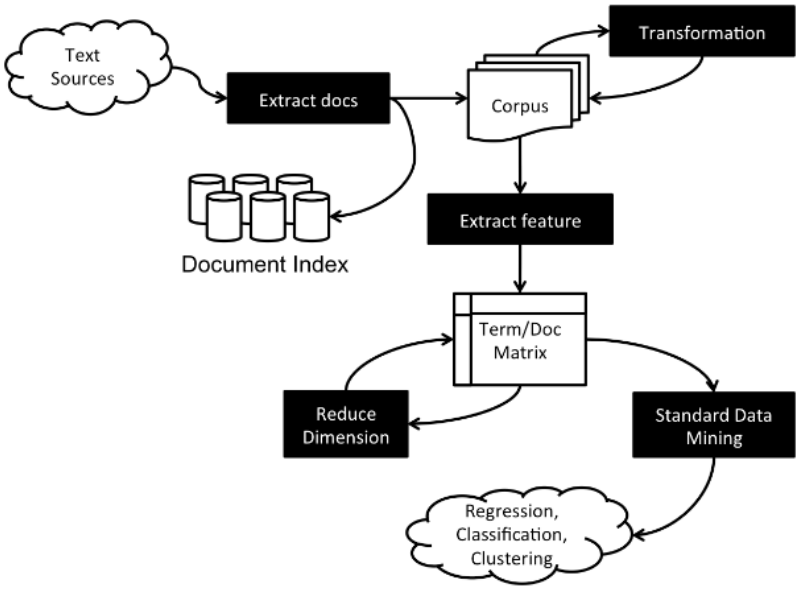


Figure 2.1: The standard text mining process.

There exist a large variety of preprocessing techniques: *Tokenization* is defined as the technique of breaking the text into sentences and words, *Part of Speech* (POS) *Tagging*, classifies words into grammatical categories based on their function in a given sentence (the most common tags are: adjectives, adverbs, nouns, verbs, prepositions, etc.), *Parsing* analyses syntactically all elements in a sentence, *Stemming* is the process of reducing words to their root, for example, the set of words {*nightmare, nighttime, nocturnal, nightlife...*} can be reduce to *night*, *Lemmatization* is a technique analogous to stemming, but in that case the word is brought into its non-inflectes dictionary form. Remove *stop words* like articles, conjunctions or prepositions.

One of the most popular visualization technique in this field is word cloud. Word cloud is a visual design that represents text by its single words which importance is displayed with font size and color. This system is very useful because in a simply glance one analyses what text is about as well as has an idea of most

relevant words that can be set as features. For instance, Figure 2.2 illustrates the world cloud of *sentiment analysis*.

Figure 2.2: Word cloud of sentiment analysis found on the web.

Because of the recent development of the ICT, the number of applications in this field has been growing exponentially. As it is shown in [Hotho et al., 2005] or in [Mostafa, 2013], there exist applications in bioinformatics, medicine, psychology, security (like anti-spam filtering), patent analysis, sociology, politics, marketing, business intelligence and so on.

2.1.2 Opinion mining and sentiment analysis

The proliferation and continuos growth of the Web 2.0 and social networks, like Facebook or Twitter, has led to a huge amount of online recorded opinion text. Nowadays, anybody with access to Internet can put into words his/her review towards restaurants, hotels, airlines, monuments, books or even rice cookers.

Because of that, the challenge resides now in organising this amount of opinions and analyse them in order to extract valuable information, like customer satisfaction, to make better decisions. Consequently, the concepts of *opinion mining* and *sentiment analysis* have come up as a new avenues of research in computer and management science.

Opinion mining is the field of knowledge that analyses people's opinions, reviews or thoughts about products, companies or experiences identifying its sentiment. Although an *opinion* refers to a person's point of view and *sentiment* refers to the feeling that somebody shows on that opinion, Bing Liu argued in [Liu, 2012] that *opinion mining* and *sentimental analysis* are merely the same concept. Therefore, throughout this project we will refer to opinion mining and sentiment analysis indistinctly.

Overall, the main goal of opinion mining is to classify the expressed sentiment. The classification problem is raised as a binary classification; and the labels of the class variable are commonly *positive* or *negative* which is referred as *sentiment polarity*. However, the class target may be multi-label: increasing the degrees of negativity and positivity or detecting the six universal emotions (anger, disgust, fear, happiness, sadness and surprise)[Ekman et al., 2013].

2.1.3 Opinion concept

As it is found in Cambridge Dictionary, an *opinion* is a thought or belief about something or someone. Thus, there exist two main attibutes in an opinion: the *entity* and the *sentiment*. The entity is the product, service, place, person, company or event which the opinion is addressed to. The sentiment is the feeling that underlies the opinion. Additionally, the *aspect* of the entity is a property related to the entity; the *opinion holder* is the person who puts into words the opinion; the *time* of the opinion is the date which the opinion was expressed.

As an example, the opinion in Figure 2.3 has some thoughts about Abaco Tea, a tea shop in the Albayzin, the famous Granada's neighbourhood.

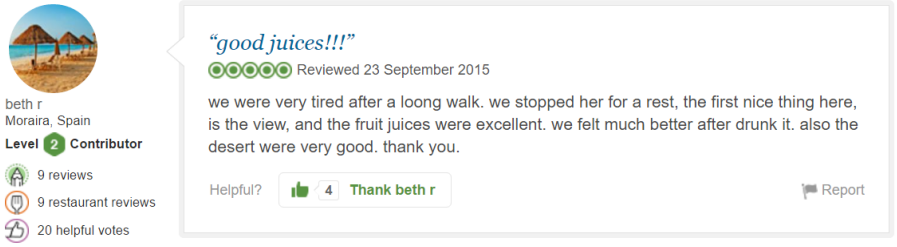


Figure 2.3: A user's opinion about a tea shop in Granada.

As we can read, the opinion holder reports feelings about different aspects of the tea shop. She talks about the views, drinks and deserts. Thus, the following Table 2.1 shows a summary of the different properties of the text.

Concept	**Value**
entity	abaco te
opinion holder	beth r
date	2015-09-23
(aspect, sentiment)	(view, positive)
(aspect, sentiment)	(fruit juice, positive)
(aspect, sentiment)	(fruit juice, positive)
(aspect, sentiment)	(desert, positive)

Table 2.1: Summary table of the opinion in Figure 2.3.

Liu proposes in [Liu, 2012] a more precise definition that has been extended to all sentiment analysis literature:

Definition 2.1 OPINION: *An opinion is a 5-tuple containing the target of the opinion (entity), the attribute of the target at which the opinion is directed, the sentiment or polarity, the opinion holder and the date when the opinion was emitted:*

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l),$$

where:

- e_i *is the i-th opinion target,*
- a_{ij} *is the j-th attribute of* e_i*,*
- s_{ijkl} *is the polarity of the opinion towards an attribute* a_{ij} *of entity* e_i *by the opinion holder* h_k *at time* t_l*,*
- h_k *is the k-th opinion holder and*
- t_l *is l-th time when the opinion was emitted.*

In that way, last opinion produces four 5-tuples:

- $(e_1, a_{11}, s_{1111}, h_1, t_1) =$ (*abacto te, view, positive, beth r, 2015-09-23*)
- $(e_1, a_{12}, s_{1211}, h_1, t_1) =$ (*abacto te, fruit juice, positive, beth r, 2015-09-23*)
- $(e_1, a_{12}, s_{1211}, h_1, t_1) =$ (*abacto te, fruit juice, positive, beth r, 2015-09-23*)
- $(e_1, a_{13}, s_{1311}, h_1, t_1) =$ (*abacto te, desert, positive, beth r, 2015-09-23*)

As one may realise, there exist different types of opinions. Generally, there are two main types of opinions:

- **Regular opinion**: Type of opinion that expresses a sentiment about an aspect of an entity.
  - **Direct opinion**: It is a direct opinion when the opinion holder makes reference to the entity directly.
    The TV serie Game of Thrones is awesome!
  - **Indirect opinion**: The reference is made undirectly.
    After having dinner at this restaurant, we got a stomachache.
- **Comparative opinion**: This type of opinion compares two entities or aspects of an entity.
  I prefer PULEVA's milk than a store brand.

On the other hand, if we take into account subjectivity:

- **Explicit opinion**: It is a regular or comparative opinion which is expressed with subjectivity.
  I don't like Cruzcampo's beer.
- **Implicit or fact-implied opinion**: It is a regular or comparative opinion expressed objectively, i.e., the opinion is given as a fact. There exist two subtypes:
  - **Personal opinion:** The opinion holder is giving his/her or someone's personal experience.
    I bought a chair in IKEA and I got backache from sitting on it.
  - **Non personal opinion:** The opinion is not implying a personal experience.
    Nintendo has doubled in value since Pokémon Go's release.

Generally, opinions which are implied with factual information are harder to analyse than explicit statements. This fact is due to the non-limited domain of implicit opinions. In addition, the study of implicit opinions entail a lack of information like who is the opinion holder.

2.2 The Sentiment Analysis Problem

The problem of sentiment analysis is very complex. Because of that fact, it is a must to understand it and organise the main ideas.

The first concept related to this problem is the *subjectivity*. In contrast with factual information, somebody express his/her opinion because is being moved by his/her beliefs, experiences and feelings. This implies that people may express different opinions about the same product, service or place. Thus, it is very important in sentiment analysis to collect a huge amount of opinions, which is called *corpus*, in order to have a global overview.

Once a large number of opinions from different people are collected, we have to assess the level of the analysis. There exist three main classification levels realted to the sentiment analysis problem:

- **Document level**: The written text to mine is a whole document. It is analysed whether the sentiment is positive or negative. This is the simplest task.
- **Sentence level**: In this case, the aim is to detect the sentiment in each sentence: positive, negative or neutral. This level of classification is related to *subjectivity classification* which discerns between sentences base on facts (*objective sentences*) or on opinions and sentiments (*subjective sentences*).
- **Aspect or entity level**: This is the most in-depth level. At this point, it is studied the sentiment and the target of the opinion. For example, if

> we analyse the following statement to the sentence-sentiment level *"The Alhambra itself was fabulous just such a shame about some of the ticket staff"* is difficult to classify due to its polarity. Nonetheless, if we analyse it to the aspect-sentiment level the opinion holder feels a positive sentiment to Alhambra monument but a negative one caused by Alhambra's staff. This level is useful for discovering aspects that cause the sentiment and thus understand the underlying problem.

Another concept related to sentiment analysis problem is the *sentiment lexicon*. The sentiment lexicon or opinion lexicon is a set of words that express a positive or negative sentiment. For example, the set {*awesome, happy, riveting*} belongs to the positive sentiment lexicon and {*deplorable, owful, sad*} belongs to the negative.

However, the sentiment lexicon is not a sufficient tool because it may involve some problems in the sentiment analysis. For example, the same word can express a positive or negative sentiment. There exist sentences without sentiment words that express an opinion or sentences with sentiment words that do not show any feeling.

The last concept is the *Natural Language Process*, NLP. As we have explained before, NLP is a branch of knowleadge that process human language to computers. The history of NLP starts in 1950 with Alan Turing and his Turing test, evaluating natural conversation between a human and a computer. This branch of research has experienced a rapid growth due to the development of machine learning algorithms. These techniques have been studied to automate language processing.

There is a large amount of NLP task: automatic summarization, discourse analysis, machine translation, named entity recognition, parsing, etc. Sentiment analysis is considered one of its major task.

Lastly, sentiment analysis has different concerns: *irony* is defined as *the use of words to convey a meaning that is the opposite*. Humans detect irony because of different facts like intonation or face expression. Because of that, it is a more difficult task for computers. However, there are different papers like [Reyes et al., 2013] that face this challenge. *Spam* is another challenge. In order to develop a robust analysis it is required to have realistic data identifying fake reviews.

2.2.1 The sentiment analysis process

As it is shown in Figure 2.4, sentiment analysis is based on different processes. In this part of the project, we insightfully explain each step.

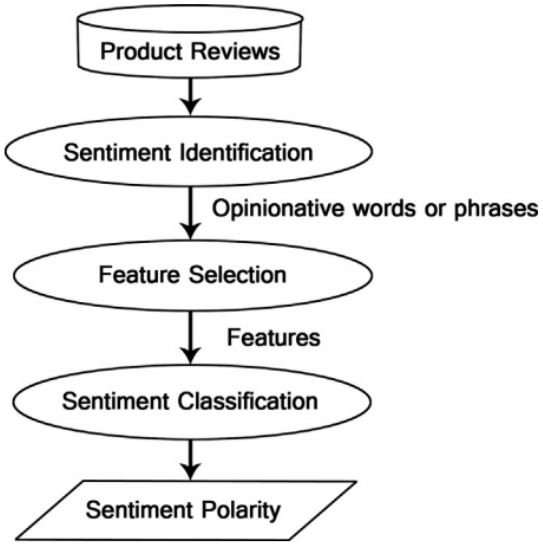


Figure 2.4: Scheme of sentiment classification process shown in [Medhat et al., 2014] for product reviews.

Data collection

Data sets used in this classification problem are very relevant. It is important to have a lot of opinions, a large corpus, from different holders in order to avoid subjectivity. Due to the recent growth of Web 2.0, user content generated and social media, Internet has become an important source of opinion and reviews.

This fact has been reflected in the research, sentiment analysis has been applied to a large variety of topics such as: movie reviews [Pang et al., 2002] or [MartíN-Valdivia et al., 2013], book reviews [Hu et al., 2012], hotels reviews [Kasper and Vela, 2011], social media of pizza companies [He et al., 2013], politics tweets [O'Connor et al., 2010] and [DiGrazia et al., 2013], etc. According to this information, this project will be the first one which analyses tourism attraction reviews.

Sentiment identification

As a second step, it is needed to identificate the sentiment of the opinion. On the one hand, the opinion may be evaluated, e.g., a 1-5 stars ranking (TripAdvisor), a thumbs up or down system (Youtube) or an emotional reaction option (Facebook). Hence, the problem may be binary-label classification task, i.e.. {positive, negative} sentiment class. Or, it may be multi-class, i.e., {positive,

negative, neutral}, {anger, disgust, fear, happiness, sadness, surprise} or {1, 2, 3, 4, 5} scale.

On the other hand, if the opinion is not ranked it may be labeled using opinion words or other techniques explained in **Classification problem**.

Feature selection

This is the main step in order to structure text. The idea is to mine opinions in order to select text features as variables. This methods can be divided into lexicon-based methods and statistical methods. On the one hand, lexicon-based methods are based on develop a large lexicon of words starting with a small sample using synonims.

On the other hand, statistical methods are automatic algorithms. The first step is to clean reviews using some preprocessing techniques (tokenization, POS tagging, lemmatization, ...) which were explained before. After that, it is selected a certain number of words, Bag of Words (BOW), for each class. This selection is made by terms presence and frequency such as term frequency-inverse document frequency ($tfidf$) value, which represents how relevant is a word in a collection of documents or corpus. Term frequency is the total number of times that a word occurs in the corpus. However, taking into account only this measure can led us to incorrect importance value because of the fact that articles, conjunctions or prepositions appears many times in text. Due to this, inverse document frequency decreases the weight given to the very recurring words and increases the weight of terms that occur rarely. Mathematically, it may be expressed as:

$$tfidf = tf_{i,j} \cdot idf_i = n_{i,j} \cdot log_2 \frac{|D|}{|\{d \mid t_i \in d\}|}$$

where:

- t_i is the term,
- in a document d_j,
- $n_{i,j}$ is the total number of times that term t_i occurs in a document d_j,
- $|D|$ denotes the total number of documents.

This measure is often used because of its simplicity. There exist other methods for feature selection like: point-wise mutual information, chi-square, latent semantic indexing and latent dirichlet allocation that are described in [Medhat et al., 2014] article.

However, there exist more sophisticated processes that discover topics or aspects which can be set as features. Latent Dirichlet Allocation (LDA), [Blei et al., 2003] is a hierarchical Bayesian model in which each item of a collection is modeled

as a finite mixture over a set of topics. Then, each topic is modeled as a infinite mixture over a set of topic probabilities. These topics try to build a representation of the documents. In this way, authors in [Jo and Oh, 2011] propose two new algorithms which discover aspects in sentences and sentiments: Sentence-LDA (SLDA) and Aspect and Sentiment Unification Model (ASUM).

In this way, text (unstructured data) is turned into document term matrix (structured data).

Classification problem

In one hand, sentiment analysis can be studied in a machine learning approach (see Figure 2.5). It is considered a supervised problem when the opinion is ranked. The aim of this problem is to build a model with the selected features so as to classify new unlabeled opinions. The main techniques that are used to classify are: decision trees, linear classifiers (SVM and neural networks), rules and probabilistic classifiers (naive bayes, bayesian networks and maximum entropy). By contrast, when there is no feedback from the opinion holder or another user the problem is unsupervised. In this case, the aim is to indentify the polarity of the text using rules or heuristics obtained from language knowledge.

On the other hand, the sentiment analysis problem can be studied as a lexicon approach which relies on a collection of known sentiment terms. There exist two techniques: dictionary-based approach and corpus-based approach. The first one refers to build a sentiment lexicon manually. This approach has the inconvenience that may not work to specific opinions. However, corpus-based approach solves this problem growing the dictionary with corpus terms.

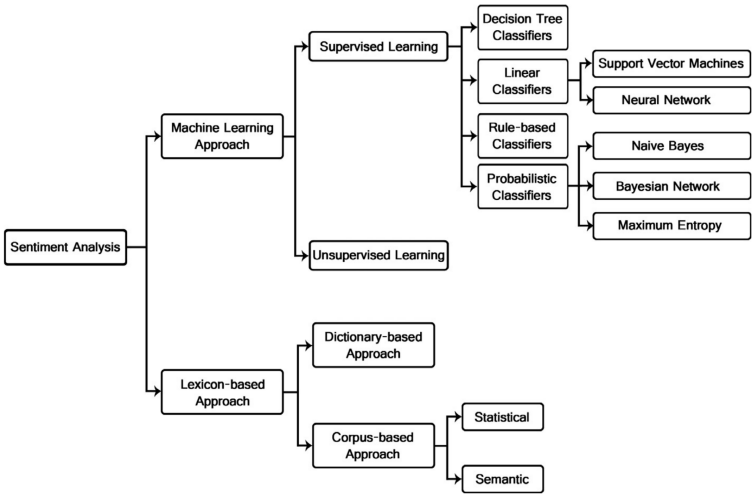


Figure 2.5: Scheme of sentiment techniques according to [Medhat et al., 2014].

Chapter 3

Data Collection

In this chapter we present how opinions are collected in order to build our data set. In section 3.1 **The real scenario** is presented the context of our text data, describing TripAdvisor and the Alhambra as our review source and entity, respectively. In 3.2 **The data** section we explain how is data structured. Firstly, we explain how we obtain Alhambra's reviews from our source: TripAdvisor. After that, we report some statistics in order to make an idea of data behaviour. This data description is necessary so as to develop an efficient analysis.

3.1 The Real Scenario

Over recent years, the rapid development of Web 2.0 and the explosive growth of tourism around the world has led to an important amount of information about travel planning on the Internet. Thus, the number of tourist that plan themself their own holidays has grown up in parallel.

Consequently, tourism has become one of the most important e-commerce services. Websites such as Booking, Kayak, Minube, TripAdvisor or Trivago have experienced an important increase of users.

3.1.1 TripAdvisor

According to Wikipedia, TripAdvisor is an American travel website company providing reviews from travellers experiences about hotels, restaurants and monuments.

TripAdvisor was founded in February of 2000 by Stephen Kaufer and Langley Steinert, among others. It started as a site with information from guidebooks, newspapers and magazines. In 2004 it was purchased by IAC and one year later,

it spun off its travel group of businesses: Expedia. After that, the website turnt to a user generated content.

The site describes itself as follows:

"TripAdvisor is the world's largest travel site[1] *enabling travelers to unleash the full potential of every trip. TripAdvisor offers advice from millions of travelers and a wide variety of travel choices and planning features with seamless links to booking tools that check hundreds of websites to find the best hotel prices."*

Figure 3.1: TripAdvisor Logo.

In general terms, this website has made up the largest travel community, reaching 340 million unique monthly visitors [2], and 350 million reviews and opinions covering more than 6.5 million accommodations, restaurants and attractions over 48 makerts worldwide.

The most interesting feature of this webpage is that its information and advice index is constructed from the accumulated opinions of million of everyday tourists. In addition, the site also publishes a ranking called *Popularity Index*. This ranking is computed using an algorithm which takes into account users information and other published sources such as guidebooks or newspaper articles. The index runs from number 1 to the overall of restaurants, hotels or other attractions within the city. In this way, a traveller can find the most interesting visitor attraction or the best appreciated restaurant.

Lastly, one of the major concerns of user-generated content is the credibility of the opinions. Awaring of it, TripAdvisor has thought up several measures in order to avoid spam and fictitious reviews such as: not allowing the use of commercial email addresses, posting warnings about the "zero tolerance for fake opinions". Regarding to this, some studies like [Jeacle and Carter, 2011] or [Ayeh et al., 2013] have carried out so as to analyse credibility and truthfulness of TripAdvisor. These research projects have concluded that TripAdvisor is an influential and trusted website.

[1]Source: comScore Media Metrix for TripAdvisor Sites, worldwide, February 2016.
[2]Source: TripAdvisor log files, Q1 2016.

3.1.2 The Alhambra

Granada is an Andalusian capital city located in the south-east of Spain, at the foot of Sierra Nevada mountains. It is ranked as the 13th largest urban area of the country, covering an area of 88.02 km^2.

The oficial population in 2015 was estimated to be 235,800 citizens[3]. However, this city holds more inhabitans due to its university. The University of Granada (UGR) is ranked as the 4th by students number: it has 60,000 enrolled students and 3,500 professors[4]. In addition, every year over 2,000 ERASMUS students study in UGR. In 2014, UGR was voted as the best Spanish university by international students.

Nowadays, the most important economic driver in the city is tourism. Last year, Granada broke a touristic record: it received more than 2.5 millions of visitors, a growth of 5.06 % with respect to 2014. It is estimated that this sector represents the 14% of the total GDP and it generates 15% of employment[5].

The most important touristic attraction in Granada is the Alhambra (illustred in Figure 3.2). This Nasrid palace was constructed in AD 889 as a small fortress. The Moorish emir Mohammed ben Al-Ahmar rebuilt it in the middle of the 13th century. After that, in 1333, the Sultan of Granada Yusuf I converted the fortress into a royal palace. Later on, after the Reconquista in 1492, the Catholic Monarchs, Isabel and Fernando, settled the Royal Court. Despite of suffering from decay, pillage and wars, Alhambra is a UNESCO World Heritage Site since 1984.

This monument covers an area of about 142,000 m^2. The palace is composed of:

- **Alcazaba**: This is a militar neighbourhood and the oldest part. The main function of this building was to protect some districts. There, we can find *La Torre de la Vela* where Isabel La Católica rised her flag as a symbol of the conquest.
- **Nasrid Palaces**: This palace was the residence of the court. It is composed of different rooms and courtyards:
  - *Hall of Ambassadors*: It is the largest room of the whole monument. It used to be the reception room. The ceiling of this throne room represents the seven heavens of Muslim cosmos.
  - *Hall of the Abencerrajes*: The name of this room derives from a noble, who were rivals of Boabdil. According to the legend, the family was

[3]Source: `http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/htm/sm18087.htm`

[4]Source: `http://www.ugr.es/pages/universidad/granada`

[5]Source: `http://www.lavanguardia.com/local/sevilla/20160122/301594948076/granada-recibe-2-6-millones-de-turistas-y-5-6-de-pernoctaciones-en-ano-record.html`

Figure 3.2: Alhambra's panoramic view.

massacred while they attended a banquet here.

  - *Court of the Lions*: The Court had a cross ground floor design with the well-known 12 Marble Lions fountain in the center. This fountain reflects the complexity of the hydraulic system operating on the site.

- **Generalife**: This villa was the summer residence of the Nasrid kings. It is plenty of fountains and gardens.
- **Palacio de Carlos V**: This palace was added by the grandson of the Catholic Monarchs, Carlos V, in 1526. It is free-entrance.

Owing to all these facts, this monument provides to the city a huge amount of tourists, in fact, it is the main reason for visiting Granada. In 2015, it counted a total of 2,474,231 visitors, the highest mark since records began. Because of this, the Alhambra was the last year most visited Spanish monument[6], followed by Sagrada Familia (Barcelona) and Museo del Prado (Madrid).

According to [de Andalucía, 2015], the huge amount of visitors are concentrated in April-May and August-September. Foreigners represented a 70% of the total of visitors in 2015: French, German, American and English are the most regular visitors.

There are 6 different channels in order to get a ticket: authorized agents, individual presale (Internet and phone), individual direct sale, entities with agreement,

[6]Source: `http://www.elmundo.es/andalucia/2016/01/28/56aa449522601d74758b457e.html`

Figure 3.3: Map of the monument.

open door days and city pass.

Specifically, ticketmaster is the official website to buy an individual presale ticket. The main types of entrance are:

- **Alhambra General**: Access to Alcazaba, Nasrid Palaces, Generalife, Carlos V Palace, Public baths and the Mosque.

- **Alhambra Gardens**: This ticket allows the entrance to Generalife, Partal and Alcazaba. In addition, it is allowed the entrance to some other city monuments such as: el Bañuelo, Chapiz House, Dar al-Horra palace, etc.

- **Alhambra at Night**: Access to some parts (Nasrid Palace or Generalife) at night.

- **Dobla de Oro**: Tour through Alhambra and Albaycin's neighbourhood. There exist three categories of Dobla de Oro entrance: General, Gardens and Night.

- **Alhambra Experiences**: This type of visit gives a chance to combine the night visit to the illuminated Nasrid Palaces and the day visit to Alcazaba, Jardines and Generalife on two consecutive days.

The average price of the Alhambra ticket is considered cheap. As summary, we can check it in Table 3.1 :

Ticket	Price
Alhambra General	14 €
Alhambra Gardens	7 €
Alhambra Night	8 €
Dobla de Oro General	19.65 €
Dobla de Oro Gardens	11.65 €
Dobla de Oro Night	14.65 €
Alhambra Experiences	14 €

Table 3.1: General prices of the Alhambra on the 31st July 2016.

In addition, this attraction offers some discounts for young people (< 30 years), retired adults (≥ 65 years), disabled and so on.

3.2 The Data

As it is known, data plays a key role in every data analysis process. To begin with, it is very important to understand the objectives of the project in order to get the required information from sources. We should be able to structure this information in case we get it unstructured. After that, a study of data variables is important in order to understand its properties and set the analysis according to it.

3.2.1 Structure of TripAdvisor webpage

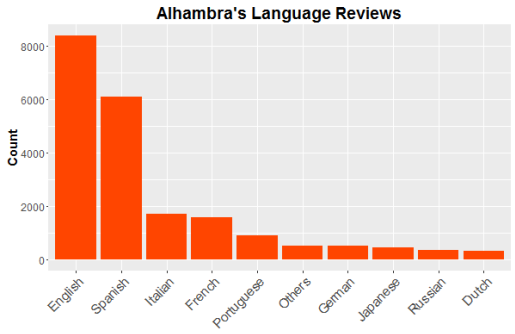
After the introduction to TripAdvisor and the Alhambra, we explain the main structure of the website which we scrap[7].

As it is shown in Figure 3.5, the first given information is the total average ranking (4.5/5), the total number of reviews (20,832) and the position in the *Popularity Index* ranking (1/194 things to do in Granada). As a result, TripAdvisor shows that the Alhambra is the first traveller's choice in 2016, i.e., the best attraction to visit in the city.[8]

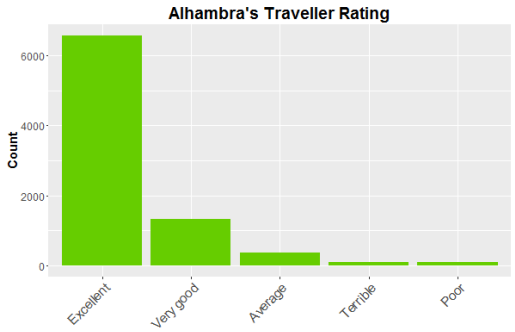
The most popular language is English, with a 39.08% of the total (Figure 3.4a). After that, Spanish is the second one, 29.37%. Considering all English reviews before 1 August 2016, TripAdvisor users show to be very satisfied with the Alhambra. Taking a glance to Figure 3.4b, a 93.73% of overall reviewers rates the visit as *Excellent* or *Very good*.

The main page is divided in five sections:

[7]Web scraping is a computer software technique of extracting information from websites.
[8]Data extracted on 31 July 2016.



(a) Barplot of total of reviews by language.



(b) Barplot of travellers visit rating.

Figure 3.4: Barplots from Alhambra's TripAdvisor webpage.

- **Overview**: This is the first section. It is shown all visitors photos, some tour prices and Alhambra's address and localization.
- **Tours & Tickets**: This is the travellers' top experiences section. Here, the user can book some tourist guide.
- **Reviews**: In this section we can read all visitors' opinions.
- **Q&A**: Questions and Answers. There, users ask questions about the attaction that other users may answer.
- **Location**: It is shown some other top-ranked attractions or restaurants nearby the Alhambra.

As we want to study Alhambra's visitors opinion, we scroll down to the *Review* section (see Figure 3.6). In this part, it appears an opinion world filter. Below, there is an another filter for traveller rating, traveller type, time of year and language of reviews. On the right side, we find information about our Facebook friends activity in this website.

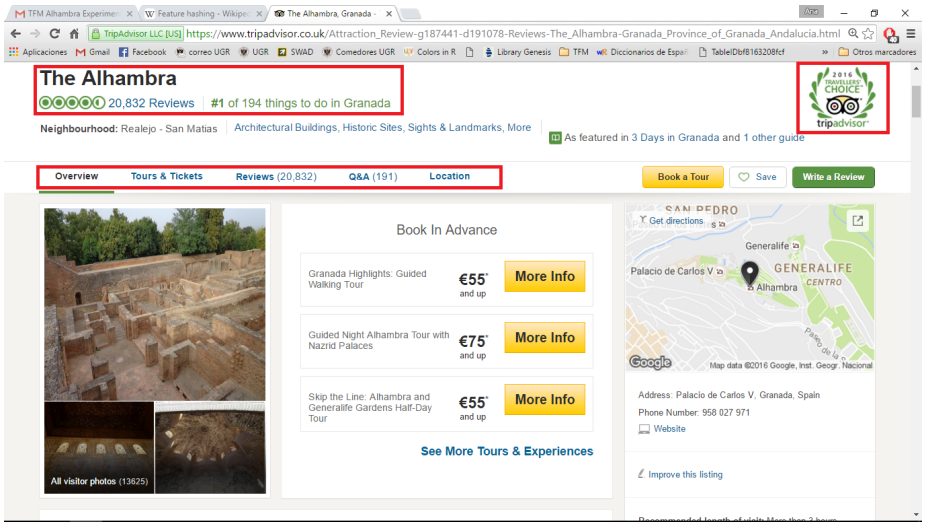


Figure 3.5: The Alhambra in TripAdvisor. Screenshot taken on 31 July 2016.

At the end of this section we can read the opinions. As we can see in Figure 3.7, TripAdvisor provides the review as well as user's information. Regarding to the opinion, it is shown the title, the user rating, the reviewing date, the channel (mobile or not) and the text. Also, we can vote about the helpfulness of the review. Moreover, we can know the TripAdvisor nickname of the user, location, level of contribution, the overall of reviews written, the number of attractions reviewed and the total of helpful votes.

3.2.2 Scrapping TripAdvisor

The most important task in data analysis is to get the data. It may seem to be a trite step, but experience teach you that it is not. For this purpose is important to settle the objectives of the research so as to collect necessary information. As we have explain in previous chapters, our purpose is to apply sentiment analysis into the Alhambra. Therefore, we settle down TripAdvisor as our source and download all information reviews of this monument from this webpage. To do so, we develop a code in R software with package `rvest` which extracts information from HTML and XML codes. For more details, code is shared in `https://github.com/anavaldi/TFM`.

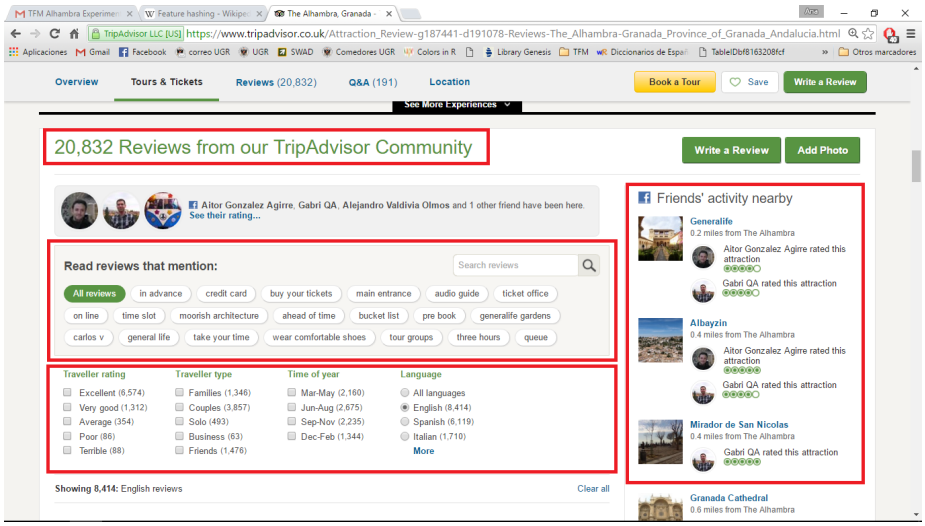


Figure 3.6: The Alhambra Review section. Screenshot taken on 31 July 2016.

3.2.3 TripAdvisor Alhambra data set

The attributes

The raw data has 8,140 instances[9] and 8 features. The description of features is as follows:

- `id`: Code of TripAdvisor user.
- `username`: Name of the user.
- `location`: Location of the user.
- `userop`: Number of written review by the user.
- `quote`: Title of the review.
- `review`: Text of the review.
- `titleopinion`: Title and text of the review.
- `rating`: Evaluation of user visit from 1 to 5.
- `date`: Date on which the user wrote the review.
- `page`: Page number of the review.

[9] All English reviews until 30 June 2016.

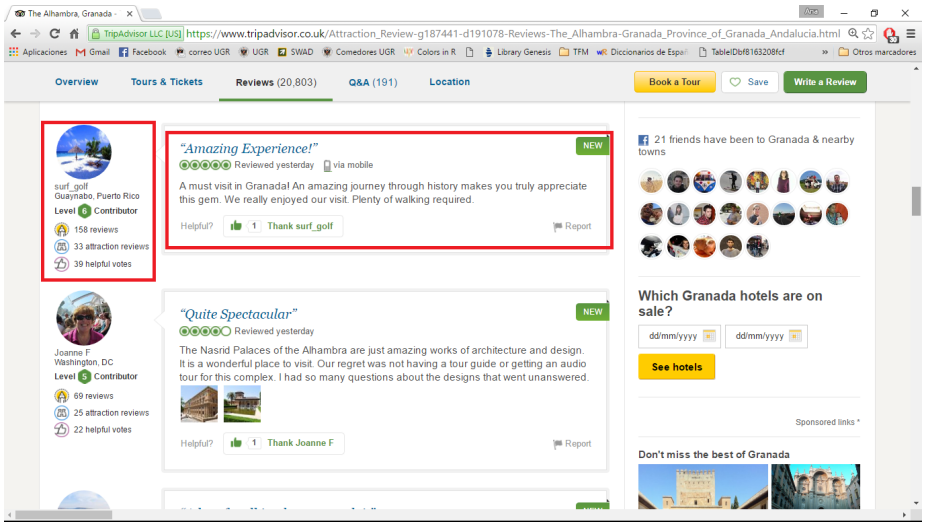


Figure 3.7: The Alhambra Review section II. Screenshot taken on 31 July 2016.

The `SentimentValue` class label

We create `SentimentValue` variable based on `rating` attribute: if `rating` is $\leq$ 2 the value is *negative*, if it is equal to 3 is *neutral* and $\leq$ 5, *positive*. We make a reference to this column hereinafter as the *expert sentiment* or *user sentiment*.

The `SentimentCoreNLP` class label

The variable `SentimentCoreNLP` shows the sentiment of reviews obtained from the Stanford CoreNLP Natural Language Processing Toolkit[Manning et al., 2014]. This is an open source NLP toolkit developed in Java by the Stanford NLP Group which provides:

- tokenization,
- sentence splitting,
- part-of-speech (POS) tagging,
- morphological analysis,
- named entity recognition[10],
- syntactic parsing,

[10] *Named Entity Recognition* (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names.

- coreference resolution[11],
- **sentiment analysis**.

Hence, we use this toolkit in order to extract the sentiment of our reviews. The sentiment analysis mechanism is described in [Socher et al., 2013]. Briefly, they developed a deep learning algorithm, Recursive Neural Tensor Network (RNTN), which outperforms old methods in different metrics. This algorithm extracts sentences sentiment representing a phrase through word vectors and a parse tree. What is new is that, unlike BOW, the algorithm is capable of capturing sentiment changes detecting scope negations and contrastive conjunctions like *but* (see attached Figure 3.8).

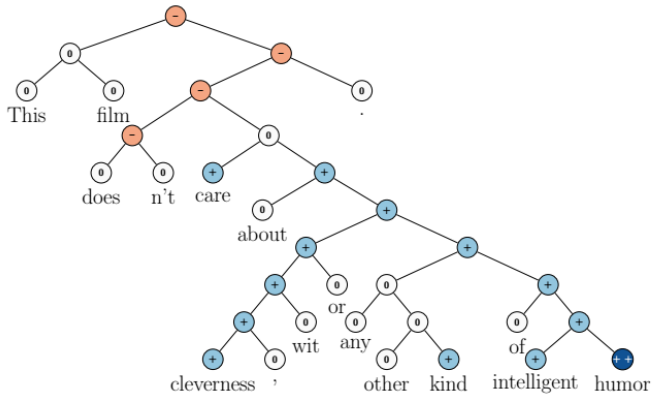


Figure 3.8: Example of RNTN extracted from [Socher et al., 2013]. The algorithm predicts 5 sentiments, from very negative (– –) to very positive (++). As it is shown in the example, the algorithm captures the negation.

We evaluate the sentiment of each sentence in a review and give the total score to the winner sentiment, recorded in `SentimentCoreNLP` class label. In case of tie, the review is set as neutral. The code implementing this process can be read in `https://github.com/anavaldi/TFM`.

This variable is refered as *automatic system sentiment* or *machine sentiment*.

Data description

As we have explained before, TripAdvisor is a webpage that has been growing over last decade. Concurrently, the number of opinions has been increasing. This fact can be contrasted in Table 3.2 which shows the number of English opinion about the Alhambra:

[11] *Coreference resolution* is the task of finding all expressions that refer to the same entity in a text.

Year	Total Opinions
2003	3
2005	1
2006	6
2007	17
2008	15
2009	42
2010	43
2011	267
2012	1,119
2013	1,190
2014	1,619
2015	2,570
2016 (until 30 June)	1,248

Table 3.2: Number of English reviews by year.

Once we have analysed the distribution of opinions over the years, it is interesting to study the distribution of ratings. As we can see in Figure3.9, we plot the average distribution since 2012, when the opinion number is higher. We can conclude that Alhambra is very well worth over TripAvisor users despite of isolated peaks in April 2012 and March 2016.

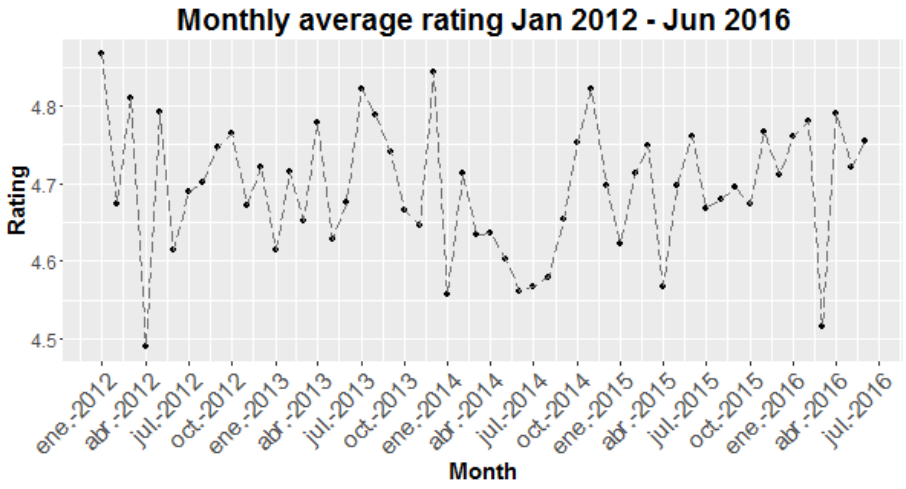


Figure 3.9: Rating averages by month since January 2012 until June 2016.

We would like to study more properties about variables such as user location (`location`) or user opinion number (`userop`), but these studies are beyond the objectives of this project. We propose the analysis for further work.

Chapter 4

The Classification Problem

The computer experiences about the classification problem are presented in this chapter. The first objective of this study is to compare the polarity of sentiment label between experts and automatic system evaluation by Stanford University, i.e., `SentimentValue` and `SentimentCoreNLP`. The second aim is to evaluate the perfomance of classification algorithms with Alhambra's TripAdvisor data set.

Firstly, we develop an analysis of the correlation polarity between the two sentiment labels in 4.1 **Polarity Label Sentiment Analysis** section. Secondly, we develop two methods in order to extract features in 4.2 **Feature Selection**. Lastly, in 4.3 **Classification Analysis**, we apply several classification algorithms in order to evaluate their performance discussing its results.

Lastly, we should comment that all experiment are developed in R software, version 3.3.1.

4.1 Polarity Label Sentiment Analysis

In this section, we analyse the correlation between users (experts) and automatic algorithm[1] sentiment labels. As a first peek, we count the number of reviews and the average of characters by (`SentimentValue`) class label. As we can observe in Table 4.1, we obtain a very unbalanced data set taking account users evaluation. The IR[2] of positive and negative is 45.14. Observing the average of characters, negative opinions have twice the average number of positive. It seems that people who is not satisfied tend to write long reviews.

[1] CoreNLP algorithm

[2] *IR* is defined as the ratio of the number of instances in the majority class to the number of examples in the minority class.

SentimentValue	# reviews	Averge # chars
positive	7628	527.34
neutral	343	728.82
negative	169	1,050.19

Table 4.1: Number of reviews and average number of characters by `SentimentValue`

Next, we evaluate the correlation between `SentimentValue` and `SentimentCoreNLP`, the users and the computer sentiment evaluations. For implementing the CoreNLP toolkit in R, we use this GitHub code: `https://github.com/statsmaths/coreNLP`.

	SentimentCoreNLP			
SentimentValue	positive	neutral	negative	Total
positive	4,049	1,071	2,508	7,628
neutral	51	32	260	343
negative	5	6	158	169
Total	4,105	1,109	2,926	8,140

Table 4.2: Correlation between `SentimentValue` and `SentimentCoreNLP`

As we can observe in Table 4.2, the correlation in positive reviews is low: a 53.08% of coincidence between expert and CoreNLP sentiment. It is interesting to remark that a 32.88% of positive reviews are classified as negative by the Stanford method. Analysing the average ratings of positive class in Table 4.3, we can conclude that if the algorithm CoreNLP detects more negative than positive sentences, the average of ratings are lower. This result can be constrasted in Figure 4.1: from all 4-ranked reviews, the 48.46% are classified as negative; if users rank the visit with 5, the negative evaluation drops to 29.79%.

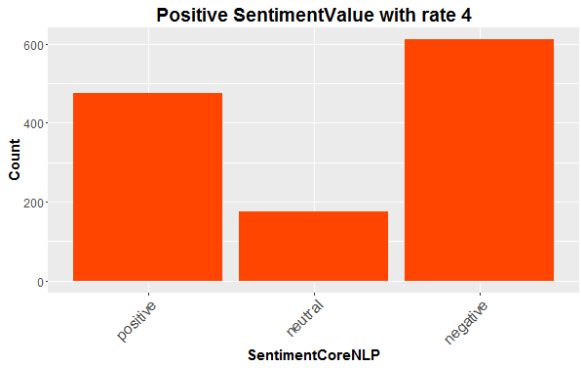
	SentimentCoreNLP		
SentimentValue	positive	neutral	negative
positive	4.883	4.836	4.756

Table 4.3: Average ratings of positive `SentimentValue`

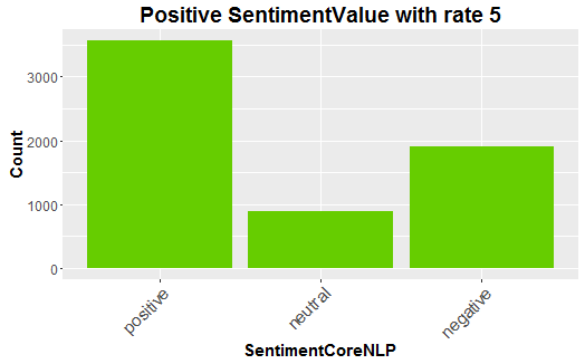
Taking a glance, it seems that in this cases the reviewer evaluates the Alhambra as *Excellent* or *Very good*, i.e., positively, but he or she reports negative sentences or gives some advices about the visit like in Figure 4.2.

In neutral opinions, the algorithm tends to evaluate the review as negative (75.80% of cases). Lastly, the correlation between negative reviews is higher, over 93.49% of matching.

From now on, neutral opinions are not going to take into account in the following experiments. The first reason is that we want to focus on the polarity of the



(a) Distribution of `SentimentCoreNLP` values with a user rate score of 4.



(b) Distribution of `SentimentCoreNLP` values with a user rate score of 5.

Figure 4.1: Distribution of user positive reviews by `SentimentCoreNLP`.

opinions, i.e., positive and negative. The second reason is that this research is setup as an introduction of sentiment analysis to tourism domain, so we want to lay the foundations and make a basic research in order to develop futures research lines. Thus, ignoring neutral class will help us in this task.

4.2 Feature Selection

The purpose of this section is to present different methods that we develop in order to extract features. As it was defined in last chapter, feature selection is the main step in order to classify sentiments because it transforms unstructured data into structured data. We develop two statistical methods that builds BOW with the most relevant words taking into account $tfidf$ measure and `SentimentValue` class label.

Figure 4.2: Review example of unalike sentiment between the expert and the coreNLP algorithm.

4.2.1 Unigram Feature Selection Method (UFSM)

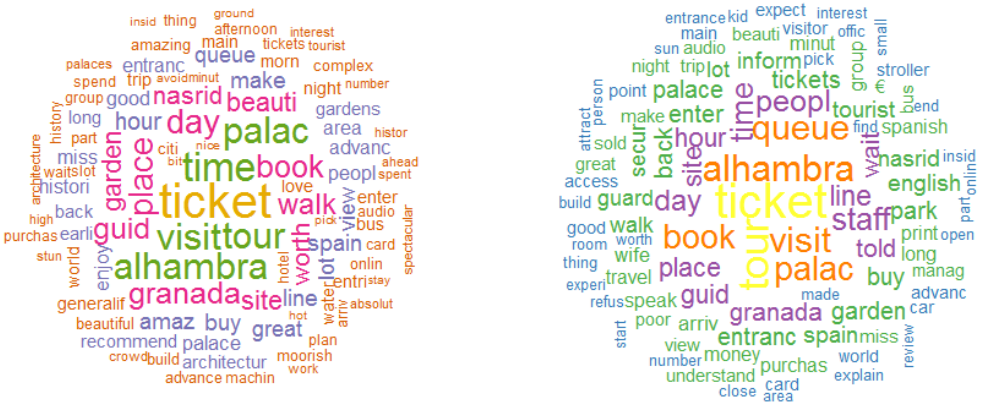
This is a first approach to convert our data set into a document term matrix. We develop a method that computes the most relevant unique words in positive and negative opinions, computing the average of $tfidf$ in each set of opinions. Firstly, it is necessary to preprocess the text, that is: remove stop words, stemming, etc. After that, we select the top 500 from one sentiment and another. We build a matrix, $M^{n\times m}$, which n is the total number of opinions and m the selected unigrams. The element a_{ij} is equal to 1 if i-unigram appears in j-document (in our case, documents refers to opinions).

Unigram Feature Selection Method (UFSM)

1. Split reviews in two sets: positives and negatives.
2. For each set, do:
   i) Preprocess the text (stemming, remove stop words, punctuations and numbers. Extract unigrams.
   ii) Compute $tfidf$ unigram of each review.
   iii) Compute average of $tfidf$ by total reviews.
   iv) Select 500 most important terms or unigrams of each set and merge both results.
3. Build the document term matrix.

Implementing this algorithm in our data set, i.e., `titleopinion` variable, we obtain 2,559 and 3,092 unigrams in positive and negative reviews. The most

prominent words are displayed in 4.3a and 4.3b for positive and negative sentiments, respectively. After that, we select top 500 and obtain 684 features. Therefore, both sentiments share 316 unigram features. In this step, we realise that some of this features are not stemmized correctly or different words represent the same idea, e.g., {*visit*, *visited*, *visitor*} or {*nasrid*, *nazari*, *nazrid*}, so we do it manually. Finally, we obtain 584 features.



(a) Word cloud of positive unigrams.

(b) Word cloud of negative unigrams.

Figure 4.3: Word clouds of unigrams.

We show the worldcloud of both sentiments in Figure 4.3. As we can observe in the images, both sentiments share a lot of common unigrams (316): *ticket*, *alhambra*, *tour*, *time*, ... Actually, the most important tag is *ticket* even before of *alhambra* and *granada* with a *tfidf* value of 1.141 and 1.457 for positive and negative opinions. If we analyse opinions, TripAdvisor users usually warns about ticket process.

In general, TripAdvisor users that rate positively tend to talk about the beauty of this monument as well as its gardens. By contrast, people complains about long queue time in ticket office and not helpful nor bilingual staff.

However, this method has some disadvantages. To begin with, it does not capture negative expressions. For instance, the term *good* is more common in positive reviews and *not good* in negative opinions, then it captures *good* in both cases and does not takes into account negative word *not*. As another drawback, it does not distinguish negative aspects or features if the global opinion is ranked positively and viceversa.

Positive			Negative		
unigram	***tfidf***	**freq**	**unigram**	***tfidf***	**freq**
ticket	1.141	7,078	ticket	1.457	366
time	0.939	5,144	tour	1.278	117
visit	0.870	6,259	alhambra	1.089	243
alhambra	0.862	6,329	visit	1.049	202
tour	0.852	3,379	queue	1.024	79
palac	0.835	4,171	palac	1.008	114
day	0.749	3,294	book	0.983	106
place	0.741	3,775	staff	0.937	71
book	0.725	2,939	time	0.929	148
granada	0.699	3,147	peopl	0.872	91

Table 4.4: Value of *tfidf* and frequencies of top 10 unigrams for positives and negative reviews.

4.2.2 Bigrams Feature Selection Method (BFSM)

Bigrams are sets of two words that appear consecutively in text. These 2-terms allow to capture negation words as well as more text information. Regarding to UFSM method, we develop a similar method for bigrams.

Bigram Feature Selection Method (BFSM)

1. Split reviews in two sets: positives and negatives.
2. For each set, do:
   i) Preprocess the text (stemming, remove stop bigrams (both words are stop words), punctuations and numbers). Extract bigrams.
   ii) Compute $tfidf$ bigrams of each review.
   iii) Compute average of $tfidf$ by total reviews.
   iv) Select 500 most important bigrams each set and merge both results.
3. Build the document term matrix.

Executing the code, we obtain 7,514 bigrams for positive reviews and 13,790 for negative. Selecting top 500 and joining results, we obtain a total of 752 bigram features. So, they share 248 tags.

As we can observe in bigram wordclouds 4.4, *the alhambra* tag is emphasised in both sentiments. That means that this bigram has higher *tfidf* and frequency

(a) Word cloud of positive bigrams.

(b) Word cloud of negative bigrams.

Figure 4.4: Word clouds of unigrams.

value compared with other features (see Table 4.5). As well as in unigram case, the most important tags are shared in both sentiments {*the alhambra*, *the palac*, *the ticket*}.

Positive			**Negative**		
bigram	***tfidf***	**freq**	**bigram**	***tfidf***	**freq**
the alhambra	0.583	4,449	the alhambra	1.053	178
the palac	0.215	1,639	the ticket	0.473	80
to visit	0.180	1,340	the palac	0.325	55
the ticket	0.170	1,293	to visit	0.320	54
your ticket	0.169	1,286	visit th	0.207	35
alhambra i	0.154	1,173	the garden	0.190	32
the garden	0.153	1,170	tickets for	0.183	31
in adv	0.141	1,080	the nasrid	0.178	30
the nasrid	0.139	1,063	the day	0.171	29
in granada	0.130	995	alhambra i	0.170	28

Table 4.5: Value of *tfidf* and frequencies of top 10 bigrams for positives and negative reviews.

4.3 Classification Analysis

In this section, we analyse the performance of several machine learning classification algorithms in a supervised problem. The aim of this section is to study

its results regarding three different class labels sets. We report several measures so as to evaluate algorithms robustness.

Firstly, we describe the different data sets that we use in the classification problem. After that, we explain the classification algorithms and predicition measures. Finally, we report the results and analyse them.

4.3.1 The three sets

In this section we define the three sets where we apply machine learning techniques. The aim of this three sets is to analyse the behaviour of the classification algorithms depending on different sentiment class labels.

The first set that we describe is the one with `SentimentValue` as the class label. In this case, the sentiment value is provided by the expert, i.e., the TripAdvisor user; the second set takes `SentimentCoreNLP` as the class label, so the CoreNLP algorithms evaluates review sentiment taking into account its content; in the third set, the class label refers to the set of instances that have same sentiment in both class labels, `SentimentValue` $\bigcap$ `SentimentCoreNLP`.

These three sets become six in total, because we build three with unigrams and three more with bigrams. In the below Table 4.6, it is shown the total number of instances and features of each set:

Data set	Class Label	Rows	Features
Set 1	`SentimentValue`	5,040	584 (unigram)
Set 2	`SentimentCoreNLP`	5,040	584 (unigram)
Set 3	`SentimentValue` $\bigcap$ `SentimentCoreNLP`	3,155	584 (unigram)
Set 4	`SentimentValue`	5,040	738 (bigram)
Set 5	`SentimentCoreNLP`	5,040	738 (bigram)
Set 6	`SentimentValue` $\bigcap$ `SentimentCoreNLP`	3,155	738 (bigram)

Table 4.6: Number of instances and features of the different data sets.

4.3.2 Machine Learning Algorithms and Classification Measures

In this section we present a brief description of the classification algorithms that we implement in our supervised problem.

J48

J48 is a decision tree[3] algorithm. This method starts with a train set, on each iteration it computes the *information gain* of each not selected attribute, a concept that measures the amount of information contained in a set of data. It gives the idea of importance of an attribute in a dataset. Then, it selects the attribute with the higher value and splits the set by this attribute. The algorithm continues iterating on each splitted set and never selecting used attributes.

Support Vector Machines (SVM)

The basic idea of Support Vector Machines is to determine linear separators in the space which best separate the different classes examples. Thus, new data is classified depending on the side they fall on.

Consider a data set where it contains information about two gens, gen X and gen Y, and the class label are normal patient or cancer patient. The aim is to classify new patients depending on their gen information. As it is seen in Figure 4.5, the data is mapped in the space as vectors. Then, SVM tries to find the linear decision surface between two classes as wide as possible. If it is not possible to find linear decision surface, the data is plotted into higher dimensional space.

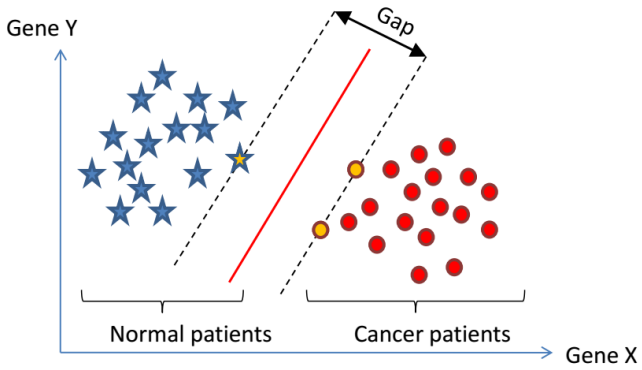


Figure 4.5: Representing patients geometrically.

The best property of this algorithm is that it was developed under a rigorous mathematical theory. In this way, there always exists a unique solution.

[3]A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. The goal is to predict the value of a target variable based on several input variables.

eXtreme Gradient Boosting (XGBoost)

XGBoost is an ensemble machine learning algorithm developed as an open-source[Chen and Guestrin, 2016]. This algorithm is defined in its github webpage as:

"XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting[4] *framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way."*

To sum up, the outline of the algorithm is to grow a tree with a maximum depth, find the best splitting point and assign weight to the two new leaves. After that, prune the tree to delete nodes with negative gain and iterate this process n-times.

The main properties of this model are: easy to use (easy to install and highly developed in R), efficient (automatic parallel computation on a single machine, can be run on a cluster), accurate (good results for most data sets) and feasible (customized objective and evaluation and tunable parameters). Through last year, this algorithm has become very popular due to its excellent results in many data science competitions[5]. The unique drawback is that this algorithm depends on nine parameters which needs to be optimized.

Measures

In order to study the robustness of the classification model, there exist different measures that reflect it. Therefore, it is reported a sample of measures that we use:

- **Confusion matrix**: Matrix of 2×2 dimension that displays the number of *true positive* (TP), *true negative* (TN), *false positive* (FP) and *false negative* (FN) classified instances.
- **Accuracy**: Proportion of correct classified.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Positive Precision**: Proportion of classified positives which are actual positive.

[4]Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

[5]Source:kdnugggets

$$positivePrecision = \frac{TP}{TP + FP}$$

- **Negative Precision**: Proportion of classified negatives which are actual negative.

$$negativePrecision = \frac{TN}{TN + FN}$$

- **Specificity or True Negative Rate (TNR)**: Proportion of actual negatives which are classified negative.

$$specificity = \frac{TN}{TN + FP}$$

- **Sensitivity, Recall or True Positive Rate (TPR)**: Proportion of actual positives which are classified positive.

$$sensitivity = \frac{TP}{TP + FN}$$

- **F-measure**: Harmonic mean between precision and recall.

$$Fmeasure = 2 \cdot \frac{precision \times recall}{precision + recall}$$

- **G-measure**: Geometric mean between precision and recall.

$$Gmeasure = \sqrt{precision \times recall}$$

- **Receiver Operating Characteristic Curve (ROC Curve)**: Curve that is created by plotting the TPR against the false positive rate (FPR) ($\frac{FP}{FP+TN}$) at various threshold settings.
- **Area Under the Curve (AUC)**: It is the area, i.e., integral value, under the ROC curve. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

4.3.3 Methodology

Before showing our experiment results, we find necessary to explain the developed model process (see Figure 4.6).

As a first step, we download all reviews from the source, TripAdvisor. After that, we build two sentiment class labels: `SentimentValue` and `SentimentCoreNLP`. The first one is the sentiment expressed by the reviewer and the second one is the sentiment evaluated by the CoreNLP algorithm. After that, we structure data in document term matrix, computing unigrams and bigrams by UFSM and BFSM. We build three different sets, or six taking account unigrams and bigrams, depending on the sentiment class label. We split them in train set (75% of instances) and test set (25% of instances) in order to evaluate the performance of predictions. In some cases, we apply oversampling so as to adjust the class distribution. We randomly resample instances of the minority class. Finally, we apply different classification algorithms for supervised models. All experiments are generated with a 5 cross validation (5cv), that is, we remove a part ($\frac{1}{5}$ of the train set in order to evaluate the model in this unlearnt set. Then we rise the last step: we predict on the test set and compute model measures so as to decide the best performance.

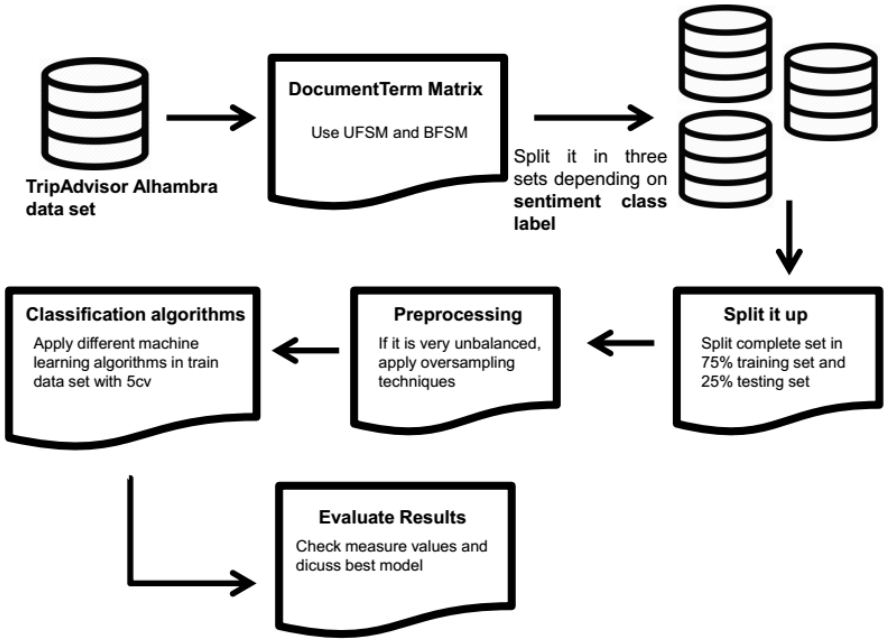


Figure 4.6: Framework of Classification Model Process.

4.3.4 Results

All experiments have been developed in R software. In order to optimise model parameters, we apply `caret` package. The `caret` package (Classification And REgression Training) is a set of functions that attempt to provide a uniform interface for constructing predictive models. In the packages, we can find several functions for creating models like: data splitting, pre-processing, feature selection, model tuning using resampling, variable importance estimation as well as other functionality. Code is shown in `https://github.com/anavaldi/TFM`.

We emphasise that positive class is the majority class and positive opinions; and negative class is the minority class and negative opinions.

In total, we create 72 experiments: we apply three different algorithms (SVM, J48 and XGBoost) on six data sets (see Table 4.6); we balance sets 1, 3, 4 and 6[6] resampling negative instances.

To begin with, we discuss general results by items:

Unigrams vs. Bigrams: As we can observe in A tables, unigrams generally perform better results than bigrams. Analysing the bigram sample, we realise that some of them share concept information. For instance:

alhambra a	alhambra and	alhambra at	alhambra but
alhambra from	alhambra i	alhambra in	alhambra it
alhambra itself	alhambra on	alhambra palac	alhambra th
alhambra to	alhambra visit	alhambra w	alhambra wa
alhambra with	alhambra you		

Therefore, we conclude that bigrams need more preprocessing.

Overfitting: Overfitting is greatly extended in our experiments. The 88.89% of the overall has a test specificity (SpecTEST) lower than 0.5 whereas AUC average is 0.827. This fact means that our models are overlearning train data which leds to a very poor performance in test.

Oversampling: We conclude that oversampled experiments tend to obtain better results than raw sets although some of these experiments are overfitted.

If we analyse results by algorithms (see Figure 4.7):

J48: We realise that models with J48 have a very wide range of AUC values. That means that this decision-tree algorithm performs non robust models.

SVM: SVM has quite high AUC values both in unigram either in bigram experiments, but SpecTEST values under 0.5 both in unigrams either in bigrams.

[6]Train sets 1 and 4 has an IR of 37.473; train sets 2 and 6, 24.040.

XGBoost: This algorithm has lower AUC and SpecTEST mean than SVM, but it has a maximum worth over all SVM SpecTEST values.

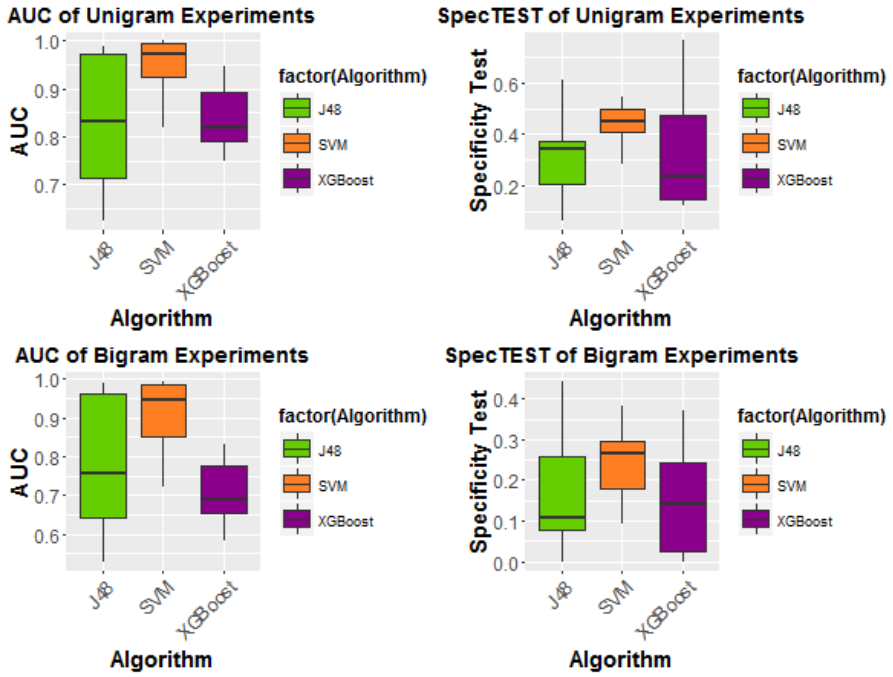


Figure 4.7: Model performance by algorithms. It is evaluated AUC and SpecTEST by features method.

Therefore, AUC and SpecTEST values in unigram experiments are always greater than in bigram experiments.

After that, we report the three best performance models sorted by AUC and SpecTest values (Figure 4.8):

1st Model (53th experiment): We conclude that best model performance is the number 53. This model is built with XGBoost algorithm and the oversampled set 1, adjusting `SentimentValue` classes to the same number of instances. It rises a 0.914 area under ROC curve and 0.763 specificity value in test. This model is basically the best because it rises a specificity value in test well above other results, which make us belive that there is no overfitting.

2nd Model (59th experiment): The second best result is experiment number 59. This model learns set 3 (`SentimentValue` $\bigcap$ `SentimentCoreNLP` with XGBoost algorithm too. The set is balanced as before. Numerically, it rises a 0.948 AUC and 0.594 test specificity.

3rd Model (7th experiment): Lastly, the third best model is experiment 7.

In this case, the model learns set 2 with a J48. The set is not balanced because `CoreNLP` class label is well balanced (the IR in train set is 1.526). However its AUC value is not really high, 0.679, it classifies correctly a 61.10% of actual negatives. Similar to this result is experiment 55. Evaluating same set with XGBoost, it obtains 0.752 AUC and 59.46% of true negative ratio percent.

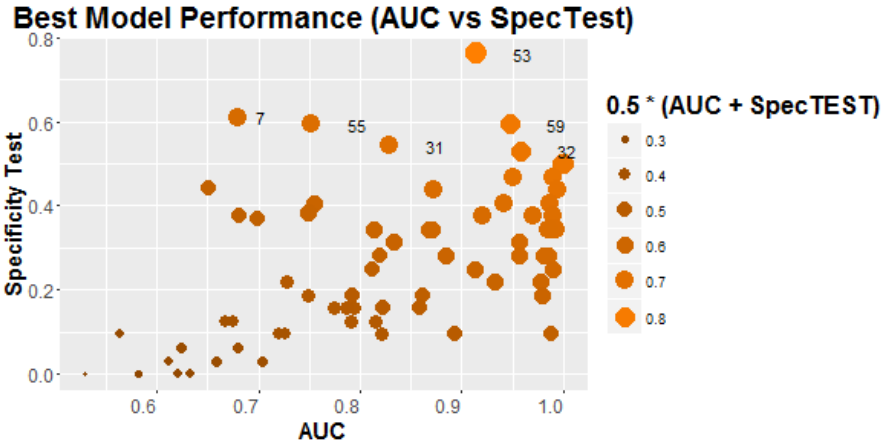


Figure 4.8: Model performance evaluated by: $0.5 \times (AUC + SpecTEST)$. Labels represent the experiment number.

Finally, we would expect to obtain better results with set 3 and 6 (`SentimentValue` $\bigcap$ `SentimentCoreNLP` as class label), because in this sets both sentiment labels (user and machine) match. However, as we have analysed, the difference between set results is tiny.

Chapter 5

Subgroup Discovery

The aim of this chapter is to describe negative opinions by its sentiment. To do so, we apply subgroup discovery, SD, methodology. This technique is aimed at discovering implicity patterns in data that describes subgroups given a target class. Therefore, at the end of this section we will be able to describe what TripAdvisor users dislike from Alhambra's visit.

This chapter is organised as follows. Section 5.1 **Introduction** is presented as a brief introduction to subgroup discovery. In following section, 5.2 **Subgroup Discovery Algorithms and Measures**, we develop a brief description of common subgroup discovery algorithms and measures. In 5.3 **Methodology**, we present the process developed in order to apply SD techniques in negative reviews. Discovered subgroups and conclusions are discussed in 5.4 **Results** section.

5.1 Introduction

As it is found in [Herrera et al., 2011], *Subgroup Discovery* is a data mining technique that discovers interesting relationships between attributes and a certain value of the target variable.

This concept was defined by [Klösgen, 1996] and [Wrobel, 1997] for the first time twenty years ago. The definition was as follows:

"Given a population of individuals and a property of those individuals we are interested in, find population subgroups that are statistically most interesting, e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

In subgroup discovery, rules have the following form:

$$R : Cond \longrightarrow Target_{value}$$

where the rule antecedent, $Cond$, is a conjunction of selected features (attribute and value pairs) describing training instances; and the rule consequent, $Target_{value}$, is the target of the variable of interest.

Subgroup discovery task attempts to cover instances from data with a fixed class label in an interpretable way. It does not focus on finding complex relations. In this sense, it is more important to find small subgroups with interesting characteristics, as seen in Figure 5.1.

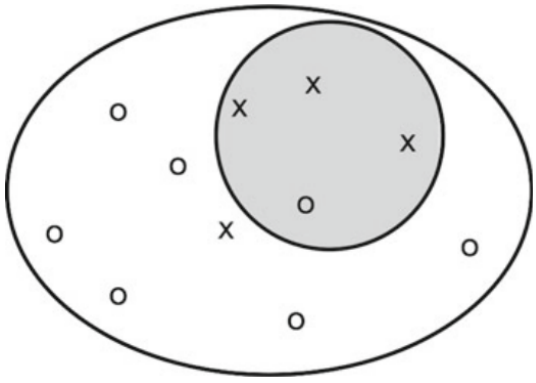


Figure 5.1: Visualization of subgroup discovery rule for $Target_{value} = \times$. Source: [Herrera et al., 2011].

Thus, subgroup discovery differs from classification methods because it tries to extract basic knowledge from data whereas a classification model tries to perfom a complex and accurate performance so as to predict new values. Moreover, this branch differs from association rules in the sense that subgroup discovery *punishes* rules with same properties in the antecedent, but different class label in the consequent.

Because of these facts, subgroup discovery produces more construable information than other methodologies. Therefore, we feel motivated to implement this method to our data and extract valuable information about disappointed tourists for touristic attraction managers.

5.2 Subgroup Discovery Algorithms and Measures

5.2.1 Algorithms

As it is described in [Herrera et al., 2011], there exist three main groups of algorithms for subgroup discovery: extensions of classification algorithms, extensions of association algorithms and evolutionary algorithms. The first group contains those methods that have been developed by adaptations of classification rule learners. EXPLORA, MIDOS, SubgroupMiner, SD, CNS2-SD and RSD are some examples of this group. The methods that are modifications of association rule algorithms are grouped in the second group. In this case, the consequent of the rule is a fixed label class. Some of this algorithms are APRIORI-SD, SD4TS, SD-Map, DpSubgroup, Merge-SD and IMR. The last group contains evolutionary and heuristic algorithms which are implemented for extracting subgroups. SDIGA, MESDIF, NMEEF-S are evolutionary algorithms that belong to this third group.

SD-Map

Here, we briefly develop an explanation of SD-Map because we will focus our experimentation on this algorithm.

As we have explain before, SD-Map is based on association rule algorithms proposed in [Atzmueller and Puppe, 2006]. More specifically, SD-Map is an exhaustive search method based on an adapted version of FP-growth[1]. It depens on a minumum support threshold in order to perform an efficient search pruning the space of features.

The main concern of this algorithm is its inefficiency covering the whole domain of rules.

5.2.2 Measures

Measures are used to extract and evaluate the rules. There exist a widely collection of measures in the literature. They can be classified by objectives: complexity, generality, precision and interest. In our experiments, we evaluate our subgroups by the Binary-Test, a precision quality function:

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}}$$

[1]FP-growth is an efficient approach for mining frequent patterns which performs a frequent pattern tree known as FP-tree that stores count information about the frequent patterns.

where,

- p is the relative frequency of the target variable in the subgroup,
- p_0 is the relative frequency of the target variable in the total population,
- N is the size of the total population,
- n is the size of the subgroup.

5.3 Methodology

As it is illustrated in Figure 5.2, we apply SD-Map over the set 3 (`SentimentValue` $\bigcap$ `SentimentCoreNLP`) because sentiment label is more consistent due to the fact that we filter out those instances with different sentiment label. After that, in order to perform an efficient search, we select a set of features through J48 algorithm tree and feature correlation[2]. We then set our study based on the negative class, i.e., the negative reviews and minority class. Finally, we discuss resulting subgroups so as to extract valuable information about disappointed users.

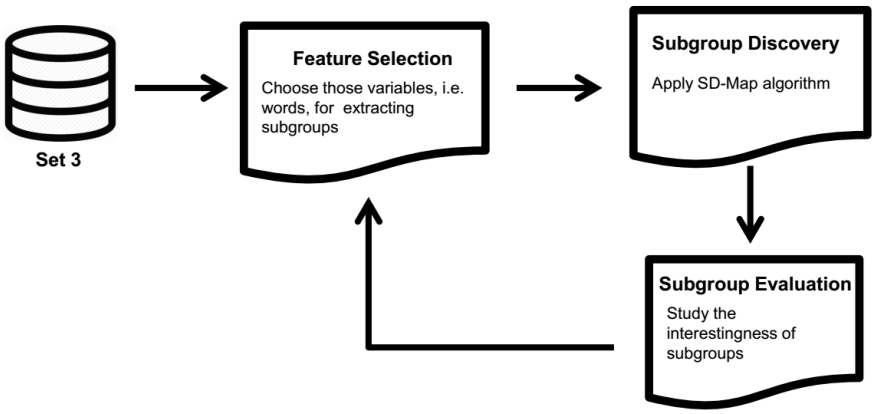


Figure 5.2: Framework of Subgroup Discovery Process.

5.4 Results

We develop this experiment in R software. In there, we use `rsubgroup` package which contains SD-Map and was developed by Martin Atzmueller. This

[2]The classification experiment of the set 3 with J48 is number 12. As we can check in A.1, the result of this model is not good at all. However, we decide to select this algorithm because we can select the most important features through its tree representation.

package is defined as *"a collection of efficient and effective tools and algorithms for subgroup discovery and analytics.* We choose this package due to its easy usability and fast computation. The developed code can be found in `https://github.com/anavaldi/TFM`.

As a first approach, we select a set of 28 features extracted from the J48-tree of 12th-experiment (see Table A.1). We are awared that this experiment has not the best performance. The subgroups extracted are:

q_{BT}	p	n	**Description**
45.24	0.49	100	told=1
45.11	0.06	2,500	beautifulX=0
42.39	0.48	96	told=1, knowX=0
41.73	0.52	87	told=1, enjoy=0
41.43	0.47	95	told=1, signX=0
41.43	0.47	95	told=1, basic=0
41.39	0.47	96	told=1, situatX=0
41.31	0.06	2,255	beautifulX=0, enjoy=0
40.89	0.05	2,426	beautifulX=0, knowX=0
40.81	0.52	85	told=1, enjoy=0, knowX=0

Table 5.1: Subgroups generated from the J48 feature set: {*told, bad, staff, signX, sevillaX, access, front, situatX, sold, entranceX, disappointX, knowX, enjoy, beautifulX, basic, problem, sell, onlineX, touristX, explain, wait, arriv, order, enter, websit, close, guideX, travel*}).

As we can observe in Table 5.1, although q_{BT} values, the subgroups extracted are not useful in the sense that they do not provide interesting information. We realise that this algorithm works better with small sets. Therefore, it is needed to change the feature selection strategy.

As a second approach, we use feature correlation in order to extract subgroups from small feature sets. In this way, we compute the matrix correlation from set 3 and order it. We then select high correlated features that we think that can provide better results.

Table 5.2 displays the results from selecting attributes base on their correlation. Analysing discovered subgroups parameters, we observe that generally group sizes are tiny; in big sets, probabilities tend to be smaller. However, we can now interpret subgroups meaning. Studying first discovered subgroup, some people in TripAdvisor complains about a guard attitude. In the same line, users evaluate negatively staff behaviour towards tourists as well as they do not speak English. The fact that there exist some parts of the Alhambra where is not allowed strollers or wheelchairs is another complaint. Suprisingly, queue and time do not genereate an interesting rule, maybe it is due to the fact that in positive reviews these words are common too. In the same way, night and light do not create high quality subgroups. Finally, it seems like people complains

about confirmation email after buying online tickets.

q_{BT}	p	n	**Description**
17.17	0.82	22	guard=1
16.21	0.81	21	terribl=1
12.29	0.68	19	rude=1
3.85	1	4	rude=1, guard=1
2.89	1	3	terribl=1, guard=1
2.77	0.5	6	babi=1
1.85	0.5	4	babi=1, strollX=1
1.6	0.06	64	strollX=1
30.33	0.46	71	staff=1
6.7	0.88	8	attitud=1
3.85	1	4	horribl=1
3.85	1	4	attitud=1, staff=1
3.85	1	4	horribl=1, staff=1
1.92	1	2	horribl=1, attitud=1, staff=1
30.33	0.14	284	queueX=1
27.06	0.06	1303	time=1
21.55	0.19	145	queueX=1, time=1
5.21	0.29	21	wheelchair=1
4.66	0.56	9	disabl=1
2.85	0.75	4	disabl=1, wheelchair=1
8.19	0.08	208	night=1
7.2	0.08	181	night=1, light=0
1.41	0.06	69	light=1
0.99	0.07	27	light=1, night=1
0.42	0.05	42	light=1, night=0
0.03	0.29	79	speakX=1
14.13	0.17	103	english=1
6.62	0.7	10	staff=1, english=1, speakX=1
7.62	0.8	10	email=1
5.47	0.43	14	confirm=1
3.85	1	4	email=1, confirm=1

Table 5.2: Subgroups generated from feature correlation.

Chapter 6

Conclusion

In our study we proposed a first approach to sentiment analysis into touristic attractions. To do so, we analysed data from TripAdvisor in the most popular monument in Spain: the Alhambra. Firstly, we scrapped TripAdvisor webpage and develop an exploratory study about this data. We found that the Alhambra is a very loved monument over TripAdvisor community: more than 90% of reviews was positively rated by users.

After that, we implemented a NLP toolkit, CoreNLP, so as to create an automatic sentiment target. Then, we analysed the correlation between this new label and user sentiments, `SentimentCoreNLP` and `SentimentValue` respectively. We found that 32.88% of user positive reviews were evaluated as negative by the Stanford toolkit. This result is basically obtained because people often give advices about the Alhambra trip such as: book in advance, wear comfortable shoes, etc.

In next step, we develop the prediction induction study building a classification model in order to predict new reviews. To do so, we developed two algorithms for feature selection, UFSM and BFSM, using text mining methods. As a result, we observed that most relevant words are repeated in both positive either negative reviews. We then created three sets depending on the sentiment label: the user, the CoreNLP and the matching between both. Thus, we obtained six sets (2 methods for feature selection $\times$ 3 different sets) and we applied three classification algorithms on this. The best model performance was XGBoost applied on balanced user's class label. It shown a very good result in testing. In this classification section, we realised that BFSM needs more preprocessing in order to obtain better results.

In next section, a description induction study was carried out through subgroup discovery method. This study was proposed so as to extract interesting patterns in negative reviews. We analysed only disappointed tourist because their opinion may help to improve Alhambra experience. In one hand, the subgroups that

we extract in the first apporach were meaninglessness. On the other hand, the subgroups in second approach does not covered a wide area of negative reviews although they were easy to interpret. So, we conclude that this chapter performs a first approach between sentiment analysis and subgroup discovery, but it can be significantly improved.

During these three months, the project has helped us to immerse ourselves in the sentiment analysis problem. We have understood the main insights of opinion mining and have developed new methodologies. Moreover, we have learnt new R packages and developed R scripts for dealing with opinion text and subgroup discovery.

In this way, we realise that this area is still developing due to the rapid growth of ICT. It has been proved that sentiment analysis has a lot of potential in our society and can be applied in many fields so as to extract valuable insight. Because of that, we know that much work still needs to be done. To begin with, we propose to develop a polarity analysis which consists in knowing which aspects are related to each sentiments. ASUM, the algorithm develop in [Jo and Oh, 2011], may help in this task. After that, we suggest to mix both sentiment labels (`SentimenValue` and `SentimentCoreNLP`) in one for building more robust models. Considering neutral class for building a multi-label classification model is another proposed work. Moreover, we proposed to build a visualization software tool to provide touristic attraction managers valuable information about their visitors.

Finally, some of these proposed works will be developed during next month in order to present a proposal research project to the Alhambra directors.

Appendix A

Classification Results Tables

This first appendix shows the numeric results of the 5cv classification experiment for each algorithm. The tables show the following information:

- **ExpNum**: Number of experiment.
- **Data set**: Data set used for building the model (see Table 4.6).
- **Instances**: Number of instances included in the data set.
- **Unbalanced**: Is the data set unbalanced? If yes (Y), the new IR.
- **AUC**: Value of the area under ROC curve.
- **SensTRAIN**: Sensitivity value in training set.
- **SpecTRAIN**: Specificity value in training set.
- **Accuracy**: Value of accuracy in training set.
- **Fmeasure**: F-measure value in training set.
- **Gmeasure**: G-measure value in training set.
- **TNTrain**: Average of true negatives instances in training set.
- **FPTrain**: Average of false positives instances in training set.
- **TNTest**: True negatives instances in testing set.
- **FPTest**: Average of false positives instances in testing set.
- **SpecTEST**: Specificity value in testing set.

1.1 J48

ExpNum	Data set	Instances	Unbalanced	AUC	SensTRAIN	SpecTRAIN	Accuracy	Fmeasure	Gmeasure	TNTrain	FPTrain	TNTest	FPTest	SpecTEST
1	Set 1	5073	Y:30	0.726	0.341	0.992	0.971	0.985	0.985	1.100	2.100	3	29	0.094
2	Set 1	5154	Y:20	0.794	0.527	0.987	0.965	0.982	0.982	2.500	2.300	5	27	0.156
3	Set 1	5400	Y:10	0.913	0.807	0.975	0.960	0.978	0.978	7.300	1.800	8	24	0.250
4	Set 1	5891	Y:05	0.977	0.979	0.973	0.974	0.984	0.984	16.300	0.400	7	25	0.219
5	Set 1	9818	Y:01	0.985	1.000	0.968	0.984	0.984	0.984	50.000	0.000	11	21	0.344
6	Set 1	5040	N	0.624	0.099	0.995	0.971	0.985	0.985	0.300	2.300	2	30	0.063
7	Set 2	5040	N	0.679	0.575	0.773	0.695	0.754	0.754	22.800	16.800	410	261	0.611
8	Set 3	3180	Y:20	0.754	0.445	0.991	0.965	0.982	0.982	2.100	2.600	13	19	0.406
9	Set 3	3332	Y:10	0.870	0.703	0.980	0.955	0.975	0.975	6.400	2.700	11	21	0.344
10	Set 3	3635	Y:05	0.970	0.944	0.969	0.965	0.979	0.979	15.700	0.900	12	20	0.375
11	Set 3	6058	Y:01	0.990	1.000	0.974	0.987	0.987	0.987	50.000	0.000	12	20	0.375
12	Set 3	3155	N	0.681	0.302	0.990	0.963	0.981	0.981	1.200	2.800	12	20	0.375
13	Set 4	5073	Y:30	0.611	0.146	0.997	0.969	0.984	0.984	0.500	2.800	1	31	0.031
14	Set 4	5154	Y:20	0.704	0.331	0.990	0.959	0.979	0.979	1.600	3.200	1	31	0.031
15	Set 4	5400	Y:10	0.893	0.733	0.981	0.959	0.977	0.977	6.700	2.400	3	29	0.094
16	Set 4	5891	Y:05	0.979	0.973	0.970	0.971	0.982	0.982	16.200	0.400	6	26	0.188
17	Set 4	9818	Y:01	0.988	1.000	0.965	0.983	0.982	0.982	50.000	0.000	3	29	0.094
18	Set 4	5040	N	0.529	0.007	0.998	0.972	0.986	0.986	0.000	2.600	0	32	0.000
19	Set 5	5040	N	0.651	0.491	0.816	0.688	0.760	0.761	19.400	20.100	297	374	0.443
20	Set 6	3180	Y:20	0.667	0.272	0.991	0.957	0.978	0.978	1.300	3.500	4	28	0.125
21	Set 6	3332	Y:10	0.812	0.555	0.985	0.946	0.971	0.971	5.000	4.100	8	24	0.250
22	Set 6	3635	Y:05	0.956	0.906	0.972	0.961	0.977	0.977	15.100	1.600	10	22	0.313
23	Set 6	6058	Y:01	0.980	0.997	0.964	0.980	0.980	0.980	49.900	0.100	9	23	0.281
24	Set 6	3155	N	0.563	0.063	0.960	0.997	0.979	0.980	0.300	3.700	3	29	0.094

Table A.1: J48 results for a 5cv classification model.

1.2 Support Vector Machine (SVM)

ExpNum	Data set	Instances	Unbalanced	AUC	SensTRAIN	SpecTRAIN	Accuracy	Fmeasure	Gmeasure	TNTrain	FPTrain	TNTest	FPTest	SpecTEST
25	Set 1	5073	Y:30	0.868	0.286	0.995	0.972	0.986	0.986	0.900	2.300	11	21	0.344
26	Set 1	5154	Y:20	0.942	0.771	0.989	0.979	0.989	0.989	3.700	1.100	13	19	0.406
27	Set 1	5400	Y:10	0.986	0.951	0.987	0.984	0.991	0.991	8.600	0.400	13	19	0.406
28	Set 1	5891	Y:05	0.994	0.998	0.987	0.989	0.993	0.993	16.600	0.000	14	18	0.438
29	Set 1	9818	Y:01	0.994	1.000	0.986	0.993	0.993	0.993	50.000	0.000	14	18	0.438
30	Set 1	5040	N	0.820	0.144	0.996	0.974	0.987	0.987	0.400	2.200	9	23	0.281
31	Set 2	5040	N	0.829	0.502	0.928	0.760	0.823	0.829	19.900	19.700	365	306	0.544
32	Set 3	3180	Y:20	0.958	0.690	0.995	0.980	0.990	0.990	3.300	1.500	17	15	0.531
33	Set 3	3332	Y:10	0.990	0.931	0.995	0.990	0.994	0.994	8.500	0.600	15	17	0.469
34	Set 3	3635	Y:05	0.998	0.990	0.995	0.994	0.996	0.996	16.500	0.200	16	16	0.500
35	Set 3	6058	Y:01	1.000	1.000	0.998	0.999	0.999	0.999	50.000	0.000	16	16	0.500
36	Set 3	3155	N	0.951	0.508	0.995	0.976	0.988	0.988	2.000	2.000	15	17	0.469
37	Set 4	5073	Y:30	0.821	0.122	0.998	0.970	0.985	0.985	0.400	2.800	3	29	0.094
38	Set 4	5154	Y:20	0.933	0.229	0.994	0.957	0.970	0.970	1.100	3.700	7	25	0.219
39	Set 4	5400	Y:10	0.981	0.947	0.986	0.982	0.990	0.990	8.600	0.500	9	23	0.281
40	Set 4	5891	Y:05	0.985	0.998	0.975	0.979	0.987	0.987	16.600	0.000	9	23	0.281
41	Set 4	9818	Y:01	0.990	1.000	0.974	0.987	0.987	0.988	50.000	0.000	8	24	0.250
42	Set 4	5040	N	0.720	0.107	1.000	0.976	0.988	0.988	0.300	2.300	3	29	0.094
43	Set 5	5040	N	0.748	0.327	0.950	0.704	0.795	0.806	13.000	26.600	257	414	0.383
44	Set 6	3180	Y:20	0.862	0.299	0.993	0.960	0.979	0.980	1.400	3.300	6	26	0.188
45	Set 6	3332	Y:10	0.957	0.888	0.983	0.974	0.986	0.986	8.100	1.000	9	23	0.281
46	Set 6	3635	Y:05	0.989	0.983	0.983	0.983	0.990	0.990	16.400	0.300	11	21	0.344
47	Set 6	6058	Y:01	0.993	1.000	0.978	0.989	0.989	0.989	50.000	0.000	11	21	0.344
48	Set 6	3155	N	0.859	0.262	0.966	0.967	0.983	0.983	1.000	2.900	5	27	0.156

Table A.2: SVM results for a 5cv classification model.

1.3 eXtreme Gradient Boosting (XGBoost)

ExpNum	Data set	Instances	Unbalanced	AUC	SensTRAIN	SpecTRAIN	Accuracy	Fmeasure	Gmeasure	TNTrain	FPTrain	TNTest	FPTest	SpecTEST
49	Set 1	5073	Y:30	0.791	0.134	0.994	0.967	0.983	0.983	0.400	2.800	4	28	0.125
50	Set 1	5154	Y:20	0.791	0.192	0.996	0.958	0.979	0.979	0.900	3.800	4	28	0.125
51	Set 1	5400	Y:10	0.815	0.417	0.991	0.939	0.967	0.968	3.800	5.300	4	28	0.125
52	Set 1	5891	Y:05	0.872	0.605	0.979	0.917	0.951	0.952	10.100	6.600	14	18	0.438
53	Set 1	9818	Y:01	0.914	0.828	0.889	0.859	0.863	0.863	41.400	8.600	29	9	0.763
54	Set 1	5040	N	0.749	0.099	0.998	0.974	0.987	0.987	0.300	2.300	6	26	0.188
55	Set 2	5040	N	0.752	0.544	0.812	0.706	0.769	0.771	21.500	18.100	399	272	0.595
56	Set 3	3180	Y:20	0.822	0.385	0.994	0.965	0.982	0.982	1.800	2.900	5	27	0.156
57	Set 3	3332	Y:10	0.884	0.495	0.991	0.946	0.971	0.971	4.500	4.600	9	23	0.281
58	Set 3	3635	Y:05	0.921	0.662	0.988	0.934	0.961	0.962	11.000	5.600	12	20	0.375
59	Set 3	6058	Y:01	0.948	0.875	0.928	0.902	0.904	0.904	43.800	6.200	19	13	0.594
60	Set 3	3155	N	0.793	0.294	0.994	0.966	0.982	0.982	1.200	2.800	6	26	0.188
61	Set 4	5073	Y:30	0.621	0.085	0.998	0.969	0.984	0.984	0.300	3.000	0	32	0.000
62	Set 4	5154	Y:20	0.632	0.086	0.998	0.954	0.976	0.977	0.400	4.300	0	31	0.000
63	Set 4	5400	Y:10	0.675	0.175	0.997	0.922	0.959	0.959	1.600	7.500	4	28	0.125
64	Set 4	5891	Y:05	0.728	0.332	0.985	0.876	0.930	0.931	5.500	11.100	7	25	0.219
65	Set 4	9818	Y:01	0.813	0.666	0.878	0.772	0.794	0.797	33.300	16.700	11	21	0.344
66	Set 4	5040	N	0.581	0.039	0.998	0.973	0.986	0.986	0.100	2.500	0	32	0.000
67	Set 5	5040	N	0.699	0.415	0.872	0.691	0.773	0.778	16.400	23.200	248	423	0.370
68	Set 6	3180	Y:20	0.680	0.186	0.995	0.956	0.977	0.978	0.900	3.900	2	30	0.063
69	Set 6	3332	Y:10	0.774	0.254	0.997	0.929	0.962	0.963	2.300	6.800	5	27	0.156
70	Set 6	3635	Y:05	0.786	0.493	0.987	0.905	0.945	0.946	8.200	8.400	5	27	0.156
71	Set 6	6058	Y:01	0.833	0.720	0.925	0.823	0.839	0.843	36.000	14.000	10	22	0.313
72	Set 6	3155	N	0.659	0.174	0.998	0.965	0.982	0.982	0.700	3.300	1	31	0.031

Table A.3: XGBoost results for a 5cv classification model.

Bibliography

[Atzmueller and Puppe, 2006] Atzmueller, M. and Puppe, F. (2006). Sd-map–a fast algorithm for exhaustive subgroup discovery. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 6–17. Springer.

[Ayeh et al., 2013] Ayeh, J. K., Au, N., and Law, R. (2013). "do we believe in tripadvisor?" examining credibility perceptions and online travelers' attitude toward using user-generated content. *Journal of Travel Research*, page 0047287512475217.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*.

[Dave et al., 2003] Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

[de Andalucía, 2015] de Andalucía, J. (2015). *Presentación datos estadísticos de la actividad cultura, educativa y turística 2015 conjunto monumental de la Alhambra y el Generalife.* Consejería de Cultura. Patronato de la Alhambra y el Generalife.

[DiGrazia et al., 2013] DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449.

[Ekman et al., 2013] Ekman, P., Friesen, W. V., and Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier.

[Elango and Narayanan,] Elango, V. and Narayanan, G. Sentiment analysis for hotel reviews.

[He et al., 2013] He, W., Zha, S., and Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472.

[Herrera et al., 2011] Herrera, F., Carmona, C. J., González, P., and Del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525.

[Hotho et al., 2005] Hotho, A., Nurnberger, A., and Paass, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62.

[Hu et al., 2012] Hu, N., Bose, I., Koh, N. S., and Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3):674–684.

[Jeacle and Carter, 2011] Jeacle, I. and Carter, C. (2011). In tripadvisor we trust: Rankings, calculative regimes and abstract systems. *Accounting, Organizations and Society*, 36(4):293–309.

[Jo and Oh, 2011] Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

[Kasper and Vela, 2011] Kasper, W. and Vela, M. (2011). Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference*, volume 231527, pages 45–52.

[Klösgen, 1996] Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, pages 249–271. American Association for Artificial Intelligence.

[Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Manning et al., 2014] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

[Marrese-Taylor et al., 2013] Marrese-Taylor, E., Velásquez, J. D., Bravo-Marquez, F., and Matsuo, Y. (2013). Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science*, 22:182–191.

[MartíN-Valdivia et al., 2013] MartíN-Valdivia, M.-T., MartíNez-CáMara, E., Perea-Ortega, J.-M., and UreñA-LóPez, L. A. (2013). Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942.

[Medhat et al., 2014] Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

[Mostafa, 2013] Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

[O'Connor et al., 2010] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[Reyes et al., 2013] Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

[Socher et al., 2013] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

[Tan et al., 1999] Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70.

[Turney, 2002] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

[Wrobel, 1997] Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer.

[WTTC, 2016] WTTC (2016). *Travel & Tourism Economic Impact 2016, Spain.* World Travel & Tourism Council.

[Yi et al., 2003] Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE.