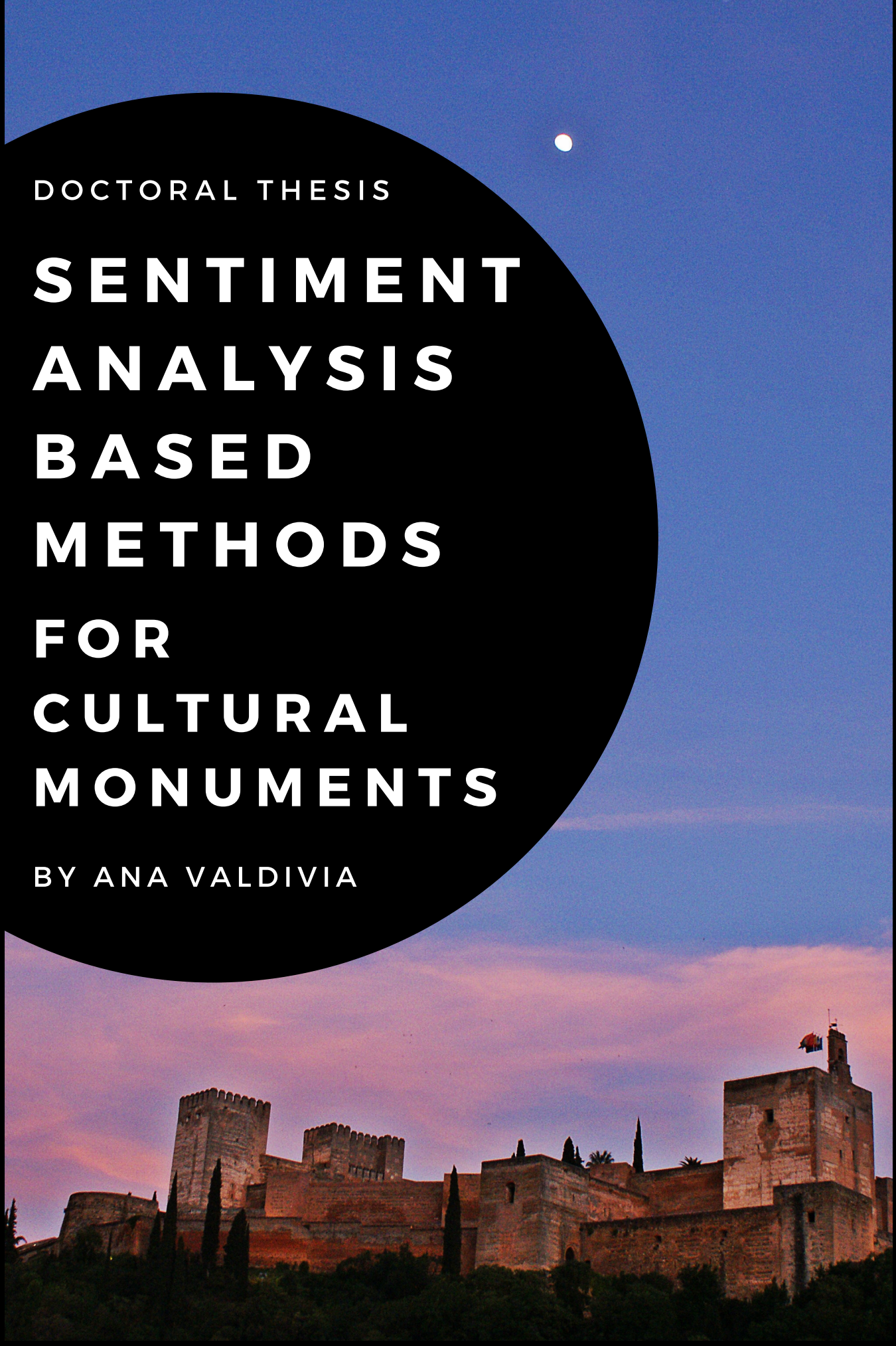
DOCTORAL THESIS

# SENTIMENT ANALYSIS BASED METHODS FOR CULTURAL MONUMENTS

BY ANA VALDIVIA



UNIVERSIDAD DE GRANADA

Editor: Universidad de Granada. Tesis Doctorales
Autor: Ana Valdivia García
ISBN: 978-84-1306-104-7
URI: http://hdl.handle.net/10481/54784

Universidad de Granada


UNIVERSIDAD
DE GRANADA

# SENTIMENT ANALYSIS-BASED METHODS FOR CULTURAL MONUMENTS

**DOCTORAL DISSERTATION**
*presented to obtain the*
**DOCTOR OF PHILOSOPHY DEGREE**
*in*
**COMPUTER SCIENCE**
*by*

**Ana Valdivia**

Granada, December 2018


UNIVERSIDAD
DE GRANADA

# SENTIMENT ANALYSIS-BASED METHODS FOR CULTURAL MONUMENTS

DOCTORAL DISSERTATION

*presented to obtain the*

DOCTOR OF PHILOSOPHY DEGREE

*in*

COMPUTER SCIENCE

*by*

**Ana Valdivia**

PhD Advisors

**María Victoria Luzón y Francisco Herrera**

COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE DEPARTMENT

This PhD dissertation titled "*Sentiment Analysis-Based Methods for Cultural Monument Reviews*", which is presented by Ana Valdivia for obtaining the PhD degree, has been carried out within the Official PhD Program "*Information and Communication Technologies*" of the Computer Sciences and Artificial Intelligence Department of the University of Granada, and under the guidance of María Victoria Luzón and Francisco Herrera.

The PhD student, Ana Valdivia, and her PhD advisors María Victoria Luzón and Francisco Herrera guarantee, by signing this doctoral thesis, that the work has been carried out by the PhD student under the direction of PhD advisors, and as far as our knowledge is concerned, the rights of authorship have been respected.

Granada, December 2018.

| The PhD student: | The PhD advisor: | The PhD advisor: |
| --- | --- | --- |
| Sgd.: Ana Valdivia | Sgd.: María Victoria Luzón | Sgd.: Francisco Herrera |

This doctoral thesis has been supported by the Spanish National Research Project TIN-2014-57251-P.

*To my mother,*
*to whom a backward society did not allow her to study.*
*I am sure that you would have been an excellent teacher.*

*To my godmother*
*who passed away while I was writing this thesis.*
*Thank you for being so strong.*

# Acknowledgements

*I would like to clearly express my gratitude to all women.*

*Many of the women in science have been made invisible by the history of humanity. I would like to dedicate my thesis to all of them, because without their struggle and perseverance, I probably would not have been here. Thanks to Ada Lovelace for showing your passion in the world of mathematics, thanks to that you created the first programming language. You have been considered the first person who programmed in history. Thanks to Hypatia for not overshadowing yourself, but to surpass many philosophers and instructors of your time. You were murdered by a society that did not understand your ideas, like so many others. You have been considered the first female mathematician. Thanks to Mileva Maric for showing your tenacity in a scientific world governed only by men. Your life has remained in the shade for so many years, and there is still controversy about your contribution to the field of relativity. However, no one can doubt your role as a scientific woman in a dark age to be so.*

*I would not like to stop thanking all those women who, from my point of view, have played a much more relevant role than science do, which is caregiving. To all those mothers who have left their dreams behind, however small, to raise and take care of their children and partners. It is the most important task to humanity and it is so invisible. Thanks to all the women of great scientists, mothers and wives, for taking care of them. Thanks to all brave women who died giving birth. Thanks to all the women who have been behind workers, laborers, etc. Thanks to all farmer women who worked in the fields. Thanks to all women who decided not to be mothers, moving against social roles. Thanks to my* yaya *Tanta, my grandmother Pepa, and my* tita *from Almería. You have been a clear example of carers. Thanks to my mom, for taking me to the doctor whenever I got sick, for encouraging me in every basketball game no matter how cold it was. You have always been there. Thanks to Musk, my other family: "We are a protest action. We are cold hands near the fire. We are granddaughters of fear and mourning. Disguised ourselves as utopia. We will take away from pain the meaning for life, riding the horse of reasoning. (Roba Estesa, Cant de Lluita, 2018)."*

*Thanks to Granada, your light will never cease to intoxicate my senses.*

*Finally, I would like to thank my PhD advisors for believing in me when proposing this thesis. Thank you for offering me your assistance and providing me with financial resources over these years.*

# Table of Contents

# Chapter I

# PhD dissertation

## 1 Introduction

A *sentiment* is an expression that is given in a conscious way due to the emotion that experiences cause. In other words, opinions are thought subjected to emotions. These opinions build our beliefs and realities. Furthermore, the decision that we take every day are previously influenced by our or others views. That is why, from the rationality that characterizes the human species, we have always sought to theorize on the basis of opinions and their emotions from different points of view, such as philosophy, anthropology, psychology or neuroscience.

One of the most relevant aspects of opinions is the way we convey them. The main channel that the human species has to show a state of mind is the facial expression. There are a large number of studies that show that combinations of facial muscles can unequivocally encode emotions [EF03, ST11]. Thus, human language is the tool used to express sentiments [LG13]. Undoubtedly, language is one of the main features that differentiate human beings from other species. Language is the concept that takes on a relevant role in the world of opinions, as it helps to communicate both orally and written sensations, thoughts or experiences, which is considered a very first need.

This basic need for communication is also embodied today. During the last decades, society has experienced a revolution in communication and information technologies, which has hit all areas of our daily lives. The Internet is a factor that has significantly influenced this process, revolutionizing the way we communicate, as did the telegraph, the radio, the telephone and television in the past. It is a global network consisting of large systems and small devices interconnected by a broad telecommunications infrastructure.

One of the ways to transfer information in this network is the web, which has also undergone a transformation process in recent years. Web 1.0 is the first format that was developed. The most relevant aspect of Web 1.0 is that the user behaves passively, receiving the information without no way of interaction with the content of such web. Afterwards, this type of web evolved into a fully dynamic and interactive website, the Web 2.0, in which people move from being passive subjects to become content generators. In this case, texts, images, and sounds appearing on these platforms have been created by the users themselves. Blogs, wikis, social media or other sharing-resources platforms are clear examples of this type of web. Due to our social need for communication, Web 2.0 has become another channel through which we express and interact, sharing opinions, thoughts, photographs, audios and videos. It has become a channel where we convey and interact, post and share opinions, thoughts, photo, audios, and videos, which evidences our need for communication.

Meanwhile, these platforms have also become huge warehouses of personal information, related in many cases with sentiments and emotions.

The content that is generated in this type of webs is very relevant, since it may affect the viability of a business or promote a great citizen mobilization, which has a great impact on society at different levels (political, economical, etc.) [AHYL18, BBL07]. Today's opinions are expose to thousands and thousands of people, and the speed with which they spread is greater than in old times. That is why both public and private organizations are interested in the development of techniques that allow analyzing the content of Web 2.0, in order to study its reputation or know the opinion that people hold towards a political party, a restaurant or a public transport company.

For more than two decades, the study of techniques capable of detecting the polarity that underlies a written text has evolved continuously, giving birth to a new branch of knowledge known as *Sentiment Analysis* (SA). The main aim of SA is to define computational tools capable of extracting subjective information from texts to create structured and actionable knowledge [PFML16]. SA has shown to have too many applications and it has attracted a great interest in the academic world. There exist a wide number of methods that have been developed to extract polarities from text [RAG$^+$16]. Pang et al. published the first study in which the polarity of reviews on films is classified [PLV02]. Taboada et al. developed a method for extracting polarities from different domains like book, music and hotel reviews [TBT$^+$11]. There are also other studies in which social media sentiments are analyzed [RFN17, WBCB17, BKW18]. These methods are also known as *Sentiment Analysis Methods* (SAMs).

TripAdvisor is a travel Web 2.0, which shows opinions of travelers towards hotels, restaurants, and tourist attractions about thousands of cities around the world. This website has become the largest online community of travelers, reaching 456 million unique visits per month[1], 661 million reviews on more than 7.7 million accommodations, restaurants or tourist attractions. TripAdvisor has also been used to analyze sentiments about restaurants or hotelss [KV11, MTVBM14]. However, to the best of our knowledge, there is no work that focuses on opinions about cultural heritage.

That is why in this thesis we have mainly focused the basis of our studies on reviews of cultural monuments. It is very important for these sites to analyze people's opinions in order to help the management operations and then improve the quality of future visitants. These heritage organizations are used to classic methods to doing so. Surveys are the most popular methodology, but its reliability depends on many factors such as the respondent's willingness to answer the truth or error in conclusions due to questions that are not asked in the interview. In this way, the instrumental hypothesis that we propose to solve is whether social media reviews, such as TripAdvisor, together with SA techniques can be proposed as alternatives to traditional methods for mining opinions of cultural visits.

The main disadvantages of SAMs are that they are strictly dependent on the corpus with which they are trained (domain adaptation problem) and they are very expensive to build. That is, if a SAM is trained with movie reviews and the same SAM is evaluated in restaurant sentiments, the method will be less precise, since the vocabulary, the way of expressing oneself and the context used in both domains can vary the meaning. Moreover, one of the requirements to develop a SAM which classifies polarities is to create an specific corpus for the problem that wants to be addressed. This procedure has a high cost, since it is necessary to label each document, sentences or phrase one by one. And also, not only by an expert, but by several, so that there is a consensus in the labeling process.

In this thesis we propose three different challenges, always based on cultural heritage reviews:

[1]Source: `tripadvisor.mediaroom.com/es-about-us`.

- The first goal is based on studying the viability of TripAdvisor as a source for cultural sentiment studies. We detect the need of a data source for developing our studies, therefore we propose TripAdvisor as a truthful platform for getting cultural heritage opinions. In this scenario, we also aim at addressing the inconsistency's problem detected. Due to the high cost of developing a SAM, we proposed to apply several off-the-shelf SAMs in TripAdvisor reviews of different monuments. We find out that the correlation between SAMs and users polarities is very low. One of the causes that have been already mentioned of these inconsistencies is the domain adaptation problem. In addition, we detect another possible reason and it is related with the variability of polarities in a same document. When someone writes a review and evaluates it with a scale of 1-5 (as in TripAdvisor), it does not express the same polarity for all the sentences of the same document. For example, if the experience is valued with a 5 (Excellent), we can find sentences with a negative connotation, and vice versa. Therefore, we propose the creation of an index that ensembles information from both sides, in order to have a better intuition of reviews and thus extract more reliable

- The (2) second specific objective is to improve the performance of polarity classification methods. For this purpose, we propose focusing on the detection of neutrality and treat it as the concept of noise in classical classification. In this way, we will obtain more precise methods when detecting positive and negative polarities.

- Finally, the (3) third objective is related to the opinion summarization. We believe that it is of utmost importance to create techniques that are capable of summarizing automatically the substantial content of opinions. These methods can help the decision-making processes of cultural organizations, detecting clearly those aspects that people like and those that need to be improved. conclusions.

Finally, this thesis consists of two different parts: the PhD dissertation and the publications. In the first part, in Section 1 provides the general context of this project. Section 2 describes the main concepts that are used in this thesis: Sentiment Analysis (Section 2.1), Machine Learning (Section 2.2), Deep Learning (Section 2.3), Fuzzy Set Theory (Section 2.4), and TripAdvisor (Section 2.5). The objectives and the methodology used to develop our ideas are described in Section 3 and Section 4, respectively. In Section 5 is proposed an introduction to publications. In Section 6 the main results of these publications are presented. Finally, in Section 7 provides the overall conclusions and open future lines derived from this thesis.

The second part of the document consists of the four international journals publications that have made this thesis possible, organized according to the proposed objectives explained before:

- Sentiment Analysis in Tripadvisor.

- The Inconsistencies on TripAdvisor Reviews: a Unified Index between Users and Sentiment Analysis Methods.

- Consensus Vote Models for Detecting and Filtering Neutrality in Sentiment Analysis.

- What do People Think about this Monument? Understanding Negative Reviews via Deep Learning and Descriptive Rules.

## Introducción

Una *opinión* es una expresión que se da de manera consciente debido a la emoción que provoca una experiencia. Es decir, que las opiniones son pensamientos sometidos a emociones. Estas opiniones son las que a su vez construyen nuestras creencias y realidades. Además, las decisiones que tomamos día a día son previamente influenciadas por opiniones, ya sean tanto propias como ajenas. Es por ello que, desde la racionalidad que nos caracteriza como especie, siempre se ha buscado teorizar sobre la base de las opiniones y sus emociones desde diferentes puntos de vista, como la filosofía, la antropología, la psicología o la neurociencia.

Uno de los aspectos más relevantes de las opiniones es la manera que tenemos de expresarlas. El principal canal que tiene la especie humana para mostrar un estado de ánimo es la expresión facial. Existen numerosos estudios que ponen de manifiesto que ciertas combinaciones de los músculos faciales codifican inequívocamente las emociones [EF03, ST11]. De tal manera, el lenguaje humano es la herramienta utilizada para expresar las opiniones [LG13]. Es indudable afirmar que el lenguaje constituye una de las manifestaciones características que diferencian al ser humano. Es el lenguaje el concepto que toma una papel relevante en el mundo de las opiniones, pues ayuda a comunicar tanto de manera oral como escrita sensaciones, pensamientos o experiencias, lo cuál se considera una necesidad primaria de nuestra especie.

Esta necesidad básica de comunicación también se manifiesta hoy en día. Durante las últimas décadas, la sociedad ha experimentado una revolución de las tecnologías de la comunicación y la información, la cuál ha afectado a todos los ámbitos de nuestra vida. Internet es un factor que ha influenciado notablemente en este proceso, revolucionando la manera que tenemos de comunicarnos, así como también lo hicieron en el pasado el telégrafo, la radio, el teléfono o la televisión. Internet es una red mundial formada por grandes sistemas y pequeños dipositivos interconectados mediante una amplia infraestructura de telecomunicaciones.

Una de las formas que existe para transferir información en esta red es la web, la cuál también ha experimentado un proceso de transformación en los últimos años. La Web 1.0 es el primer formato que se desarrolló. En ella, la persona usuaria se comporta de manera pasiva, recibiendo la información sin que exista posibilidad alguna de interacción con el contenido. Posteriormente, este tipo de web evolucionó a una versión totalmente dinámica e interactiva, la Web 2.0, en la cuál las personas pasan de ser sujetos pasivos para convertirse también en generadoras de contenido. En este caso, los textos, imágenes, sonidos que aparecen en la plataforma han sido creados por la misma persona usuaria del sitio web. Blogs, wikis, redes sociales o entornos para compartir recursos son claros ejemplos de este tipo de web. Debido a nuestra necesidad social de comunicación, la Web 2.0 se ha convertido en otro canal más por el que nos expresamos y relacionamos, compartiendo tanto opiniones, como pensamientos, fotografías, audios y vídeos. Es así como estas plataformas web también se han transformado en almacenes de una gran cantidad de datos, muchos de ellos relacionados con sentimientos y estados de ánimo.

El contenido que se genera en este tipo de webs es de total relevancia, pues puede llegar a condicionar la viabilidad de un negocio o promover una gran mobilización ciudadana, lo cuál tiene un gran impacto a nivel social y económico como recientemente se ha demostrado [AHYL18, BBL07]. Las opiniones que anteriormente se transmitían de *boca en boca*, hoy están expuestas a millones de personas en diferentes plataformas webs. Es por ello que tanto organizaciones públicas como privadas están interesadas en el desarrollo de técnicas que permitan analizar el contenido de Webs 2.0, para así estudiar su reputación o conocer la opinión que se tiene respecto a un producto, un

partido político, un restaurante o una empresa de transporte público.

Desde hace más de dos décadas, el estudio de técnicas capaces de detectar la polaridad que subyace de un texto escrito ha evolucionado continuamente, creando una nueva rama del conocimiento nombrada *Análisis de Opiniones* (AO). El objetivo del AO es el de definir herramientas automáticas capaces de extraer información subjetiva a partir de textos para crear conocimiento estructurado y procesable [PFML16]. Debido a las muchas aplicaciones que esta rama ha demostrado tener y el gran interés que ha despertado en el mundo académico, son muchos los métodos que se han desarrollado para, por ejemplo, extraer la polaridad de un texto [RAG+16]. Pang et. al publicaron el primer estudio en el que se clasifica la polaridad de reseñas sobre películas [PLV02] . Taboada et al. desarrollaron otro método para extraer polaridades en el que utilizaban reseñas de diferentes fuentes como: libros, automóviles, música, hoteles, etc [TBT+11]. También existen otros estudios en el que se analizan los sentimientos de textos en redes sociales [RFN17, WBCB17, BKW18]. Este tipo de métodos son conocidos como *Métodos de Análisis de Opiniones* (MAO).

TripAdvisor es una Web 2.0 de viajes, la cuál muestra opiniones de viajantes sobre hoteles, restaurantes o puntos turísticos de miles de ciudades alrededor del mundo. En términos generales, esta web se ha convertido en la mayor comunidad online de viajantes, llegando a albergar 456 millones de visitas únicas al mes[2], 661 millones de reseñas sobre más de 7,7 millones de alojamientos, restaurantes o atracciones turísticas. TripAdvisor también ha sido utilizado para analizar el sentimiento que se expresa sobre restaurantes u hoteles [KV11, MTVBM14]. No obstante, en la literatura no existe ningún artículo que se centre en las opiniones respecto a puntos de interés patrimonial.

En esta tesis hemos centrado la base de nuestros estudios en reseñas de monumentos culturales. Para este tipo de instituciones es de total relevancia conocer la opinión del visitante para mejorar la calidad de su experiencia. Principalmente, estas organizaciones utilizan métodos clásicos para conocer la valoración de sus visitantes. La encuesta es la metodología más popularizada, pero su fiabilidad depende de muchos factores como: la disposición del encuestado/a por responder la verdad, error en las conclusiones debido a preguntas que no se plantean en el cuestionario, etc. De esta manera, la pregunta fundamental que nos proponemos resolver es si las opiniones almacenadas en redes sociales, como TripAdvisor, junto con técnicas de AO pueden ser propuestas como alternativas a estos métodos tradicionales, y así conocer tanto las preferencias como las insatisfacciones de las visitas culturales.

No obstante, las principales desventajas de los MAO es que dependen estrictamente del conjunto de datos con el que se entrene (*problema de adaptación al dominio)* y son muy costosos de construir. Si se desarrolla un MAO entrenado con opiniones de películas y se evalúa ese mismo método en opiniones referentes a un restaurante, el método será menos preciso, pues el vocabulario, la forma de expresarse y el contexto que se utilizan en ambos temas pueden cambiar de significado. Además, uno de los requisitos para desarrollar un método que clasifique la polaridad de un conjunto de opiniones es el de crear un corpus específico para el problema que se quiera abordar. Este procedimiento tiene un elevado coste, pues deben etiquetarse uno a uno los sentimientos de cada comentario; y además, no solo por un experto, sino por varios, para que exista consenso en el etiquetado y los resultados sean robustos.

En esta tesis proponemos tres retos diferentes, siempre basados en opiniones de monumentos culturales:

- El (1) primer objetivo se basa en estudiar la viabilidad de TripAdvisor como una fuente de opiniones para monumentos culturales. La necesidad de tener una fuente de datos de confianza para desarrollar nuestros estudios es crucial y de total relevancia. En este escenario, también

[2]Fuente: `tripadvisor.mediaroom.com/es-about-us`.

proponemos la necesidad de remediar el problema de las inconsistencias. Debido al alto coste que tiene desarrollar un MAO, propusimos aplicar diferentes ya desarrollados en nuestro corpus de monumentos. Descubrimos que las inconsistencias entre la polaridad extraida por las MAO y la valoración del usuario eran muy altas. Una de las causas ya comentadas de estas inconsistencias es el problema de adaptación al dominio que muestran la mayoría de MAO. No obstante, detectamos otra causa y es la variabilidad de polaridades en un documento, es decir, cuando alguien escribe una reseña y la evalúa con una escala del 1-5 (como en TripAdvisor), no utiliza el mismo sentimiento para todas las frases del mismo documento. Por ejemplo, si la experiencia está valorada con un 5 (Excelente), podemos encontrar frases con connotación negativa, y viceversa. Por ello, proponemos la creación de un índice que recoja información de ambas partes, para así tener una idea más ajustada del contenido del texto.

- El (2) segundo objetivo específico es el de mejorar el rendimiento de los métodos de clasificación de polaridades. Para ello, proponemos centrarnos en la detección de la neutralidad y tratarla como el concepto de ruido en clasificación clásica. De esta manera, obtendremos métodos más precisos a la hora de detectar polaridades positivas y negativas.

- Por último, el (3) tercer objetivo está relacionado con la sintetización de opiniones. Además de mejorar los resultados de métodos de clasificación de polaridades, y extraer información detallada de estas, creemos que es de total relevancia crear métodos que sean capaces de resumir automáticamente el contenido substancial de las opiniones. Estos métodos pueden ayudar a los procesos de toma de decisiones de estas organizaciones, detectando de una manera clara aquellos aspectos a mejorar.

Finalmente, esta tesis está formada de dos partes diferenciadas: la tesis doctoral y las publicaciones. En la primera parte, en la Sección 1 describimos el contexto general de este proyecto. En la Sección 2 describimos los principales conceptos que soportan esta tesis: Análisis de Opiniones (Sección 2.1), Aprendizaje Automático (Sección 2.2), Aprendizaje Profundo (Sección 2.3), Teoría de Conjuntos Difusos (Sección 2.4) y TripAdvisor (Sección 2.5). Los objetivos y la metodología empleada para el desarrollo de nuestras ideas se describen en la Sección 3 y en la Sección 4, respectivamente. En la Sección 5 proponemos una introducción a las publicaciones relacionadas con esta tesis. Por otro lado, en la Sección 6 explicamos los principales resultados de estas publicaciones. Finalmente, en la Sección 7 describimos las principales conclusiones extraídas, así como las futuras línias abiertas que deja este extenso trabajo.

La segunda parte del documento consta de las cuatro publicaciones en revistas indexadas e internacionales, organizadas según los objetivos propuestos:

- Sentiment Analysis in Tripadvisor.

- The Inconsistencies on TripAdvisor Reviews: a Unified Index between Users and Sentiment Analysis Methods.

- Consensus Vote Models for Detecting and Filtering Neutrality in Sentiment Analysis.

- What do People Think about this Monument? Understanding Negative Reviews via Deep Learning and Descriptive Rules.


Figure 1: Evolution of interest of *Sentiment Analysis* in Google.

# 2 Preliminaries

This section presents the main concepts of the topics that have driven this thesis. First, Section 2.1 gives an overview of the Sentiment Analysis, explaining the problem and the different challenges which are part of it. Section 2.2 describes Machine Learning approaches for Sentiment Analysis tasks, as well as in in Section 2.3 for Deep Learning. Section 2.4 introduces some concepts from fuzzy theory that can be applied to face Sentiment Analysis open problems. Finally, Section 2.5 details the most relevant features of the TripAdvisor webpage that makes it a proper source for Sentiment Analysis research.

## 2.1 The Sentiment Analysis

Sentiment Analysis (SA) is a research line that has experienced an important growth over the last two decades (see Figure 1), which is due in part to the content generation effect of Web 2.0 and partly due to the new challenges that this SA has opened. This field is considered a subtask of Natural Language Processing (NLP) [3] that tries to help computers to understand and interpret human language. The main goal of SA is to extract computationally meaningful knowledge from opinions. It was defined by Liu as [Liu15]:

*"It is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text."*

However, we find that this definition fits more to the current state of SA [CH12]:

*"The set of computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various on-line news sources, social media comments, and other user-generated contents."*

The object of the study itself is an opinion. As it was defined in [QGLS85], an opinion is the evidence or expression of a private state. In general, there exist two main attributes 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 comes from the opinion.

[3] `en.wikipedia.org/wiki/Natural_language_processing`.

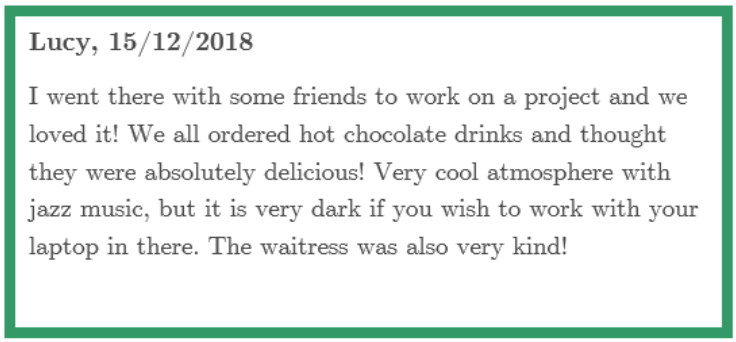



Figure 2: Example of an opinion about a coffee shop.

However, to analyze opinions is also relevant to know about the opinion holder, the date when the opinion was expressed, etc. Specifically, Liu proposes in [Liu12] a formal definition of an opinion widely recognized by the community:

**Definition 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 entity or opinion target,*
- $a_{ij}$ *is the j-th attribute of* $e_i$*, i.e., a characteristic of the entity that is expressed in the opinion,*
- $s_{ijkl}$ *is the sentiment or polarity of the opinion towards the aspect* $a_ij$ *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 the l-th time when the opinion was emitted.*

As an example, the Figure 2 shows an example of a review about a coffee shop (*entity*). As we can read, the *opinion holder*, Lucy, reports thoughts about several aspects of the coffee shop on the 15th of December, 2018 (*time*). She talks about the place, the drinks and the service (*aspects*). This opinion can be structured by Liu's definition as shown in Table I.1.

### 2.1.1 The Sentiment Analysis Problem: Tasks and Levels

The number of SA tasks and applications is growing considerably because of the variety of topics people write about on the Internet (politics, science, tourism, health care, finances, etc.). This information can be process and improve the decision making process of organizations or government agencies. Also, these applications are evolving daily becoming more challenging, due to the richness of the human language, i.e., people shorten words, use irony and sarcasm, or make use of emojis[4] to

[4] `en.wikipedia.org/wiki/Emoji`.

| 5-Tuple | Entity | Aspect | Sentiment | Opinion holder | Time |
|---|---|---|---|---|---|
| $(e_1, a_{11}, s_{1111}, h_1, t_1)$ | $e_1$ = coffee shop | $a_{11}$ = drinks | $s_{1211}$ = positive | $h_1$ = Lucy | $t_1$ = 15/12/2018 |
| $(e_1, a_{12}, s_{1211}, h_1, t_1)$ | $e_1$ = coffee shop | $a_{12}$ = drinks | $s_{1211}$ = positive | $h_1$ = Lucy | $t_1$ = 15/12/2018 |
| $(e_1, a_{13}, s_{1211}, h_1, t_1)$ | $e_1$ = coffee shop | $a_{13}$ = place | $s_{1311}$ = brightness | $h_1$ = Lucy | $t_1$ = 15/12/2018 |
| $(e_1, a_{14}, s_{1411}, h_1, t_1)$ | $e_1$ = coffee shop | $a_{14}$ = service | $s_{1411}$ = positive | $h_1$ = Lucy | $t_1$ = 15/12/2018 |

Table I.1: Example of structured opinion about review in Figure 2.

express feelings and sentiments. However, there exists a taxonomy of the main tasks related with SA that we can find in [PFML16] (see Figure 3):

- **Subjectivity classification:** This task aims at classifying opinionated sentences, i.e., sentences that express a positive or negative polarity.
- **Polarity classification:** This is considered the most popular task of SA. It is focused on detecting the polarity of a text. The text can be a whole document, a sentence or phrase or an aspect of the review. There exist different tags for polarities: binary label (positive, negative), trio label (positive, neutral, negative), numeric scale ($\{1, 2, 3, 4, 5\}$), text scale (very negative, negative, neutral, positive, very positive), etc. The methods based on this task are named *Sentiment Analysis Methods* (SAMs).
- **Opinion summarization:** It is the task of automatically summarize a large quantity of reviews. The result of the summary can be shown in different formats like visualizations, texts, etc.
- **Opinion visualization:** It is the task of reflecting the content of a text as a graph or plot. In this works, it is very common to build dashboards that can assist humans in a decision making process.
- **Sarcasm detection:** This task aims at detecting irony in texts. Humans usually detect irony or sarcasm by intonation of the voice or face's signals, which makes the tasks of detecting irony in plain text very challenging.
- **Entity, opinion holder and time extraction:** It aims at extracting detailed information from the review. If we analyze opinions, it is important to know who wrote the opinion, when, and what he/she was talking about.
- **Coreference resolution and word sense disambiguation:** When we talk about an idea or concept, we can express it in multiple ways. In many languages, a same word can have different meanings, which are defined as polysemic words. Therefore, it is important to automatically classify the different meanings of words in a text if we want to obtain consistent conclusions.

Once we have defined the main tasks of SA, it is important to describe the levels to which an analysis of sentiments can be developed. To do so, we focus on the most popular task, the Polarity

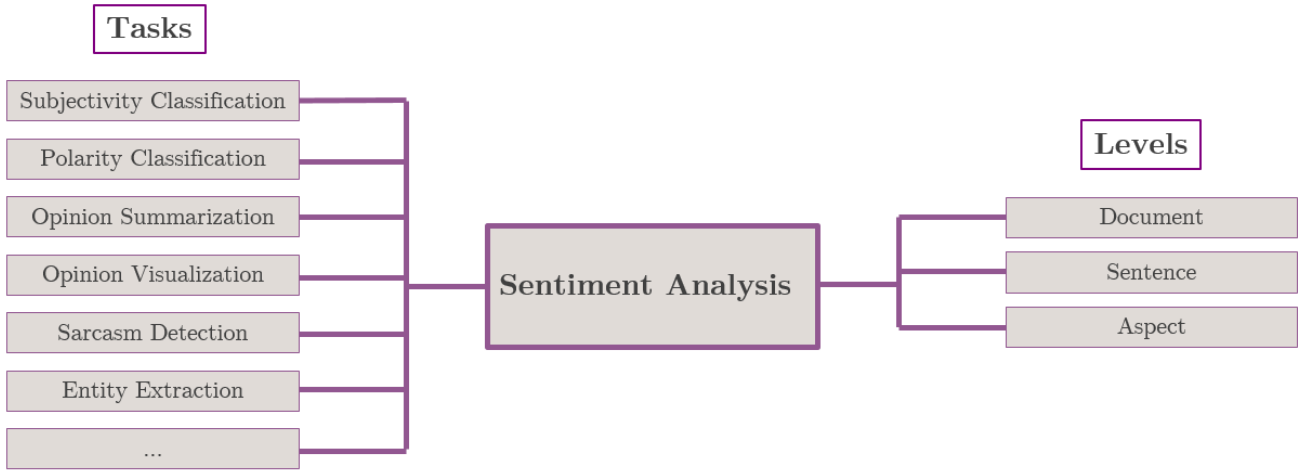



Figure 3: Sentiment Analysis tasks and levels.

Classification. As we observe in Table I.1, the polarity or sentiment of the opinion corresponds to the aspect. However, we can also extract the sentiment of the whole text, a sentence, etc. Because of that, in the SA literature considers three analysis levels:

- **Document-Based Sentiment Analysis (DBSA):** This level aims at classifying the overall sentiment of the text. It does not analyze other properties of the opinion like aspects or entities. Therefore, it assumes that opinions only refer to an entity and it is expressed by a single opinion holder. It is considered the most studied task as well as the easiest one. For instance, in Figure 2 we can assume that the whole sentiment is positive.

- **Sentence-Based Sentiment Analysis (SBSA):** When we convey an opinion, we usually do not use the same sentiment for all the sentences. Although the main sentiment of the document can be positive, we can express negativity in some sentences, and vice versa. Moreover, we can observe neutral sentences, i.e., sentences that do not express any sentiment. It is in this task where neutral polarities are also considered.

- **Aspect-Based Sentiment Analysis (ABSA):** When we review a product, we can express a sentiment towards several aspects of it. For example, if we review a coffee shop as in Figure 2, we can talk about the drinks, the service, the cleaning, the bathrooms, etc. Therefore, this level has two main approaches: *Aspect Extraction* and *Aspect Sentiment Classification*. Aspect Extraction is to automatically identify aspects and entities that have been evaluated in the opinion (see Table I.1). Aspect Sentiment Classification is based on the Polarity Classification task, it determines the polarity towards an aspect. The task of extracting fine-grain information has seized the attention of researchers due to the highly requirements needed for performing this task.

In order to prove the constant evolution of this field, there exist an ongoing of computational analysis tasks SemEval[5]. Every year this association proposes new challenges regarding several

[5] `en.wikipedia.org/wiki/SemEval`.

tasks of computational semantic analysis. Over the last years, they have proposed problems such as semantic role labeling, relations between sentences, multilingual textual similarity, etc. One of the most active areas of study of SemEval is SA, proposing evaluations based on polarity classification in Twitter, or aspect sentiment classification of hotel reviews.

## 2.2 Machine Learning for Sentiment Analysis

Machine Learning (ML) is defined as the field of artificial intelligence that uses statistical techniques to give computer systems the ability to learn from data. ML has solved many problems that have no known algorithm to solve it, by developing algorithms that learns by itself. Formally, Tom M. Mitchell defined ML algorithms as [M+97], a computer program based in ML is said to learn from *experience* $E$ with respect to some class of *task* $T$ and *performance* measure $P$, if its performance at tasks in $T$ as measured by $P$, improves with experience $E$.

There are mainly two basic categories of ML algorithms:

- **Supervised Learning:** The algorithm already know the correct output of the data. The goal is then to discover patterns in the inputs that characterize the desired outputs. The main examples are regression (continuous output) and classification (discrete output) algorithms.
- **Unsupervised Learning:** In this case, the algorithm has no idea about the output. So the aim is to find structured outputs from the inputs. A clear example of this type is clustering algorithms.

ML has proven to be very effective in a wide number of real problems. As it learns from large amounts of data, it can help to discover patterns that are invisible to humans. It also can help to show how it learns, in order to extract robust conclusions about data and its trends. Nowadays, these algorithms are applied for different purposes like: product recommendation, weather prediction, detection of diseases, etc. Moreover, they are also used in the field of NLP and SA, contributing to several tasks like spam detection. Among SA applications, the developing of SAMs for polarity classification is the most known task [RAG+16].

### 2.2.1 Sentiment Analysis Methods (SAMs)

ML has influenced the development of SA techniques, outperforming the state-of-art of many problems. Concretely, ML has been directly related with the polarity classification task, proposing learning methods that determines the sentiment of an opinion (SAMs) [MHK14]. The process of these algorithms is as follows:

1. **Data collection:** A huge amount of data is needed in order to build a method capable of extracting sentiments from text. Typically, this data is obtained from social networks, blogs or other platforms.
2. **Sentiment labeling:** The algorithm needs labeled data (*the class*) to learn the distribution of it and understand correlations between the input and the output. In this case, the class can be binary (positive, negative), or multiple (positive, neutral, negative, ...). This process can be done by the opinion holder. If this is not the case, it is necessary to run a crowdsourcing tagging to asses the polarity of opinions by several experts.

3. **Feature selection:** One of the main points of these processes is that data is unstructured. Therefore, we need to give a structure in order to apply a ML algorithm. Such structure is called *term-document matrix*, where the element $a_{ij}$ is 1 if the $i$th-feature appears in the $j$th-document, and 0 otherwise. Features are the most significant words from each polarity or class. Traditionally, it is selected the most relevant terms on the collection of documents, which are measured by metrics like the term frequency-inverse document frequency (*tfidf*). Terms and words are also transform using text mining techniques like *lemmatization*, *stemming*, *part-of-speech*, *tokenization*, etc.

4. **Classification problem:** In one hand, sentiment analysis can be studied in a machine learning approach. 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), association 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 identify the polarity of the text using rules or heuristics obtained from language knowledge.

It is also worth mentioning that SAMs can be build as a lexicon-based approach. In this case, these methods are developed from a term dictionaries, i.e., collection of words. There exist two main techniques: dictionary-based approach and corpus-based approach. The first one is about building a sentiment lexicon manually. This approach has the inconvenience that may not work on concrete opinions, where terms does not appear. However, corpus-based approaches solve this problem by growing the dictionary with terms that are specific from the text that we are evaluating.

The main disadvantage of SAMs is that they are **content dependent** and they strongly depend on the learning approach. These methods are built from scratch, using text from a specific domain such as movie reviews, restaurant opinions, product sentiments, etc. Therefore, if we evaluate one of these methods in a distinct domain where it has been trained, it will show a poor performance. The fact is that words can be used in different contexts indicating an opposite polarity. This lack of generalization has become one of the main concerns of the SA community, becoming an authentic challenge.

## 2.3 Deep Learning for Sentiment Analysis

Deep Learning (DL) is a sub-field of ML that aims at developing algorithms for extracting and transforming new features. These algorithms work in a layers's system, simulating the basic function of a human brain. Classic ML approaches work well in a wide range of AI problems, however they have not shown successful results in areas like computer vision or NLP. In contrast, DL architectures have proven to obtain impressive results, outperforming the state-of-art of many open problems. In recent years, Convolutional Neural Networks (CNNs), Recurrent Networks (RNNs) and Recursive Networks (RecNN) have been performing outstanding results on various NLP tasks like named-entity recognition, semantic role labeling, and POS tagging [YHPC18]. There architectures have also enhance the results of many SA challenges, mainly polarity classification (SAMs) and aspect extraction and sentiment classification (ABSA) [RB16, TZ18, ZWL18]

One of the keys to the success of these systems are *word embeddings*. Word embeddings are distributed representations of words which follows the hypothesis that words with similar meanings tend to occur in similar contexts [BDVJ03]. They are mainly used as an input for the neural networks. Mikolov et al. proposed in 2013 the `word2vec` model which is based on the

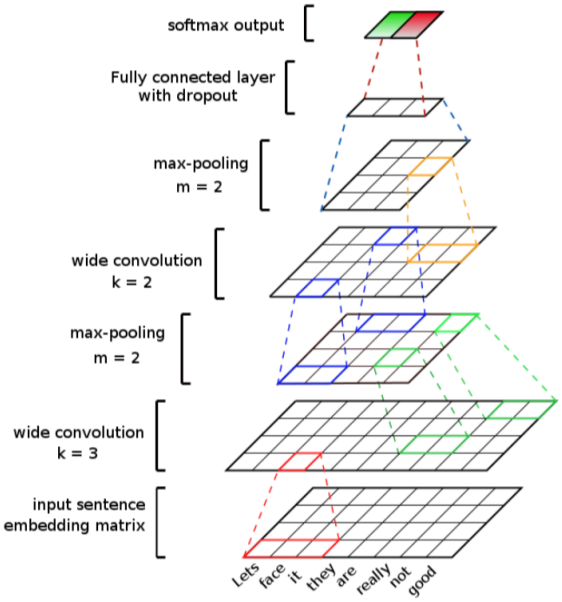



Figure 4: CNN proposed in [PCG16] for aspect extraction.

concatenation of continuous bag-of-words (CBOW) and skip-gram, two architecture networks that learns vector representations capturing a large number of precise syntactic and semantic word relationships [MSC+13]. Levy et al. proposed dependent-based word embeddings, generalizing the linearity of the skip-gram model [LG14].

DL have also prompted the development of more precise ABSA-based systems. Word embeddings have enhanced this task because they can represent the relation of words as vectors, which is extremely helpful for extracting aspects. Moreover, neural networks can focus on aspect-related words. For instance, Poria et al. developed a system based on a CNN (see Figure 4) and linguistic rules for extracting aspects [PCG16].

Although these results, DL has not exhibited a stunning behaviour in sentiment classification. As we observe in several polarity classification tasks of SemEval[6], teams used ML approaches like SVM or Conditional Random Fields (CRFs) rather than neural networks architectures. However, there exist different SAMs focus on document, sentence or aspect level. For instance, CoreNLP [MSB+14] is a SAM sentence-based level, based on a RNTN, which outperforms old methods in different metrics. This architecture classifies polarities representing sentences as *treebanks* which are representations of sentence's chunks where is labeled the polarity and the POS for each unit. At each leaf node of the treebank, the RNTN computes the overall polarity of the node of the treebank, taking into account the words that depend on the node and the outcome of the previous one iteration (bottom-up system). This system is able of capturing polarity changes on a same sentence, i.e., it effectively detects scope negations and contrasting conjunctions like *but* (see attached Figure 5).

As we have explained in Section 2.2, one of the main drawbacks of SAMs is the content dependent problem. If the training domain is different from the test context, the performance of the SAM sharply decreases. Unfortunately, this adaptation problem has not been yet solved with DL techniques. For instance, CoreNLP which is trained with movie reviews has a poor performance in other contexts like micro-blogging messages. Note that the way we review a movie or express our sentiments in Twitter towards a politician can be very different (use of emojis, short texts, ...).

[6]See SemEval-2014 Task 5.

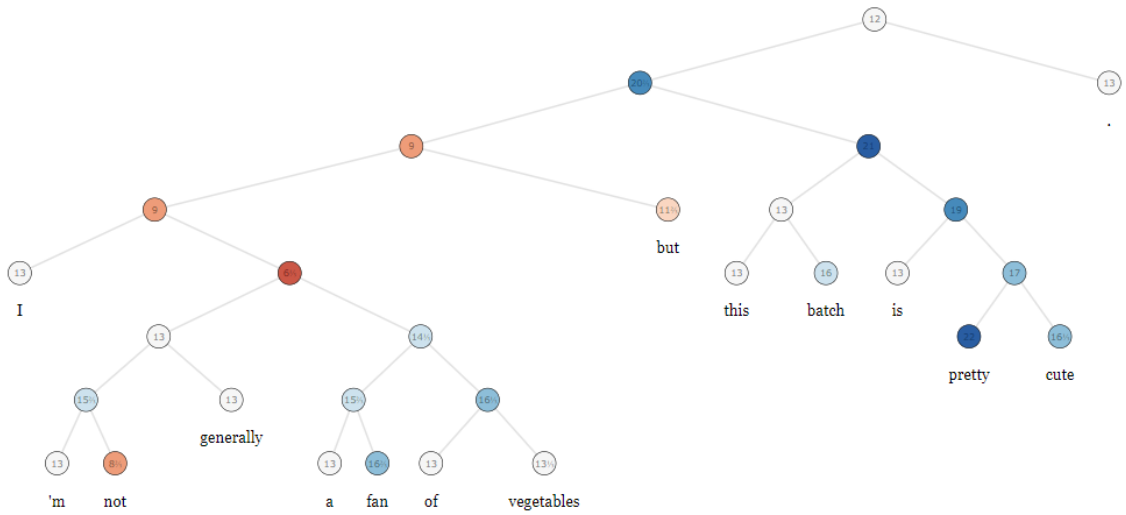



Figure 5: Example of sentence based on CoreNLP architecture [MSB+14]. For more examples visit: `nlp.stanford.edu/sentiment`.

## 2.4 Fuzzy Sets Theory for Sentiment Analysis

Decision making is the process of identifying the decision, gathering the required information, identifying the alternatives from several agents and choosing the best option. Using this process can increases the probability of selecting the most successful choice. Some examples are a group of doctors determining the best treatment to a patient, a group of investors deciding where to invest money, a popular jury deciding a final decision, etc. *Fuzzy set theory* has been widely applied for different decision making problems because in general, the opinion of the majority is the key for obtaining the most satisfying alternative.

As we explained in Section 2.2, over the last years a large numbers of SAMS have been developed to analyze sentiments from text with different contexts and domains. Thus, we propose in this thesis to ensemble several SAMs (agents) and enhance polarity-detection classification results through fuzzy set theory. The idea is to aggregate the output of several methods and compute the opinion of the majority by giving different weights to each SAM's polarity. The most used aggregation methods are the tendency values, mean or median, but in our approach we propose different mechanisms that enrich the process of decision making in SA by majority vote [ACCF17].

Yager defined a set of fuzzy operators based on the implementation of the concept of fuzzy majority, which is defined as Ordered Weighted Averaging (OWA) operators [Yag88]. These operators are used in several multi-options decision making processes, obtaining an accurate representation of the majority. More formally, these operators can be express as:

**Definition 2.** *OWA [Yag88, CHVHA07].* An *OWA operator* of dimension $n$ is a mapping $\phi : \mathbb{R}^n \to \mathbb{R}$ that has an associated weighting vector $W$ such that $w_i \in [0, 1]$, $\sum_{i=1}^{n} w_i = 1$, and is defined to aggregate a list of values $\{p_1, \ldots, p_n\}$ following this expression:

$$\phi(p_1, \ldots, p_n) = \sum_{i=1}^{n} w_i p_{\sigma(i)},$$

being $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ a permutation such that $p_{\sigma(i)} \geq p_{\sigma(i+1)}$, $\forall i = 1, \ldots, n-1$.

Lately, Yager et al. proposed the Induced Ordered Weighted Aggregation (IOWA) operators are a generalization of OWA operators [YF99]. They are defined as an aggregation operator which one component induces an ordering over the second component. More precisely:

**Definition 3.** *IOWA [YF99, CHVHA07].* An *IOWA operator* of dimension $n$ is a mapping $\Psi : (\mathbb{R} \times \mathbb{R})^n \to \mathbb{R}$ that has an associated weighting vector W such that $w_i \in [0, 1]$, $\sum_{i=1}^{n} w_i = 1$, and it is defined to aggregate the set of second arguments of a list of $n$ 2-tuples:

$$\Psi(\langle u_1, p_1\rangle, \ldots, \langle u_n, p_n\rangle) = \sum_{i=1}^{n} w_i p_{\sigma(i)},$$

being $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall i = 1, \ldots, n-1$.

The vector of values $U = (u_1, \ldots, u_n)$ is defined as the *order-inducing* vector and $(p_1, \ldots, p_n)$ as the *values of the argument variable*. In this way, the *order-inducing* reorders the *values of the argument variable* based on its magnitude.

There are mainly two ways to compute the weighting vector (W) of OWA and IOWA operators. The first approach is to use the inherent properties of data. The second one is to give weights a semantic meaning, like *most SAMs obtained a neutral polarity*, or *at least half SAMs agree on detecting neutrality*. Therefore, we may assign weights according to some importance's criteria, like in previous examples *detecting neutrality*. This approach is defined as *fuzzy linguistic quantifiers* [Zad83]:

**Definition 4.** *Fuzzy Linguistic Quantifier [Zad83].* A *fuzzy linguistic quantifier* $Q : [0, 1] \to [0, 1]$ is a monotonic and parametrized function where $Q(0) = 0$ and $Q(1) = 1$, and $Q(x) \geq Q(y)$ for $x > y$. It is used for modeling the concept of quantification to represent the fuzzy majority. *At least half*, *Most of* and *Many as possible* are some examples of these quantifiers (see Figure 6), which can be modeled explicitly as:

$$Q_{(a,b)}(x) = \begin{cases} 0 & \text{if } 0 \leq x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } b \leq x \leq 1 \end{cases}$$

The values that are used for the pair $(a, b)$ are [Kac86]:

$$Q_{At\ least\ half}(x) = Q_{(0,0.5)}(x)$$

$$Q_{Most\ of}(x) = Q_{(0.3,0.8)}(x)$$

$$Q_{Many\ as\ possible}(x) = Q_{(0.5,1)}(x)$$

For instance, *Most of* may be interpreted as follows: if at least 80% of some elements satisfy a property (such as obtaining a neutral polarity), then most of them certainly (to degree 1) satisfy it, when less than 30% of them satisfy it, then most of them certainly do not satisfy it (satisfy to degree 0), and between 30% and 80% the more of them satisfy it, the higher the degree of satisfaction by most of the elements [KFN92].

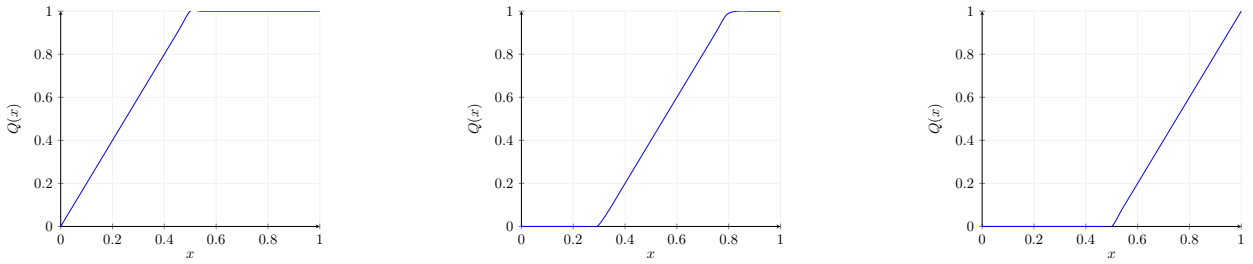



Figure 6: Linguistic Quantifiers Represented as Fuzzy Sets: *At least half*, *Most of* and *Many as possible*, respectively.

## 2.5 TripAdvisor

According to Wikipedia, TripAdvisor is an American travel website company providing reviews from travellers experiences about hotels, restaurants and monuments[7]. 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 turned to a user generated content.

The site describes itself as follows:

*"TripAdvisor is the world's largest travel site1 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."*

In general terms, this website has made up the largest travel community, reaching 409 million average unique monthly visitors, and 702 million reviews and opinions covering more than 8 million accommodations, restaurants and attractions over 48 markets worldwide[8]. It is considered a Web 2.0 for tourism domain and, more specifically, as a social network, virtual community and blog [O'C08]. The most interested feature of this platform is the user-generated factor. Hundreds of thousands of tourists post every day reviews about either restaurants, hotels or touristic attractions. Reviewers are also ask to evaluate their experience on a numeric scale (from one to five points), which is used for generating an overall rate of the place. These evaluation aspects determine a wide amount of people who make their decisions based on the sentiments and ratings published on TripAdvisor.

The description of the main features of TripAdvisor's reviews is as follows (see Figure 7):

(1) `bubble rating`: Evaluation of the user. Numeric scale from 1 to 5 (from Terrible to Excellent).

(2) `time`: Date when the review is published.

(3) `review`: Text of the review.

(4) `username`: Nickname of the user.

(5) `location`: Location of the user.

[7] `en.wikipedia.org/wiki/TripAdvisor`.
[8] `tripadvisor.mediaroom.com/uk-about-us`.

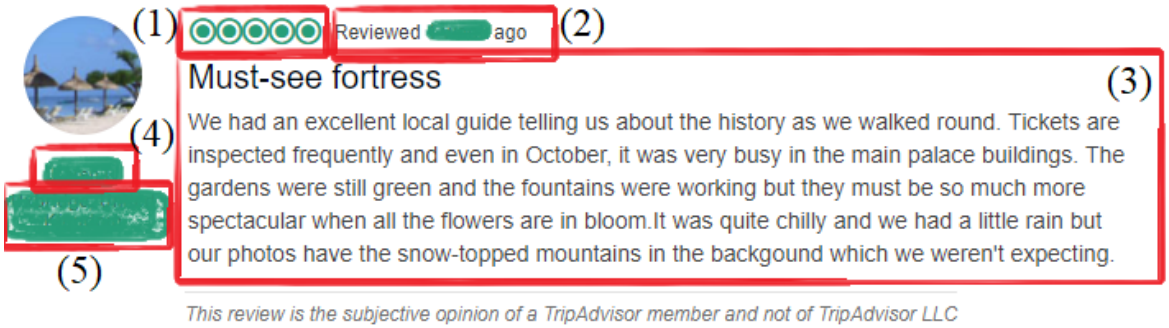



Figure 7: Example of TripAdvisor's review and its features. (1) is the `bubble rating`, (2) the `time`, (3) is the `review`, (4) is the `username`, and (5) the `location`.

Lastly, one of the major concerns of user-generated content is the credibility of opinions. Awarded of this, 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 have carried out so as to analyze credibility and truthfulness of TripAdvisor [JC11, DR18, O'C08]. These works have concluded that TripAdvisor has a consistent and truthful policy system.

# 3 Objectives

After introducing the main concepts related to this work, we present the main objectives that have driven this thesis. They include the use of ML techniques for addressing SA open problems focusing on cultural monument reviews. More specifically the objectives are:

- **To consider the advisability of TripAdvisor as a source for cultural monument reviews**. TripAdvisor has been used for many studies, mining opinions from restaurants or hotels. However, touristic attractions reviews have never been analyzed from a SA perspective. Therefore, we seek to study the viability of setting TripAdvisor as a data source for mining opinions from several cultural monuments.
- **To develop techniques to address the domain adaptation problem of SAMs**. Supervised sentiment methods depend strongly on the domain, whereas unsupervised methods depend on the linguistic resources. Moreover, the development of SAMS is very costly because the preparation of a corpus is not trivial, it is a task that can last several months. In this thesis we aim at developing more efficient techniques to (1) classify polarities without the need of building an exclusive SAM for it and (2) address the lack of generalization.
- **To address sentiment's inconsistencies detected on cultural monument reviews**. We detect that there exist inconsistencies between the User Rating and some sentences of its review. This is due to the fact that when people write do not contend the same sentiment of the overall text. Therefore, it is desirable to build a unified index that combines the polarity of the expert (User Rating) and the polarity of the sentences.
- **To improve polarity classification by filtering neutral polarities**. Neutral reviews are in most cases removed from classification models. This is mainly due to the fact that some studies claim that neutrality lack of information due to its ambiguity. We aim to consider neutral reviews as noise to enhance binary classification results (positive and negative polarities).
- **To develop a methodology that summarizes a large quantity of cultural reviews**. Most of the SA studies focus on either classification of polarities or extraction of detailed information from reviews like aspects. We identify a lack of methods capable of summarizing a large amount of reviews. Therefore, we propose to build a novel methodology that fills this gap. The end goal is to provide a system that will guide the decision-making process of cultural operators, enhancing the aspects that characterize negative reviews.

# 4 Methodology

This thesis requires the application of a methodology that is both theoretical and practical. Therefore, we need a strategy that, while maintaining the guidelines of the traditional scientific method, is able to provide the special needs of such methodology. In particular, the following guidelines for the research work and experiments will be applied:

1. **Observation**: through the study of the Sentiment Analysis problem and its specific characteristics in different domains, as well as the evaluation of several off-the-shelf techniques on cultural monument reviews.
2. **Hypothesis formulation**: design of new Sentiment Analysis methods that make use of the approaches that have been highlighted as promising to improve the process of open Sentiment Analysis problems focusing on cultural monuments. The new methods should implement the characteristics described in the previously mentioned objectives to face the problem of inconsistencies and domain adaptation of TripAdvisor reviews.
3. **Observation gathering**: getting the results obtained by the application of the new methods, on different cultural monument reviews and using different types of performance measures. Both the accuracy and the precision have to be taken into account.
4. **Contrasting the hypothesis**: Comparison of the results obtained with those published by other methods related to Sentiment Analysis state-of-the-art.
5. **Hypothesis proof or refusal**: Acceptance or rejection and modification, in due case, of the developed techniques as a consequence of the performed experiments and the gathered results. If necessary, the previous steps should be repeated to formulate new hypothesis that can be proven.
6. **Scientific thesis**: Extraction, redaction and acceptance of the conclusions obtained through out the research process. All the proposals and results gathered along the entire process should be synthesized into a memory of the thesis.

# 5 Summary

This Section 5 presents a summary of the publications associated to this thesis. After that, in Section 6 we described the main results obtained by these proposals. Both the research carried out for this thesis and the associated results are collected into the published journal publications listed below:

- A. Valdivia, MV. Luzón, F. Herrera. Sentiment analysis in tripadvisor. IEEE Intelligent Systems 32 (4), 72-77 (2017). doi: doi.org/10.1109/MIS.2017.3121555.
- A. Valdivia, E. Hrabova, I. Chaturvedi, MV. Luzón, L. Troiano, E. Cambria, F. Herrera. The Inconsistencies on TripAdvisor Reviews: a Unified Index between Users and Sentiment Analysis Methods. Neurocomputing.
- A. Valdivia, MV. Luzón, E. Cambria, F. Herrera. Consensus vote models for detecting and filtering neutrality in sentiment analysis. Information Fusion 44 126-135 (2018). doi: doi.org/10.1016/j.inffus.2018.03.007.
- A. Valdivia, E. Martínez-Cámara, I. Chaturvedi, MV. Luzón, E. Cambria, YS. Ong, F. Herrera. What do people think about this monument? Understanding Negative Reviews via Deep Learning and Descriptive Rules. Journal of Ambient Intelligence & Humanized Computing.

The remainder of this section is organized following the objectives presented in Section 3 and these publications. Firstly, Section 5.2 presents different models to improve polarity classification by filtering neutral reviews. After that, Section 5.3 provides a method to summarize negative reviews combining three algorithms of: aspect extraction, clustering, and rules association. Finally, Section 5.1 presents a SA-based study focus on TripAdvisor and cultural monument reviews and describes the inconsistencies found between users and SAMs polarities and presents a numerical index to address this problem.

## 5.1 The Problem of Inconsistencies on Cultural Monument Reviews

Many applications of SA has focused on analyzing sentiments of products, companies, political parties, etc. To the best of our knowledge, there was in the literature a gap for cultural monuments reviews. We detected that these organizations need for a digitization process in order to enhance the cultural experience of visitors. We aimed at proposing an alternative methodology to surveys through SA.

We studied the feasibility of TripAdvisor as a source of information for SA-based studies. SAMs are methods capable of extracting polarities from reviews which are necessary to understand people's thoughts towards the entities that they are reviewing about. However, the development process of SAMs is time-consuming and costly (corpus creation, labelling process, ...). We proposed to analyze the performance of several off-the-shelf SAMs on TripAdvisor's monument review. We showed that the domain adaption problem of these methods is generalized, it does not depend on the method used for the development of the SAM. We detected highly inconsistencies between users (bubble rating) and SAMs polarities. We also highlighted that these inconsistencies can also be motivated by the fact that TripAdvisor users do not usually write sentences with the same polarity of the global evaluation.

To tackle this problem, we designed a function that aggregates user and SAMs polarities based on the geometrical mean. We found that the mathematical properties of this mean are suitable for our purposes. For example, it is usually used when the numbers show different behaviours, which is the case of SAMs outputs.

The experimental scenario applied on 6 monuments and more than 80,000 reviews have shown an improvement on user-SAM inconsistencies. We evaluate our model in an aspect scenario and the results clearly shown that the aggregated polarities obtained by our model absorbed the information from the two sources.

The works that are associated to this part are:

A. Valdivia, MV. Luzón, F. Herrera. Sentiment analysis in tripadvisor. IEEE Intelligent Systems 32 (4), 72-77 (2017). doi: doi.org/10.1109/MIS.2017.3121555.

A. Valdivia, E. Hrabova, I. Chaturvedi, MV. Luzón, L. Troiano, E. Cambria, F. Herrera. The Inconsistencies on TripAdvisor Reviews: a Unified Index between Users and Sentiment Analysis Methods. Neurocomputing.

## 5.2 Neutrality Detection Guided by Consensus Models

The most known task of SA is polarity classification which aims at detecting polarities in texts. These polarities are generally defined as *positive*, *negative*, and *neutral*, but in many cases neutral reviews are separated from the analysis, becoming insignificant. Neutrality is also considered to be an ambiguous class. This has meant that SAMs usually show a low consensus when detecting it.

In this work we aimed at considering neutrality as the noisy class in order to enhance classification results at document level. Because we detected that SAMs do not agree on detecting neutral polarities, we proposed to aggregate their opinions to better detect neutrality, and then filter it. We claimed that this preprocessing step improve polarity classification between positive and negative opinions.

To do so, we propose IOWA operators guided by two weight aggregations: (1) distance-based function to neutrality and (2) linguistic quantifiers. The first type of operators define weights guided by the distance of SAMs polarities to the neutral point, which is considered to be the 0.5 in the $[0, 1]$ interval. In this type of aggregation we also considered the average aggregation where all weights have the same value. In the second case, the weight function is built by three linguistic quantifiers: *at least half*, *most of*, *many as possible*. Overall, we build 7 different aggregation systems.

In order to evaluate our models, we applied 5 SAMs on 4,500 reviews from different domains like product reviews, movie sentiments, cultural monument opinions, etc. We showed that there exists a low consensus on neutral reviews between these 5 SAMs. Therefore, we detect and filtered neutrality by aggregating polarities driven by each of the 6 IOWA operators. After that, we studied the behaviour of two ML classifiers (SVM and XGBOOST) in order to evaluate the performance of our preprocessing models (see Figure 8).

The research journal associated to this part is:

A. Valdivia, MV. Luzón, E. Cambria, F. Herrera. Consensus vote models for detecting and filtering neutrality in sentiment analysis. Information Fusion 44 126-135 (2018). doi: doi.org/10.1016/j.inffus.2018.03.007.

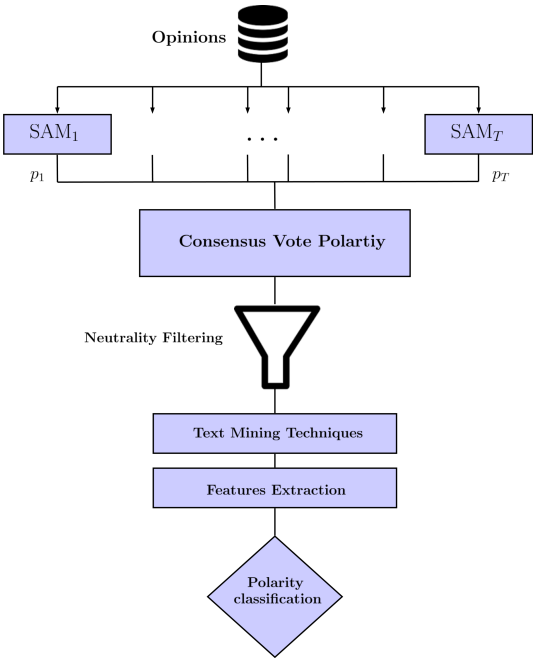



Figure 8: Proposed methodology to evaluate detection and filter of neutral reviews.

## 5.3 Opinion Summarization

ABSA is aimed at extracting the most detailed information about reviews, it links aspects that are mentioned in reviews with a polarity. ABSA is also considered the most challenging task of SA due to its in-depth level of analysis. Owing to this fact, SA researchers has focused on developing techniques based on this task for the last few years.

We detected that despite the fact that the information obtained by ABSA-based models is very specific, it does not allow to summarize the content of reviews. We claimed that it is necessary to develop methods that contribute to obtain an overview of the main features of sentiments. Moreover, we also assert that this summary should take into account the polarity, so we can detect the most significant entities and aspects of positive, negative or even neutral reviews.

We proposed a novel methodology that extract aspects and summarize reviews in a rule-based visualization. This method combines three algorithms from different areas: Aspect Extraction, Aspect Clustering and Descriptive Rules (see Figure 9). The first algorithm is a CNN described in Section 2.3 which focus on extracting aspects from reviews. However, we detect that many of the aspects referred to a same concept. The human language is very rich and we may use different words for conveying a same idea. Therefore, we detected the necessity of clustering aspects and decrease its high-dimensionality. After that, aspects are grouped into clusters using the elbow method and k-nn algorithm. Finally, we used a rule association algorithm adjusted for subgorup discovery, apriori-sd, to extract the most relevant aspects regarding a polarity. Rules consists of a group of aspects as the antecedent and the value of polarities as the consequent.

We focused our experiments on cultural monument reviews and its negative reviews. The purpose was to figure out if our methodology is able to help cultural operators in decision-making processes by detecting features that people complain about their monuments.

The publication related to this section is:

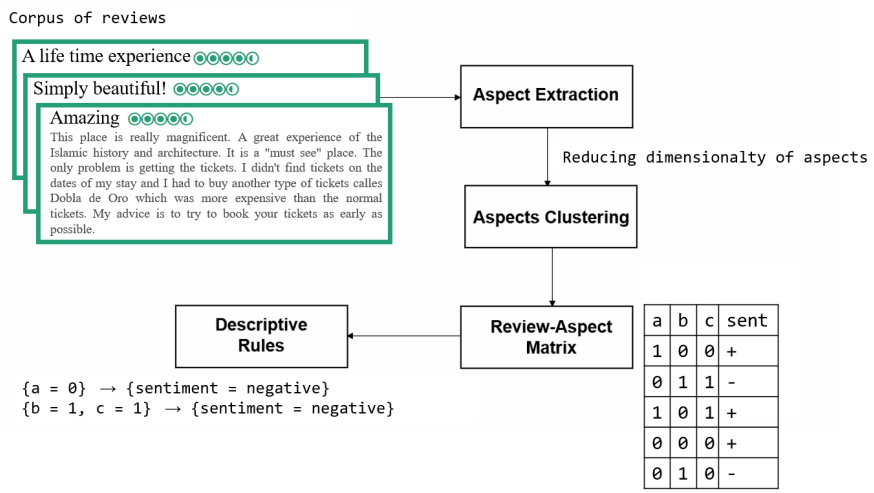



Figure 9: Proposed workflow for summarizing reviews based on aspect level.

A. Valdivia, E. Martínez-Cámara, I. Chaturvedi, MV. Luzón, E. Cambria, YS. Ong, F. Herrera. What do people think about this monument? Understanding Negative Reviews via Deep Learning and Descriptive Rules. Journal of Ambient Intelligence & Humanized Computing.

# 6 Discussion of results

The following sections present the main results and further discussion motivated by the research conducted in this thesis.

## 6.1 The Problem of Inconsistencies on Cultural Monument Reviews

SAMs that are trained in a certain domain may not be applied in other context reviews because of two facts: the domain adaptation problem and the diversity of polarities in a same document. The first fact is well-known in the SA community and it is caused by the lack of generalization capacity of ML models and the lexical coverage limitation of linguistic resources. Users and SAMs polarity distributions on cultural monuments reviews shown a widely varying behaviour.

We first analyzed four off-the-shelf SAMs (Bing, CoreNLP, SentiStrenght, and Syuzhet) on three monument reviews (Alhambra, Mezquita, and Sagrada Família) from TripAdvisor. We compared SAMs distributions with users distributions and observed that whereas users generally express a positive polarity (more than 90% of reviews where evaluated with 4 or 5 bubbles), SAMs obtained more negative reviews (CoreNLP detected almost 40% of negative reviews on Alhambra's dataset). We then proposed to replicate these study with more SAMs and monument reviews. The results shown that the inconsistencies problem is generalized: there is no SAM standing out on detecting the same polarity of the user. Moreover, neutral reviews are those that obtain worst results, being SentiStrength the method that better detects user's neutrality with 37.20 % of accuracy.

Thus, we claim to propose a unified index to tackle the correlation problem of polarities. We aimed at combining both polarities with a parameterized function of the geometric mean. We run several analysis for studying the performance of our aggregation model. The first analysis is to study the relationship of our proposal with users and SAMs polarities. We shown that our aggregation model clearly obtains in-between polarities. After that, we run an aspect analysis taking into account our model. We conclude that aspects that obtain opposite polarities (by users and SAMs) end up getting averaging scores (by our model) which led us to obtain more reliable conclusions.

## 6.2 Neutrality Detection Guided by Consensus Models

In this work we claimed that there is a need for a consensus voting model to detect neutrality. We aimed at improving polarity classification by first detect and filter neutral polarities. We evaluated the consensus of 7 SAMs and shown the low agreement that exists.

To address this problem, we proposed several aggregation models based on fuzzy operators. We clearly observed that the agreement rate increases using the aggregators. After that, we run several classification models preprocessing neutral reviews by setting them as the noisy class, i.e., filtering them. We used to ML classifiers (SVM and XGBOOST) and compared the results with each of the individual SAMs. The results of our comparision analysis highlighted that:

- There was no outstanding SAM on the classification performance of individual methods. This led us to conclude that SAMs strongly depend on the dataset (the domain adapation problem).

- Linguistic Quantifiers shown better results, and more specifically *at least one* model obtained the best average classification results. We asserted that good results are owing to the fact that this operator did not take into account extreme polarities, weighting only more neutral SAMs than other operators
- We then compared individual SAMs results with the *at least one* model and observed that this model outperforms in many cases. It also obtained better results on average.

SAMs aggregation models have proven to obtain great results in classification tasks [CHVHA07]. Our results also let us concluded that they can also be applied to detect neutrality, which can be filter and enhance classification results. Finally, we proposed to compare the aggregation model with a ground truth. In this sense, we proposed to label opinions by different experts and then build the same models to compare classification results.

## 6.3 Opinion Summarization

In this last part of the thesis, we have presented a novel methodology for summarizing reviews. The idea presented aimed at identifying the most relevant features of a polarity. We proposed to combine three different methods: neural network for extracting aspects, clustering algorithm for gathering those that are similar, and subgroup discovery technique to find the most substantial rules.

We designed an experimental framework based on three monument datasets. We focused on extracting the most significant aspects of negative reviews, since we claimed that those reviews are the key for obtaining insightful information to enhance the visitor's experience on these organizations. The clustering results shown that aspects are effectively grouped into clusters that represent a same general idea. The rules obtained very low scores of precision, confidence, weighted relative accuracy. We detected that this is due to two main facts: negative reviews are underrepresented and aspects occurrence is very low. Negative sentiments represented the 6% of reviews at most, which implies that support measure got very low values. Aspects presented also very low frequencies, which led to sparse matrices. This also affected our methodology which discovered no meaningful rules like: $\{\texttt{aspect = 0}\} \longrightarrow \{\texttt{negative}\}$.

Our results shown that our methodology is suitable for opinion summarization. We clearly detected the most relevant features of negative polarities like: *staff*, *queue*, *guard*, etc. The most remarkable feature of our methodology is that it does not depend on the context, which is a well-known limitation of many SA-based applications. The weakness of our experimentation setup is that it takes into account users polarities, which may led misinterpretations since we detected that users usually convey different polarities on a same review. We then suggested to evaluate our methodology with polarities based on sentence or aspect level.

# 7 Concluding remarks

This thesis presents a broad study about SA-based methods (SA) following a common goal: the evaluation of these techniques applied on sentiments about monumental institutions. The digitization of cultural heritage has become one of the main objectives of public policy, promoting regions, making institutions visible and protecting their cultural diversity. That is why we believe it is important to use algorithms to extract insights from opinions conveyed on Internet's platforms, and thus improve the cultural experience in these institutes.

The first objective we have faced is to study the possibility of proposing TripAdvisor as a reliable source of opinions. It is relevant to find truthful information to carry out SA studies, so that results can be consistent. To do this, we proposed an analysis of off-the-shelf methods in sentiments of three cultural monuments taken from this source. The results showed a high inconsistency between the polarity between users and algorithms.

We detected that the low correlation is due to two main factors: (1) the problem of domain adaptation and (2) the variability of polarities on sentences of a same document. The first problem is widely known by the academic community and is one of the biggest challenges to face in this field. It has been shown that the performance of polarity extraction methods trained in a particular context drops considerably if evaluated in other contexts. If we apply a method trained with movie opinions and we evaluate it in restaurant reviews, the accuracy and precision will be generally lower. The second factor that affects these inconsistencies is the fact that people convey different polarities in a same text. In our study we detected that, for example, people visiting the Alhambra evaluate their visit with a 5, but express disagreement with the tiicket system.

Another goal addressed in this thesis has been to address the inconsistencies problem. For this, we proposed the development of an index that collected both polarities (users and algorithms). This index was constructed from the parametrization of the geometric mean, which has been widely used in aggregation models for its mathematical properties. The results obtained showed that this index effectively unifies the polarities obtained by both agents.

The study of neutrality is considered other of the objectives of this thesis. Neutral opinions are in many cases avoided from opinion mining and decision making processes. We also have detected the low consensus that exists between very different methods of polarity extraction when detecting neutral opinions. That is why in our study we proposed to give neutral opinions a value to improve the classification of positive and negative polarities. The idea was to consider neutrality as the noisy class, detect them with different consensus models, and then filter it. We evaluated of our models with two classification models. The results showed that preprocessing based on consensus aggregations improved the performance of many individual methods. Thus, we concluded that neutral class should not be eluded, but it can improve the detection of both positive and negative polarities.

The last objective proposed was to create a methodology capable of summarizing a large number of opinions. One of the gaps that we detect in the SA literature is the need to create methods that are capable of characterizing opinions according to their polarity. The proposed idea tries to find those most significant aspects given a polarity in the form of a rule. The method combined three algorithms that come from three different areas: deep learning, clustering y subgroup discovery. The results obtained show that it is clearly possible to draw valuable conclusions for decision making processes in cultural monuments, detecting the most relevant aspects of negative

opinions.

From the conclusions drawn from this thesis, new and promising lines of research can be proposed. These lines are based on the creation of methods and techniques that face the challenges that are still open within the SA literature:

- **Evaluation studies between the proposed and classic methods.** In this thesis we have developed alternative methods for extracting insights from visitors' sentiments. The results show that the conclusions drawn are consistent and solid to assist in decision-making processes. However, once this thesis is finished, we find the need to elaborate studies that contrast these results with those obtained by classical methodologies such as surveys.
- **Extension of SA-based methods for other languages.** Another gap detected in the development of this work is the creation of a corpus based on cultural heritage opinions. This task is of utmost importance for the development of more SA-based studies at different levels (document, phrase and aspect). The publication of this corpus will also mean the generation of new methods that can improve the results obtained by those proposed in this thesis.
- **Extension of SA-based methods for other languages.** Another shortcoming detected in the development of this thesis is to extend studies and methods to other languages than English. All the works that built this thesis have been based on English sentiments. Thus, it is necessary to apply our methodology to opinions written in other languages, in order to evaluate and take into account other ways of expression.
- **Adaptation domain problem.** As we have already mentioned, the lack of generalization of SA-based methods is one of the main open problems in this area. Therefore, the study of new techniques that confront this conflict is of total relevance. We believe that the path to overcome this problem can be marked by architectures based on deep learning, because over last years they have shown an efficient performance in many of the open problems within the Natural Language Processing, beating in many cases the results of the state of the art.

## Conclusiones

Esta tesis presenta un amplio estudio sobre métodos basados en AO siguiendo un objetivo común: la evaluación de estas técnicas usando opiniones entorno a instituciones monumentales. La digitalización del patrimonio cultural se ha convertido en uno de los objetivos principales de la mayoría de políticas públicas de muchos países, promoviendo regiones, visibilizando instituciones y protegiendo su diversidad cultural. Es por ello que creemos relevante el uso de algoritmos para extraer conocimiento a partir de opiniones volcadas en la red, y así mejorar la experiencia cultural en estos centros.

El primer objetivo que hemos afrontado es el de estudiar la posibilidad de proponer TripAdvisor como una fuente fiable de opiniones. Es de total relevancia encontrar información fiable y contrastable para realizar nuestros estudios de análisis de opiniones, para que así los resultados sean consistentes. Para ello, propusimos realizar un análisis de diferentes métodos ya desarrollados en opiniones de tres monumentos culturales extraídas de esta fuente. Los resultados mostraton una alta incosistencia entre la polaridad del/la usuario/a y el algoritmo, es decir, existía poco consenso a la hora de evaluar el texto de una reseña.

En el estudio desarrollado, detectamos que la baja correlación es debida a dos factores: (1) el problema de adaptación al dominio y (2) la variabilidad de polaridades en frases de un mismo documento. El primer problema es ampliamente conocido por la comunidad académica y es uno de los mayores retos afrontar dentro de este campo. Se ha demostrado que el rendimiento de métodos de extracción de polaridad entrenados en un contexto en concreto baja considerablemente si se evalúan en otros contextos. Es decir, si utilizamos un método entrenado con opiniones de películas y lo evalúamos en opiniones de restaurantes, el acierto a la hora de predecir la polaridad será mucho menor. El segundo factor que afecta a estas inconsistencias es el hecho que las personas expresamos diferentes polaridades en un mismo texto. En nuestro estudio detectamos que, por ejemplo, muchas de las visitas al monumento de la Alhambra evalúan su visita con un 5, pero expresan disconformidad con el sistema de entradas.

Otro de los objetivos afrontados en esta tesis ha sido el de de crear una solución a este problema de las incosistencias. Para ello, propusimos el desarrollo de un índice que recogiera ambas polaridades (personas y algoritmos). Este índice fue construido a partir de la parametrización de la media geométrica, la cuál ha sido ampliamente utilizada en modelos de agregación por sus propiedades matemáticas. Los resultados obtenidos mostraron que este índice unifica efectivamente las polaridades obtenidas por ambos agentes, y que por lo tanto corrigen el problema de las incosistencias.

Otro de los objetivos a tratar ha sido el del estudio de la neutralidad. Las opiniones neutras son en muchos casos menospreciadas de los procesos de análisis de opiniones y tomas de decisiones. Debido a ello, en esta tesis detectamos y porbamos el bajo consenso que existen entre métodos muy variados de extracción de polaridad a la hora de detectar opiniones neutras. Es por ello que en nuestro estudio propusimos darle a las opiniones neutras un valor para mejorar la clasificación de polaridades positivas y negativas. La idea fue considerar la neutralidad como la clase ruido, detectarla con diferentes modelos de consenso, para posteriormente filtrarla. Luego se evaluaba el rendimiento de dos modelos de clasificación. Los resultados mostraron que el preprocesamiento basado en las agregaciones por consenso mejoraban el rendimiento de la mayoría de métodos por individual. Así, concluimos que la clase neutral no debe ser eludida, sino que puede mejorar la detección de polaridades tanto positivas como negativas.

El último objetivo propuesto fue el de crear una metodología capaz de resumir una gran cantidad de opiniones. Uno de los vacíos que detectamos en la literatura del AO es la necesidad de crear métodos que sean capaces de caracterizar las opiniones según su polaridad. La idea propuesta trata de encontrar aquellos aspectos más significativos dada una polaridad en forma de regla. El método combina tres algoritmos que provienen de tres áreas diferentes: *deep learning*, *clustering* y *subgroup discovery*. Los resultados obtenidos muestran que claramente se pueden sacar conclusiones valiosas para la toma de decisiones en monumentos culturales, detectando aquellos aspectos más relevantes de las opiniones negativas, por ejemplo.

De las conclusiones extraídas por esta tesis se pueden proponer nuevas y prometedoras línias de investigación. Estas vías estan basadas en la creación de métodos y técnicas que afronten los retos que aún quedan abiertos dentro del AO:

- **Estudios de evaluación entre los métodos propuestos en esta tesis y los clásicos.** En esta tesis hemos desarrollado métodos alternativos para la extracción de conocimiento a partiro de opiniones de visitantes. Los resultados muestran que las conclusiones extraídas son consistenes y sólidas para asistir en procesos de toma de decisiones. No obstante, una vez finalizada esta tesis encontramos la necesidad de elaborar estudios que contrasten estos resultados con los obtenidos por metodologías clásicas como encuestas.

- **Creación de un corpus sobre patrimonio cultural.** Una de las carencias básicas que hemos detectado en el desarrollo de este trabajo es la creación de un corpus basados en opiniones de patrimonios culturales. Esta tarea es de total importancia para el desarrollo de más estudios basados en AO a diferentes niveles (documento, frase y aspecto). La publicación de este corpus supondrá también la generación de nuevos métodos que puedan mejorar los resultados obtenidos por los propuestos en esta tesis.

- **Extensión de métodos basados en AO para otros idiomas.** Otra de las carencias detectadas en el desarrollo de esta tesis es la de extender estudios y métodos a otros idiomas que no sean el inglés. Todos los trabajos que forman esta tesis se han basado en opiniones de monumentos culturales publicadas en inglés. No obstante, nos parece necesario aplicar toda nuestra metodología a opiniones escritas en otros idiomas, para así evaluar y tener en cuentra otras maneras de expresarse.

- **Métodos para afrontar el problema de adaptación al dominio.** Como ya hemos comentado anteriormente, la falta de generalización de los métodos de AO es uno de los principales problemas abiertos de esta área. Por lo tanto, el estudio de nuevas técnicas que afronten este conflicto es de total relevancia. Creemos que la senda a seguir para vencer este problema puede venir marcada por arquitecturas basadas en *deep learning*, pues durante los últimos años han mostrado un eficiente rendimiento en muchos de los problemas abiertos dentro del Procesamiento del Lenguaje Natural, batiendo en muchos casos los resultados del estado del arte.

# Chapter II

# Publications

# 1 Sentiment Analysis in Tripadvisor

- A. Valdivia, MV. Luzón, F. Herrera. Sentiment analysis in tripadvisor. IEEE Intelligent Systems 32 (4), 72-77 (2017)
  - Status: **Published**.
  - Impact Factor (JCR 2017): **2.596**
  - Subject Category: **Computer Science, Artificial Intelligence**
  - Rank: **38/132**
  - Quartile: **Q2**

# SENTIMENT ANALYSIS IN TRIPADVISOR

A PREPRINT

**Ana Valdivia***
Department of Computer Science
and Artificial Intelligence
Universidad de Granada
Granada, Spain 18014
avaldivia@ugr.es

**M. Victora Luzón**
Department of Software Engineering
Universidad de Granada
Granada, Spain 18014
luzon@ugr.es

**Francisco Herrera**
Department of Computer Science
and Artificial Intelligence
Universidad de Granada
Granada, Spain 18014
herrera@decsai.ugr.es

***K*eywords** Sentiment Analysis · TripAdvisor

## 1 Introduction

The number of Web 2.0 has recently experienced important growth. These websites emerged as an evolution of Web 1.0 or static websites. In Web 2.0 the user is not only a content consumer, but also is able to generate the content and collaborate with other users. In this way, the users take an active role and create a virtual community. There are different types of Web 2.0: blogs (Blogger or Wordpress), media content (Prezi, Youtube, Flickr), wikis (Wikipedia, WikiSpace), collaborative (Dropbox, GoogleDocs) and social networks (Twitter, Facebook, Google+). The burgeoning information explosion offered through the Web 2.0 has implied that end customers often check other user's opinions in forums, blogs and social networks before buying a product or contracting a service.

TripAdvisor emerged in 2004 as a Web 2.0 for the tourism domain. This user-generated content website offers a large amount of reviews from travelers' experiences regarding hotels, restaurants and tourist spots. TripAdvisor has since been ranked as the most popular site for trip planning. Nowadays, million of tourists arrange their holidays taking into account TripAdvisor reviews (see Figure1).

Sentiment Analysis (SA) is extremely useful when monitoring Web 2.0, allowing us to know public opinion of a large number of issues without the need for satisfaction inquiries [1, 2]. According to the Oxford dictionary, sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text to determine whether the writer's attitude toward a particular topic, product, and so on is generally positive, negative, or neutral. The interest in sentiment analysis has increased significantly over the last few years due to the large amount of stored text in Web 2.0 applications and the importance of online customer opinions. As a result, more than 1 million research papers contain the term "sentiment analysis," and various start-ups have been created to analyze sentiments in social media companies.

Multiple studies on TripAdvisor exist, but there is no complete analysis from the sentiment analysis viewpoint. This article proposes TripAdvisor as a source of data for sentiment analysis tasks. We develop an analysis for studying the matching between users' sentiments and automatic sentiment-detection algorithms. Finally, we discuss some of the challenges regarding sentiment analysis and TripAdvisor, and conclude with some final remarks.

*Corresponding author

Figure 1: Granada (Paseo de los Tristes). TripAdvisor is the most popular site for planning a trip.

## 2 TripAdvisor and Sentiment Analysis

According to Wikipedia, TripAdvisor is an American travel website company providing reviews from travelers about their experiences in hotels, restaurants, and monuments. Stephen Kaufer and Langley Steinert, along with others, founded TripAdvisor in February 2000 as a site listing information from guidebooks, newspapers, and magazines. InterActiveCorp purchased the site in 2004, and one year later, spun off its business travel group, Expedia. After that, the website turned to user-generated content. It has since become the largest travel community, reaching 390 million unique visitors each month and listing 465 million reviews and opinions about more than 7 million accommodations, restaurants, and attractions in 49 markets worldwide. Figure 2 shows the Google search rate for TripAdvisor, illustrating its popularity around the world.

Because it has so much data, TripAdvisor has become extremely popular with both tourists and managers. Tourists can read the accumulated opinions of millions of everyday tourists. They can also check the popularity index, which is computed using an algorithm that accounts for user reviews and other published sources such as guidebooks and newspaper articles. This index runs from number 1 to the overall total number of restaurants, hotels, or other attractions within the city. Travelers can find the most interesting visitor attraction or most popular restaurant. Linked to this is the bubble rating (user rating), a 1–5 scale where one bubble represents a terrible experience and five bubbles an excellent experience. All reviewers are asked to use this scale to summarize their feedback. Together with this rating, users include their opinions, which can cover the performance of a restaurant, hotel, or tourist spot. Therefore, reading and analyzing reviews can help develop a business.

The World Travel & Tourism Council report shows that tourism generates 9.8 percent of the wider gross domestic product and supports 248 million jobs [2]. These numbers suggest that the tourism industry is the most important economic driver of many economies. Therefore, it's important to understand the main drivers of the tourist flow as well as tourists'opinions about a city's restaurants, hotels, and tourist attractions.

TripAdvisor has enough standing to be used as a text source [3] storing numerous reviews of tourist businesses around the world. Sentiment analysis extracts insights from this data. Sentiment classification, the best-known sentiment analysis task, aims to detect sentiments within a document, a sentence, or an aspect. This task can be divided into three steps: polarity detection (label the sentiment of the text as positive, negative, or neutral), aspect selection/ extraction (obtain the features for structuring the text), and classification (apply machine learning or lexicon approaches to classify the text). Sentiment analysis methods (SAMs), which are trained for sentiment polarity detection [4, 5, 6], can automatically detect sentiments from documents, sentences, or words. A large variety of SAMs address the different categories of texts (blogs, reviews, tweets, and so on). However, the analysis of feelings is not a perfect science,

[2] www.wttc.org/-/media/files/reports/economic%20impact%20research/regions%202016/world2016.pdf.

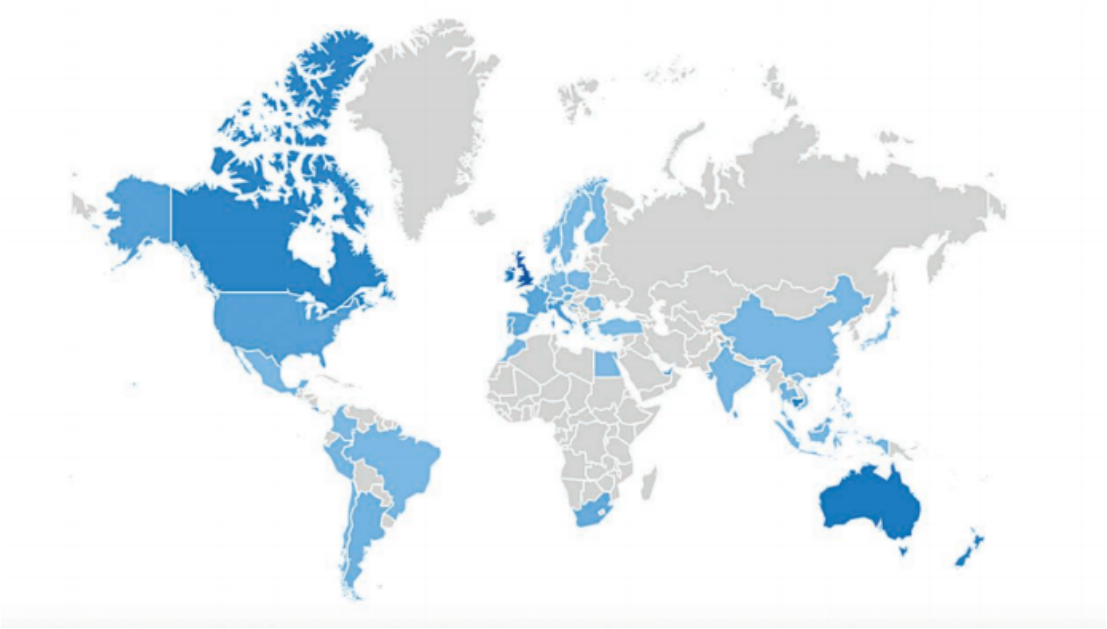


Figure 2: Map of the popularity of TripAdvisor searches in Google over the last five years.

especially when applied to the unstructured texts that predominate in social networks [7]. Human language is complex, so teaching a machine to detect different grammatical nuances, cultural variations, jargon, and misspellings in messages on the network is a difficult process, and it is even more difficult to automatically understand how the context can affect the message's tone. Because humans can apply a contextual understanding, we can intuitively interpret the intentionality of any writing. Computers, however, have difficulty understanding the context in which a phrase is expressed and detecting whether a person is being sarcastic or not.

A few sentiment analysis studies set TripAdvisor as a data source. For example, researchers have collected reviews about the TripAdvisor app in Google Play Store for extracting app features to help developers [8]. Others analyzed TripAdvisor's hotel reviews for classifying good and bad customer opinions [9]. Still others propose a system to summarize comments from travel social networks, such as TripAdvisor for analysis [10]. Similarly, other researchers developed a tool to analyze tourists' opinions of restaurants as well as hotels from a region in Chile [11].

## 3 TripAdvisor: A study for calibrating user's polarity

We scrape TripAdvisor webpages on three well-known monuments in Spain: Alhambra, Mezquita Córdoba, and Sagrada Familia. We consider user ratings of one and two bubbles as negative, three as neutral, and four and five as a positive sentiment. We then apply four SAMs (SentiStrength [12], Bing [13], Syuzhet [14], and CoreNLP [15]) and extract the overall polarity on each opinion.

Figure 3 shows the results. We observe that the distributions of polarities are different from user ratings. The user ratings bar plot shows that users tend to rate their visits to the three monuments positively, with more than 90 percent of ratings having four or five bubbles. SentiStrength and Syuzhet methods reach a similar distribution to the user rating. However, Bing and CoreNLP detect more negativity in the TripAdvisor opinions. In all cases, the number of neutral polarities is higher than the neutral ratings (three bubbles).

Next, we studied the distribution of bubble ratings over the negative SAM polarities. We thus analyzed the behavior of user feedback against the SAM evaluations. Figure 4 presents 12 bar plots (four SAMs for each of three monuments) containing the shares of all negative SAM polarity over the bubble evaluation (percent over the original user ratings). Analyzing this data, we observe that SentiStrength and Syuzhet detect at best 57.48 and 45.10 percent of negative reviews. However, they misclassify on average 20 percent of positive user reviews (three and four bubbles). On the other hand, Bing and CoreNLP methods detect as negative more negative user ratings, but misclassify over 30 percent of positive reviews. Bing and CoreNLP tend to highlight the negative opinions.

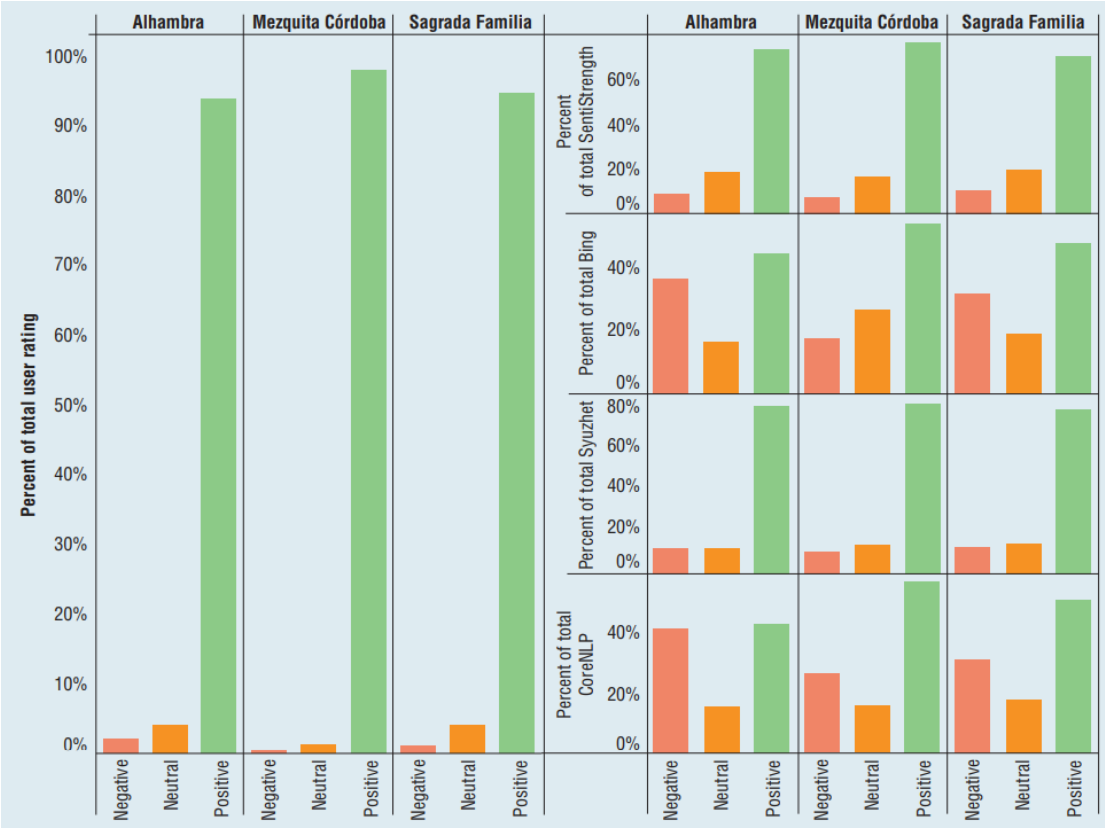



Figure 3: Distribution of sentiments between TripAdvisor users (bubble ratings) and four sentiment analysis methods (SAMs): SentiStrength, Bing, Syuzhet, and CoreNLP. Red indicates negative sentiments, orange neutral, and green positive.

In general, we observe that users tend to write negative sentences on positive user ratings, and vice versa. Therefore, we suggest not setting the user rate as a label sentiment for the whole review and analyzing the opinions in depth.

This study clearly shows the need to analyze opinions beyond user ratings. As a practical methodology, we propose following three steps: handle the negative opinions identified with SAMs via learning models, get a good clusterization according to consensus degrees among SAMs, and discover relationships among common aspects to characterize the cause behind the negative comments.

## 4 Challenges

Several challenges arise when TripAdvisor uses sentiment analysis, due to the specific content in TripAdvisor-based opinions. Although some related topics have been extensively studied in the literature, their adaptation to the context of TripAdvisor opinions requires revisiting them.

Aspect-based sentiment analysis (ABSA) is an important sentiment analysis task [16]. An aspect refers to an attribute of the entity, for example, hotel room cleanliness, the staff at a tourist spot, or the service at a restaurant. ABSA aims to identify the sentiment toward an aspect and extract fine grained information about specific TripAdvisor-based opinions (hotels, monuments, restaurants, and so on). Recent relevant studies are based on deep learning [17], which should be analyzed in the TripAdvisor context.

ABSA is helpful to business managers because it allows for the extraction of transparent customer opinions. Discovery knowledge techniques such as subgroup discovery [18] can be applied to discover relationships among common aspects and get aspect associations for both positive and negative opinions.

The detection of irony and sarcasm is a complex sentiment analysis task. The detection of ironic expressions in TripAdvisor reviews is an open problem that could help to extract more valuable information about the study's

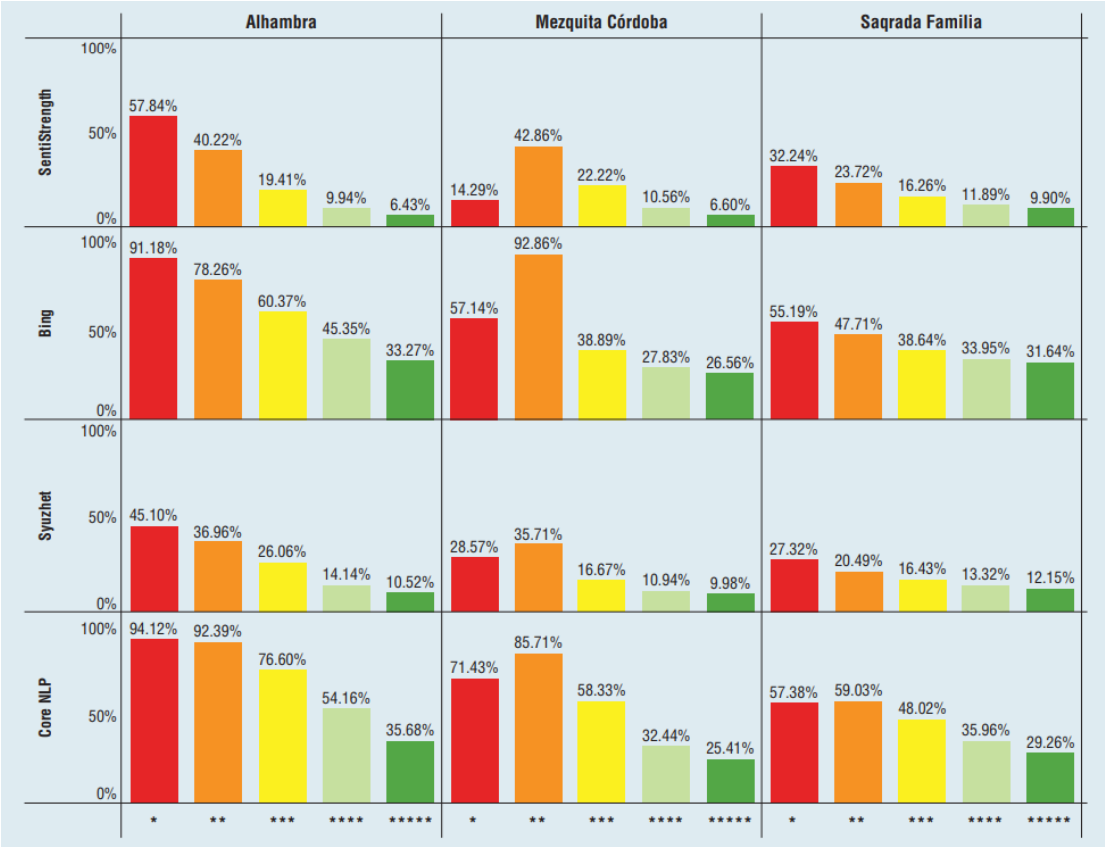



Figure 4: Distribution of SAMs' negative polarity by user rating in TripAdvisor.

subject [19]. Spam is another sentiment analysis-related concern. Some authors have developed studies to measure the credibility of TripAdvisor with satisfactory results [20].

A novel approach is the extraction of aspects/features from opinions to raise the issue as a bag of feature vectors, considering the problem as multi-instance learning [21]. This might provide a robust approach from the classification viewpoint.

Sentiment analysis is an incipient research field. It is difficult to determine how it will evolve in the future, although there is a general belief that this analysis needs to go beyond a simple classification of texts on a positive and negative one dimensional scale. Over the last few years, the list of sentiment analysis related challenges has grown (subjectivity classification, opinion summarization, opinion retrieval, and so on).

Through Web platforms such as TripAdvisor, tourists can openly describe their experiences and thus affect a business's viability. Therefore, the implementation of sentiment analysis techniques to mine sources of opinion is crucial to understanding the faults and assets of a tourist service. Given the large number of applications in the tourist domain, sentiment analysis has great potential to directly influence quality improvement in tourism.

Because of inconsistencies between user ratings and SAM evaluations, with users often writing negative sentences in positive opinions and vice versa, we need new approaches to fix the positive, negative, and neutrality via consensus among SAMs, as well as design models to discover relationships among common aspects to characterize the reasons behind negative comments.

## Acknowledgments

This work has been supported by FEDER (Fondo Europeo de DEsarrollo Regional) and the Spanish National Research Project TIN2014-57251-P.

## References


[1] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[2] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[3] Peter O'Connor. User-generated content and travel: A case study on tripadvisor. com. *Information and communication technologies in tourism 2008*, pages 47–58, 2008.

[4] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[5] Jesus Serrano-Guerrero, Jose A Olivas, Francisco P Romero, and Enrique Herrera-Viedma. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38, 2015.

[6] Erik Cambria and Amir Hussain. *Sentic computing: a common-sense-based framework for concept-level sentiment analysis*, volume 1. Springer, 2015.

[7] Farhan Hassan Khan, Saba Bashir, and Usman Qamar. Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57:245–257, 2014.

[8] Emitza Guzman and Walid Maalej. How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, pages 153–162. IEEE, 2014.

[9] Dietmar Gräbner, Markus Zanker, Gunther Fliedl, Matthias Fuchs, et al. *Classification of customer reviews based on sentiment analysis*. Citeseer, 2012.

[10] Srisupa Palakvangsa-Na-Ayudhya, Veerapat Sriarunrungreung, Pantipa Thongprasan, and Satit Porcharoen. Nebular: A sentiment classification system for the tourism business. In *Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on*, pages 293–298. IEEE, 2011.

[11] Edison Marrese-Taylor, Juan D Velásquez, and Felipe Bravo-Marquez. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, 41(17):7764–7775, 2014.

[12] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength, 2013. *Cyberemotions: Collective emotions in cyberspace*, 2013.

[13] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[14] Matthew Jockers. Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*, 2017.

[15] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[16] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge & Data Engineering*, (1):1–1, 2016.

[17] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

[18] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[19] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*, 2016.

[20] Raffaele Filieri, Salma Alguezaui, and Fraser McLeay. Why do travelers trust tripadvisor? antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. *Tourism Management*, 51:174–185, 2015.

[21] Francisco Herrera, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, and Sarah Vluymans. *Multiple instance learning: foundations and algorithms*. Springer, 2016.

# 2 The Inconsistencies on TripAdvisor Reviews: a Unified Index between Users and Sentiment Analysis Methods

- A. Valdivia, E. Hrabova, I. Chaturvedi, MV. Luzón, L. Troiano, E. Cambria, F. Herrera. The inconsistencies on TripAdvisor reviews: a unified index between users and sentiment analysis methods. Neurocomputing.
  - Status: **Accepted**.
  - Impact Factor (JCR 2017): **3.241**
  - Subject Category: Computer Science, Artificial Intelligence
  - Rank: **27/132**
  - Quartile: **Q1**

# THE INCONSISTENCIES ON TRIPADVISOR REVIEWS: *A Unified Index between Users and Sentiment Analysis Methods*

A PREPRINT

**Ana Valdivia**[*]
Department of Computer Science
and Artificial Intelligence
Universidad de Granada
Granada, Spain 18014
avaldivia@ugr.es

**Emiliya Hrabova**
Department of Engineering
University of Sannio
Bnevento, Italy 82100
emiliya.hrabova@gmail.com

**Iti Chaturvedi**
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore 639798
iti@ntu.edu.sg

**Luigi Troiano**
Department of Engineering
University of Sannio
Bnevento, Italy 82100
luigi.troiano@gmail.com

**M. Victora Luzón**
Department of Software Engineering
Universidad de Granada
Granada, Spain 18014
luzon@ugr.es

**Erik Cambria**
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore 639798
cambria@ntu.edu.sg

**Francisco Herrera**
Department of Computer Science
and Artificial Intelligence
Universidad de Granada
Granada, Spain 18014
herrera@decsai.ugr.es

## ABSTRACT

TripAdvisor is an opinion source frequently used in Sentiment Analysis. On this social network, users explain their experiences in hotels, restaurants or touristic attractions. They write texts of 200 character minimum and score the overall of their review with a numeric scale that ranks from 1 (Terrible) to 5 (Excellent). In this work, we aim that this score, which we define as the User Polarity, may not be representative of the sentiment of all the sentences that make up the opinion. We analyze opinions from six Italian and Spanish monument reviews and detect that there exist inconsistencies between the User Polarity and Sentiment Analysis Methods that automatically extract polarities. The fact is that users tend to rate their visit positively, but in some cases negative sentences and aspects appear, which are detected by these methods. To address these problems, we propose a Polarity Aggregation Model that takes into account both polarities guided by the geometrical mean. We study its performance by extracting aspects of monuments reviews and assigning to them the aggregated polarities. The advantage is that it matches together the sentiment of the context (User Polarity) and the sentiment extracted by a pre-trained method (SAM Polarity). We also show that this score fixes inconsistencies and it may be applied for discovering trustworthy insights from aspects, considering both general and specific context.

***K*eywords** sentiment analysis cultural monuments e-tourism polarity aggregation aspect based sentiment analysis

[*]Corresponding author

# 1 Introduction

Sentiment Analysis (SA), also referred to as Opinion Mining, is a branch of Affective Computing research (52) that has experienced an important growth through the last few years due to the proliferation of the Web 2.0 and social networks. This area has been established as a new Natural Language Processing (NLP) research line which broadly processes people's opinions, reviews or thoughts about objects, companies or experiences identifying its sentiment (14; 43; 44; 51). Several teams have developed algorithms, Sentiment Analysis Methods (SAMs), capable of automatically detecting the underlying sentiment of a written review (30; 31; 45). Many companies are deploying these algorithms in order to make better decisions, understanding customers behavior or thoughts about their company or any of their products.

TripAdvisor has become a very popular e-tourism social network. It provides reviews from travelers experiences about accommodations, restaurants and attractions. In this website, users write opinions and rank their overall experience in the TripAdvisor Bubble Rating: a score ranging from 1 to 5 bubbles where 1 represents a Terrible and 5 an Excellent opinion. TripAdvisor has therefore become a rich source of data for SA research and applications (6; 49).

In past works, we shown the problem of using the TripAdvisor Bubble Rating, which we refer to as the *User Polarity* (42). This polarity represents a global evaluation of users towards a restaurant, hotel or touristic attraction, but users usually write negative sentences despite reporting 4 or 5 bubbles. In this work, we dive deeper into this problem and propose an original solution for tackling this problem. Therefore, we articulate the following research questions:

1. *"Do users usually write sentences with opposing polarities in the same opinion?"*
2. *"Is the TripAdvisor Bubble Rating a good indicator of the polarity of every sentences within an opinion?"*

We aim at answering these questions with the detection of inconsistencies between Users and SAMs polarities. SAMs are able to detect polarities of each sentence. By checking that the average of the polarities of all sentences in an opinion has a very different score from that labeled by the user, we show the presence of sentences with opposite polarities. Therefore, the TripAdvisor Bubbles Rating cannot be selected as a representation of the polarity for all sentences or aspects. We also claim that a negative aspect within a positive review should have a different score than a negative aspect within a negative review. Consequently, we propose a Polarity Aggregation Model to take into account both sentiments, the overall and the specific. This function is driven by geometric mean between User and SAM polarity which enhances the aggregation of very small values, i. e. negative polarities. It aims at obtaining a unified and robust score for facing these inconsistencies. The main contributions of this paper can be shown in the following two main aspects:

1. This model is presented as an aggregation of both expert and methods polarities, which enhance the precision of the polarity of a certain aspect in the review. We parametrized the weight of the method with a parameter $\beta$ which calibrates the contribution of that polarity.
2. We propose this model for assigning polarities to aspects. In this work, we show that our aggregation model encompass together the User and SAM polarity, which first addresses the inconsistencies problem and second, led to a better understanding of the aspect's context.

For the experimentation, we scrap the TripAdvisor website of six Italian and Spanish monuments obtaining a total of 88,882 reviews. We apply eight SAMs and study the correlations between their polarities and users ratings. Our experiments clearly show a low matching on detecting positive, neutral and negative reviews, which led us to confirm that there exists a latent inconsistency between them. We then study the behavior of the proposed polarity model taking into account its parameters, and analyze its performance on an Aspect Based Sentiment Analysis (ABSA) framework. We extract aspects and assign to them the polarities of the model. We show that aspects with very different scores between Users and SAMs obtain new polarities. Finally, we conclude that the Polarity Aggregation Model solves the inconsistency's problem and helps to extract more reliable conclusions.

The rest of this work is organized as follows: Section 2 briefly introduces the SA problem and the SAMs used for the study; Section 3 proposes TripAdvisor as our data source; Section 4 presents the results that show the inconsistencies between polarities; Section 5 proposes the Polarity Aggregation Model to face this problem and evaluates its results in an aspect extraction framework; lastly, Section 6 presents conclusions and suggests future research lines.

# 2 Sentiment Analysis

The main concepts for understanding the present work are contained in this section. Section 2.1 is a brief introduction to the SA problem. Section 2.2 presents a summary of the 8 SAMs applied in this work. Finally, in Section 2.3 we explain the algorithm for extracting aspect that we used to evaluate our model.

### 2.1 The Sentiment Analysis Problem

SA is a new research line of NLP which aims at studying people's opinion towards a product, service, organization, topic or human being in written text. The idea is to develop computational methods capable of detecting sentiments and thus extract insight to support decision makers.

Mathematically, an *opinion* can be defined as a 5-tuple (14):

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

where $e_i$ is the $i$-th opinion *entity*, $a_{ij}$ is the $j$-th *attribute*, a property related to the entity $e_i$; $s_{ijkl}$ is the *sentiment* 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* or reviewer and $t_l$ is $l$-th *time* when the opinion was emitted. Over this problem, the *sentiment* can be qualified in different ways: polarity ({positive, neutral, negative}), numerical rating ({1, 2, ..., 5} or [0, 1]) or emotions ({anger, disgust, fear, happiness, sadness, surprise}).

While most works approach it as a simple categorization problem, sentiment analysis is actually a suitcase research problem that requires tackling many NLP tasks, including subjectivity classification (47), polarity classification (25), opinion summarization (26), sarcasm detection (27), word sense disambiguation (28), opinion spam detection (29), etc. Another fact that makes this problem complex is that there exist several types of opinions (15): *regular opinions* express a sentiment about an aspect of an entity, *comparative opinions* compare two or more entities, *subjective opinions* express a personal feeling or belief and thus are more likely to present sentiments and *objective sentence* present factual information.

### 2.2 Sentiment Analysis Methods

Polarity detection has focused on the development of SAMs that can be able to detect polarity in an automatic and efficient way. These SAMs are developed to process different types of texts, from tweets (short texts containing hash-tags and emojis) to reviews (long texts talking about a movie, restaurant or hotel). In the literature we can find several studies that analyze the performance of different SAMs over multiple texts (30; 31).

Generally, these methods can be divided in three groups:

**Lexicon Dictionary Based Method:** It mainly consists of creating a sentiment lexicon, i.e., words carrying a sentiment orientation. These methods can create the dictionary from initial seed words, corpus words (related to a specific domain) or combining the two. Frequently, the dictionary is fed with synonyms and antonyms. These methods are unable to capture the underlying structure of grammar in a sentence.

**Machine Learning Based Method:** It develops statistical models with classification algorithms. These methods can be divided into supervised and unsupervised. The main difference is that the first group uses labeled opinions to build the model. One of the most important steps in these methods is the feature extraction for representing the classes to be predicted.

**Deep Learning Based Method:** Over last years Deep Learning has experienced an important growth due to its good performance in many fields of knowledge. SAMs based on neural networks learning have been shown to obtain very good results compared to other methods, discovering correlations starting from raw data. Due to the revolution of Deep Learning inside NLP and SA areas, we propose to separate it form the Machine Learning Based Methods.

Moreover, Table 1 shows a summary of all SAMs used in this work which contains references for further reading of these methods.

Table 1: Summary of the eight SAMs that we apply in our study.

| SAM | Group | Numerical Output | Reference |
|---|---|---|---|
| Afinn | LD | {-5, . . ., 5} | (36) |
| Bing | LD | {-1, 0, 1} | (11) |
| CoreNLP | DL | {0, 1, 2, 3, 4} | (21; 17) |
| MeaningCloud | ML | [0, 1] $\in \mathbb{R}$ | (41) |
| SentiStrength | LD & ML | {-1, 0, 1} | (33; 34) |
| SenticPattern+DL | DL | {0, 1, 2} | (37; 38) |
| Syuzhet | LD & ML | [0, 1] $\in \mathbb{R}$ | (13) |
| VADER | LD | [-1, 1] $\in \mathbb{R}$ | (32) |

### 2.3 Aspect Based Sentiment Analysis (ABSA)

One important fact of SA is that there exist different levels of analysis to tackle this problem. The *document level* extracts the sentiment of the whole opinion. This is considered to be the simplest task. The *sentence level* extracts a sentiment in each sentence of the text. Finally, the *aspect level* is considered the fine-grained level. This is the most challenging analysis because it extracts the entity or aspect related to the sentiment which the opinion refers to.

Over last years, the research in SA has been focusing in the aspect level (48), due to the fact that it is a more granular task and the information obtained is more detailed. Related to the extraction of aspects within an opinion, the first methods were based setting the most frequent nouns and compound nouns as aspects (10). These methods have been improved by adding syntactical relations that can enhance the task of extracting the correct aspect. However, these methods have a high number of drawback, i.e., do not detect low frequency aspects or implicit aspects, need to describe a high number of syntactical rules for detecting as many aspects as possible.

Recently, deep learning has enhanced the results of several computer science problems, and NLP is not an exception (5). Poria et al. proposed a CNN algorithm which extract aspects from reviews (40). They also used some additional features and rules to boost the accuracy of the network. The results shows that this algorithm overcome most of the state-of-the-art methods for aspect extraction.

More concretely, the network contained:

- **One input layer**. As features, they used word embeddings trained on two different corpora. They claimed that the features of an aspect term depend on its surrounding words. Thus, they used a window of 5 words around each word in a sentence, i.e., $\pm 2$ words. They formed the local features of that window and considered them to be features of the middle word. Then, the feature vector was fed to the CNN.

- **Two convolution layers**. The first convolution layer consisted of 100 feature maps with filter size 2. The second convolution layer had 50 feature maps with filter size 3. The stride in each convolution layer is 1 as they wanted to tag each word. The output of each convolution layer was computed using a non-linear function, which in this case was the $tanh$ function.

- **Two max-pools layers**. A max-pooling layer followed each convolution layer. The pool size they use in the max-pool layers was 2. They used regularization with dropout on the penultimate layer with a constraint on L2-norms of the weight vectors, with 30 epochs.

- A **fully connected layer** with *softmax* output.

In aspect term extraction, the terms can be organized as chunks and are also often surrounded by opinion terms. Hence, it is important to consider sentence structure on a whole in order to obtain additional clues. Let it be given that there are $T$ tokens in a sentence and $y$ is the tag sequence while $h_{t,i}$ is the network score for the $t$-th tag having i-th tag. We introduce $A_i$, $j$ transition score from moving tag $i$ to tag $j$. Then, the score tag for the sentence $s$ to have the tag path $y$ is defined by this formula which represents the tag path probability over all possible paths:

$$s(x, y, \theta) = \sum_{t=1}^{T} (h_{t,y_t} + A_{y_{t-1}y_t}).$$

We propose to use this model to evaluate the performance of our proposed index. We aim to analyze which polarity (User Polarity, SAM Polarity and our proposed index) obtains the most accurate score that represents the sentiment of the aspect within the opinion (See Section 5.3 and 5.4).

## 3 TripAdvisor as an Opinion Source

In this section, we describe TripAdvisor as our data source. We first give an introduction to this social network website in Section 3.1. Then, we explain how we get the data in Section 3.2. Finally, we explain the structure of the datasets in Section 3.3.

### 3.1 Why TripAdvisor?

TripAdvisor[2] is one of the most popular travel social network websites (46) founded in 2000. This Web 2.0 contains 570 million reviews about 7.3 million restaurants, hotels and attractions over the world[3]. Travelers are able to plan their trip checking information, ranking lists and experiences from others. In this website, users write reviews of minimum 100 characters and rank their experience in the TripAdvisor Bubble Rating, which is a scale from 1 to 5 points (from *Terrible* to *Excellent*). TripAdvisor are considered one of the first Web 2.0 adopters: its information and advice indices is constructed from the accumulated opinions of millions of tourists. For this reason, this website has made up the largest travel community. Due to these facts, this website has been used in the state-of-the-art of the SA (42). Examples of works analyzing hotels reviews are (1; 3; 4; 6; 7; 16; 19; 22). Restaurant reviews are analyzed in (7; 9; 24). Monument reviews are analyzed in (39; 42).

One of the major concerns of user-generated content is the credibility of the opinions. Many websites have to deal with fake or spam opinions, as their presence decreases the level of users' confidence towards their pages. Aware of it, TripAdvisor has designed several measures like verifying that customers stayed in the place their review or checking that hotels or restaurants don't review themselves. Besides that, several studies for analyzing credibility and truthfulness of this website has been carried out (2; 8; 12; 23).

### 3.2 Web Scraping

All monument pages are structured in the same way. On the top, they display the total number of reviews, written in different languages, and a *Popularity Index ranking*. After that, the page is divided in five sections: Overview, Tours&Tickets, Reviews, Q&A and Location. In the review section we find all the opinions written by users. A review is formed by:

**User Name:** The name of the user in TripAdvisor.

**User Location:** The location of the user.

**User Information:** The total number of reviews, attraction reviews and helpful votes of the user.

**Review Title:** A main title of the text.

**TripAdvisor Bubble Rating:** The writer's overall qualification of the review. It is expressed as a *bubble* scale from 1 to 5 (from *Terrible* to *Excellent*).

**Review Date:** The reviewing time.

**Review:** The text of the opinion.

Finally, we develop a code in R software with `rvest` package which allows us to extract the TripAdvisor reviews from HTML and XML sources. We analyze **User Polarity** and **Review** (see Figure 1).

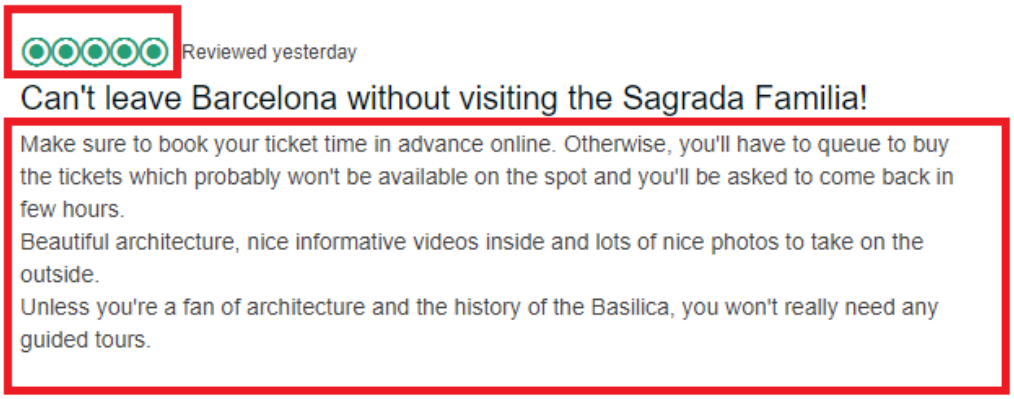



Figure 1: Information of a review in TripAdvisor. For this study, we analyze the bublle scale and the text of the review.

[2] `https://www.TripAdvisor.com`

[3] Source: `https://TripAdvisor.mediaroom.com/uk-about-us`

### 3.3 The Data

We base our experiments on TripAdvisor English reviews of three monuments in Italy (Pantheon, Trevi Fountain and Grand Canal) and other three monuments in Spain (Alhambra, Sagrada Familia and Mezquita de Córdoba). Therefore, we created six datasets with reviews from July 2012 until June 2016 and collect a total of 88,882 reviews.

Table 2: Summary of text properties of the six datasets.

| | Reviews | Words | Sentences | Avg. # words | Avg. # sentences | Avg. User Polarity |
|---|---|---|---|---|---|---|
| Alhambra | 7,217 | 676,398 | 35,867 | 93.72 | 4.97 | 4.69 |
| Grand Canal | 10,730 | 539,465 | 47,943 | 50.28 | 4.47 | 4.67 |
| Mezquita de Córdoba | 3,526 | 217,640 | 13,083 | 61.72 | 3.70 | 4.84 |
| Pantheon | 17,279 | 774,765 | 76,720 | 44.84 | 4.44 | 4.68 |
| Sagrada Familia | 34,558 | 2,220,719 | 136,181 | 64.26 | 3.94 | 4.72 |
| Trevi Fountain | 15,572 | 764,998 | 70,407 | 49.13 | 4.52 | 3.93 |

As we observe in Table 2, Sagrada Familia contains the largest number of opinions (38.88% of the total). Alhambra contains in average the longest reviews, with average words of 93.72 and average sentence of 4.97. Note that the average of the User Polarity in all datasets is very high, most of them surpass the 4.5. The best valued monument in TripAdvisor is Mezquita de Córdoba with an average rate of 4.84. Trevi Fountain is the worst valued monument with a 3.93. This is the fact that makes us wonder if in all these opinions, sentences are always positive.

## 4 A Study on the Inconsistencies between User and SAMs Polarities

TripAdvisor's opinions have been the source of data for many research works. In them, users' opinions are analyzed to extract information on what they think about a restaurant, hotel or touristic attraction. However to the best of our knowledge, it has never been analyzed the relationship between User Polarity and polarities of each sentence within the opinion. Many of the businesses that appear on the web can believe that the visitor is satisfied just by observing the average rating, but perhaps they are losing useful information by not going deeper into each opinion. We therefore believe that it is necessary to carry out a study that compares the relationship between the User Polarity and SAMs. Finally, we also think that it is interesting to focus the study on cultural monuments, since few studies in the field of SA have been carried out using them as the object of study.

In this section, we present an extended study of (42). The idea is to analyze the correlation of the User Polarity with the SAM polarities and conclude if there exist inconsistencies between them. In this work, we extend the analysis to several monuments from different countries, analyzing almost 100k reviews.

We first study the polarity label distribution of User Polarity. To do so, we label the TripAdvisor Bubble Rating of 1 and 2 bubbles as negative, 3 as neutral, and 4 and 5 as positive. We apply each of the SAMs to the whole set of opinions and scale polarities to $[0, 1]$, setting values in $[0, 0.4]$ as negative, $(0.4, 0.6)$ as neutral and $[0.6, 1]$ as positive polarity. Thereby, we get 8 polarities from 8 SAMs within the range $[0, 1]$.

We detect that the most of TripAdvisor user feedbacks are positive which means that users are satisfied with their visit (Table 3). However, this distribution is not maintained throughout SAMs. We observe that Afinn (Table 4) and MeaningCloud (Table 7) obtain a similar polarity distribution to the Users. However, Afinn does not detect any negative opinions and MeaningCloud detects 1,985 more negative reviews in Sagrada Familia dataset. Bing (Table 5), CoreNLP (Table 6) and SentiStrength (Table 8) display very different distributions: they detect many more neutral and negative reviews. Finally, Syuzhet (Table 9) and VADER (Table 10) also have a slight tendency to detect more neutral and negative opinions than users. So generally, looking at the polarity distributions between users and SAMS, we observe little similarities between them. Users have more positive and SAMs more neutral and negative opinions. This fact reflects a clear mismatching in determining the sentiment of an opinion which may be due to the different polarities that exist in sentences. It is also exposed on Figure 1 where user rates Sagrada Familia with 5 bubbles (positive opinion) but there are sentences with a negative polarity within the same opinion.

Table 3: Distribution of polarities of monuments reviews. User Polarity.

| User Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 6,781 | 293 | 143 |
| Grand Canal | 13,832 | 548 | 104 |
| Mezquita de Córdoba | 3,454 | 55 | 17 |
| Pantheon | 23,635 | 1,087 | 107 |
| Sagrada Familia | 32,664 | 1,443 | 451 |
| Trevi Fountain | 19,515 | 3,363 | 2,513 |

Table 4: Distribution of polarities of monuments reviews. Afinn.

| Afinn Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 5,395 | 1,383 | 439 |
| Grand Canal | 9,821 | 682 | 227 |
| Mezquita de Córdoba | 2,808 | 547 | 171 |
| Pantheon | 15,868 | 1,042 | 369 |
| Sagrada Familia | 31,725 | 2,833 | 0 |
| Trevi Fountain | 11,854 | 2,103 | 1,615 |

Table 5: Distribution of polarities of monuments reviews. Bing.

| Bing Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 3,310 | 1,252 | 2,655 |
| Grand Canal | 12,531 | 1,505 | 448 |
| Mezquita de Córdoba | 1,918 | 642 | 966 |
| Pantheon | 22,235 | 2,085 | 509 |
| Sagrada Familia | 16,541 | 6,644 | 11,373 |
| Trevi Fountain | 18,320 | 4,806 | 2,265 |

Table 6: Distribution of polarities of monuments reviews. CoreNLP.

| CoreNLP Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 3,154 | 1,143 | 2,920 |
| Grand Canal | 7,283 | 4,483 | 2,718 |
| Mezquita de Córdoba | 1,992 | 577 | 957 |
| Pantheon | 14,491 | 7,168 | 3,170 |
| Sagrada Familia | 17,561 | 6,007 | 10,990 |
| Trevi Fountain | 10,281 | 8,134 | 6,976 |

Table 7: Distribution of polarities of monuments reviews. MeaningCloud.

| MeaningCloud Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 6,050 | 730 | 437 |
| Grand Canal | 12,458 | 1,284 | 742 |
| Mezquita de Córdoba | 3,062 | 290 | 174 |
| Pantheon | 22,487 | 1,572 | 770 |
| Sagrada Familia | 28,124 | 3,998 | 2,436 |
| Trevi Fountain | 19,379 | 3,139 | 2,873 |

Table 8: Distribution of polarities of monuments reviews. SentiStrength.

| SentiStrength Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 5,277 | 1,341 | 599 |
| Grand Canal | 8,777 | 5,153 | 554 |
| Mezquita de Córdoba | 2,674 | 585 | 267 |
| Pantheon | 17,476 | 6,584 | 769 |
| Sagrada Familia | 23,964 | 6,880 | 3,714 |
| Trevi Fountain | 14,490 | 8,715 | 2,186 |

Table 9: Distribution of polarities of monuments reviews. Syuzhet.

| Syuzhet Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 5,423 | 1,252 | 2,655 |
| Grand Canal | 13,000 | 1,176 | 308 |
| Mezquita de Córdoba | 2,704 | 466 | 356 |
| Pantheon | 22,925 | 1,601 | 303 |
| Sagrada Familia | 25,379 | 4,805 | 4,374 |
| Trevi Fountain | 19,722 | 4,211 | 1,458 |

Table 10: Distribution of polarities of monuments reviews. VADER.

| VADER Polarity | Positive | Neutral | Negative |
|---|---|---|---|
| Alhambra | 6,505 | 362 | 350 |
| Grand Canal | 13,368 | 753 | 363 |
| Mezquita de Córdoba | 3,206 | 200 | 120 |
| Pantheon | 23,319 | 1,042 | 468 |
| Sagrada Familia | 30,485 | 2,450 | 1,623 |
| Trevi Fountain | 20,979 | 2,093 | 2,319 |

Figure 2 shows the matching ratio between User and SAMs polarities: each row of the matrix represents the classified polarities by users while each column represents the classified polarities by each SAMs. In order to optimize the layout (8 SAMs $\times$ 6 monuments = 48 matrices), we display the average rates over the six monuments. This is justified since the distribution on the six tables are very close (the maximum standard deviation of all monuments is 0.176).


Figure 2: Percentage of matching between Users (rows) and SAMs (columns) polarities. The values are the average over the six monuments. A more orangeade color on cells indicates higher correlation, bluer lower correlation.

SAMs have an acceptable performance detecting positivity as orange tones predominate in almost all positive-positive cells. On the other hand, bluish tones are the most predominant on neutral-neutral and negative-negative cells, indicating a low correlation ratio. VADER is the one that best qualifies positive user reviews (92.10 %) and CoreNLP the worst one (47.60 %). This one obtains better results detecting negative user reviews (68.10 %) but all others get poor results (ratios beneath 38 %). Most of them tend to classify them as positive. Neutrality is the polarity which shows the worst outcomes. There is no SAM standing out on detecting this middle polarity (39).

As can be hinted from Figure 2, data reveals a clear disparity between users and SAMs polarities. We show that there is a low level of matchings when detecting polarities. Analyzing text data we discover that users may tend to write negative sentences on positive reviews, and vice versa. Therefore, we should recommend not to set users polarity as the overall sentiment of their reviews because otherwise, we will be missing a lot of information.

## 5 A Polarity Aggregation Model for Reviews: Calibrating the Polarity Between Users and SAMs

In this section, we propose a solution to address the problem of inconsistencies. As we shown in last section, the correlation of polarities between Users and SAMs is low. This is mainly driven by the fact that users tend to write negative sentences in positive opinions and vice versa. Therefore, we propose a model (Polarity Aggregation Model) which aggregates both polarities and straddles the general context of the opinion (User Polarity) with the specific context (SAM Polarity) (Section 5.1). Then, we propose to test our model with TripAdvisor's reviews from the Alhambra and the Pantheon monuments (Section 5.2).

After that, we develop an analysis to show how our model behaves within an aspect scenario. Firstly, we study the performance of our model assigning scores on aspects that are extracted with the algorithm presented in previous Section 2.3 (Section 5.3). Secondly, we present a most detailed analysis within this scenario, reporting two aspects in particular (Section 5.4).

### 5.1 The Polarity Aggregation Model

In Section 4, we show that there is a low correlation between User and SAMs polarities. We discuss that users tend to rank their visit with high punctuations, which connotes a positive sentiment. However, users do not usually use positive sentiment in every sentence, which leads to SAMs detecting more neutral or negative polarities.

In order to tackle this problem, we create a new polarity index that takes into account both user and SAMs for overcoming the inconsistency problem. For this reason, we propose an aggregation model guided by the geometrical mean, a variant including a parameter to control one variable influence. This type of mean indicates the central tendency by using the product of their values and it is defined as the $n$th root of the product $n$ numbers[4]. It is often used when the numbers have very different properties. One of the main properties of this mean is that it strengthens values close to 0, for example, the arithmetic mean between 0 and 1 is 0.5 but the geometric mean is 0. This function is expressed as follows:

$$f(x, y) = \sqrt{xy^{\beta}}$$

where:

- $x = \frac{p_i^{USER} - min(\{p_1^{USER}, ..., p_N^{USER}\})}{max(\{p_1^{USER}, ..., p_N^{USER}\}) - min(\{p_1^{USER}, ..., p_N^{USER}\})}$ is the *Normalized User Polarity* of the $i$th-opinion and $x \in [0, 1]$.
- $y = \frac{p_i^{SAM_k} - min(\{p_1^{SAM_k}, ..., p_N^{SAM_k}\})}{max(\{p_1^{SAM_k}, ..., p_N^{SAM_k}\}) - min(\{p_1^{SAM_k}, ..., p_N^{SAM_k}\})}$ is the $k$th-*Normalized SAM Polarity* of the $i$th-opinion and $y \in [0, 1]$.
- $\beta$, is the parameter to control the SAMs polarity influence and $\beta \in \mathbb{R}^+$.
- $p_i^{USER}$ is the User Polarity of the $i$th-opinion.
- $p_i^{SAM_k}$ is the k$th$-SAM Polarity of the $i$th-opinion.

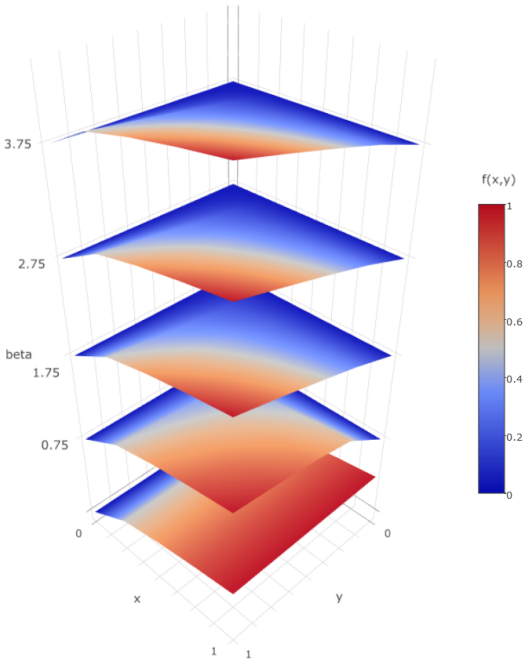



Figure 3: Distribution of the Polarity Aggregation Model for different $\beta$ values (0, 0.75, 1.75, 2.75 and 3.75). Bluer colors represent more negative aggregated polarities, more orange colors more positive aggregated polarities.

In Figure 3, we present the behavior of that function. In this 3D figure, the Normalized User Polarity ($x$) is represented on x-axis, the Normalized CoreNLP Polarity ($y$) on the y-axis and $\beta$ parameter on the z-axis for a certain set of values. As we can observe, the surface that shows the distribution of polarities for small values of $\beta$ contained more red, which means that it gets more positive scores. As we increase the value of $\beta$, surfaces contains more blues, which means that the function obtains more negative scores. This Figure clearly shows how can we adjust the distribution of the scores, setting the $\beta$ parameter.

More concretely, this function works as follows:

- If $\beta < 1 \implies f(x, y) > \sqrt{xy}$. In that case, we observe that for $\beta = 0$ (see the bottom surface) most scores are close to 1 (red colors) because $\sqrt{y^{\beta}}$ is always 1. Then, $\sqrt{x}$ rules the final value of the function obtaining more positive scores. The negative scores are only obtained with small values of $x$. If we increase the value of that parameter, we obtain more negative values for small values of $x$ and $y$ (see the second surface where $\beta = 0.75$), but the positive polarities still predominate.

[4]Source: `https://en.wikipedia.org/wiki/Geometric_mean`

- If $\beta \geq 1 \implies f(x,y) \leq \sqrt{x}y$. In that case, the value of $y$ gains relevance in the final score. If we observe the top surfaces on Figure 3, final negative polarities (blue colors) are obtained with a wide range of $y$ values. As we increase the value of $\beta$, more negative scores are obtained. In fact, the blue strip on the y-axis gains ground as we increase that parameter. Hence, we are able to model the function for obtaining pro-positive or pro-negative polarities setting parameter $\beta$

Once we have show the behavior of the Polarity Aggregation Model taking account the value of User and SAM Polarities, we seek to analyze how it behaves with real values. For that, in next section we present the values of the proposed model taking into account the polarities of the User and CoreNLP in the datasets of the Alhambra and Pantheon.

### 5.2 A case study on the datasets of the Alhambra and the Pantheon

We analyze the behavior of the Polarity Aggregation Model (with CoreNLP as the selected SAM) on reviews of the Alhambra and Pantheon datasets. Figure 4 shows the relationship between this SAM, the User Polarity and the Polarity Aggregation Model. The instances are ordered along the x axis, taking into account the Normalized User Polarity Rating, from the most positive to the most negative. We select different $\beta$ values between 0 and 4. We observe that when $\beta \in [0, 1]$, the polarity trend of the model is between the User and CoreNLP. When $\beta \geq 1$, its polarity score tend to be more negative, under the CoreNLP line.

**Figure 4 (top)**: In the Alhambra's dataset, from the 1st to the 5,660-th instance the value of the Normalized User Polarity is always 1 (positive), but on the other hand, CoreNLP values are decreasing to 0 (negative). Then we observe that when the User values go to 0.75 (still positive), CoreNLP goes up to positive values and then decreases to negative values again. At negative User values, CoreNLP detects some reviews as positive.

**Figure 4 (bottom)**: In the Pantheon's dataset we observe a similar behavior, although CoreNLP decreases more slowly. In the previous case, CoreNLP goes from positive to neutral before the 2,000-th row, in this case, after the 7,500-th row. We also observe that the behavior of the CoreNLP trend is more staggered than in the Alhambra.

Table 11: Mean of CoreNLP Polarity taking account the User Polarity.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Alhambra | 0.321 | 0.348 | 0.393 | 0.475 | 0.534 |
| Pantheon | 0.389 | 0.356 | 0.477 | 0.583 | 0.597 |

For the positive User Polarity range, CoreNLP decreases faster on the Alhambra's dataset. This can be observed also in Table 11, where the CoreNLP mean on this range is lower (4 and 5 bubbles). On the neutral range (3 bubbles), CoreNLP decreases very fast on the Pantehon's dataset and there are more values above 0.5, which is reflected on its mean (0.477). On the negative range (1 and 2 bubbles) both CoreNLP Polarity plots jumps, which means that this SAM detects positive and neutral polarities in opinions labeled negative by the user.

We study the behavior of $\beta$ also in Table 12. For low $\beta$ values (0.25, 0.75, 1), the Polarity Aggregation Model obtains higher average scores (more positive), refolding the trend of the User Polarity. For higher values (2, 3), the model obtains lower average scores (more negative), refolding the trend of the CoreNLP Polarity. In fact, for reviews scored as positive (4 and 5 bubbles) this model obtains neutral and even negative scores. This fact was also reflected in Figure 3.

Table 12: Mean of the Polarity Aggregation Model taking account the User Polarity.

| | 1 | 2 | 3 | 4 | 5 | beta |
|---|---|---|---|---|---|---|
| Alhambra | 0 | 0.437 | 0.620 | 0.781 | 0.919 | 0.25 |
| | 0 | 0.334 | 0.490 | 0.645 | 0.782 | 0.75 |
| | 0 | 0.292 | 0.436 | 0.588 | 0.723 | 1 |
| | 0 | 0.174 | 0.278 | 0.411 | 0.534 | 2 |
| | 0 | 0.105 | 0.181 | 0.294 | 0.403 | 3 |
| Pantheon | 0 | 0.436 | 0.637 | 0.802 | 0.930 | 0.25 |
| | 0 | 0.333 | 0.523 | 0.695 | 0.810 | 0.75 |
| | 0 | 0.292 | 0.476 | 0.649 | 0.759 | 1 |
| | 0 | 0.178 | 0.338 | 0.505 | 0.597 | 2 |
| | 0 | 0.113 | 0.250 | 0.405 | 0.483 | 3 |

Finally we point out that the inconsistencies between both polarities are evident. We also conclude that the Polarity Aggregation Model clearly averages the two polarities when $\beta \in [0, 1]$. Thus, this new aggregation model can be useful for reassessing review sentiments across different monuments.

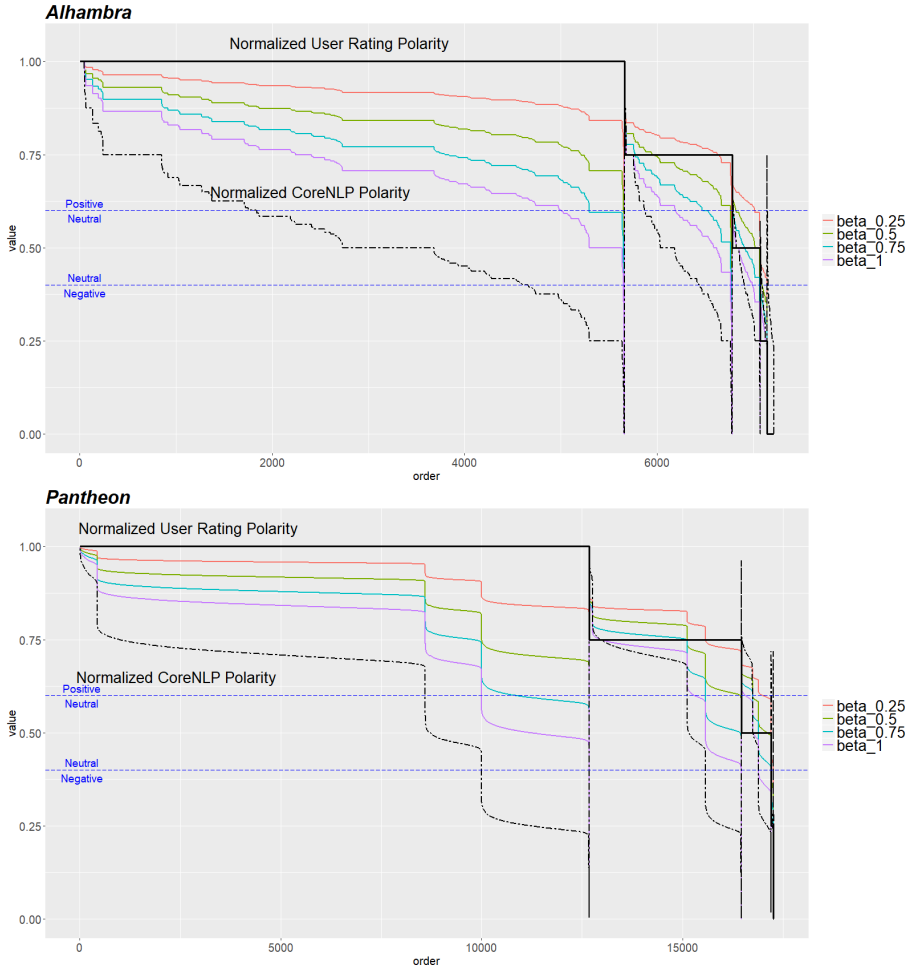



Figure 4: Different Polarity Aggregation Models taking account beta's values. Reviews are sorted on the x label in ascending order, from most positive (left) to most negative (right). The thick line represents the Normalized User Polarity. The two-dash line represents the Normalized CoreNLP Polarities.

### 5.3 An Aspect Analysis on the Three Polarities: User, CoreNLP and Polarity Aggregation Model

The aim of this study is to analyze polarities (User, CoreNLP and Polarity Aggregation Model) on ABSA framework. The idea is to study the inconsistencies on the extracted aspects and find out if they actually occur in sentences with a different polarity to the overall. We will then study whether the Polarity Aggregation Model helps to solve the problem. For this, we extract aspects with a deep learning approach developed by Poria et al. in (40). We then compute the average polarity of User, CoreNLP and the Polarity Aggregation Model for each aspect. For the model, we select $\beta = 0.75$ because it is the value which obtains polarity scores in between users and CoreNLP (see Figure 4). We base these experiments on one monument from Spain and other from Italy: the Alhambra and the Pantheon.

Our first analysis aims at studying the polarities incoherences on aspects extracted. The idea is to find and analyze those aspects that have a very positive User Polarity and very negative CoreNLP Polarity or vice versa. Figure 5 shows the relationship between User and CoreNLP Polarity on Alhambra's and Pantheon's aspects appearing at least twice.

> **Figure 5 (top)**: As we can observe, *Alhambra* is the aspect that most often appears (it is the one on the far right). Although this aspect has a Normalized User Polarity of 0.9 (positive), its color reveals that CoreNLP

only gives it a 0.47 (neutral). It is interesting to note that aspects such as *ticket* or *queue* also appear with a very high User's polarity (from 0.90 and 0.84 respectively). However, its CoreNLP's polarity is 0.43 and 0.39, which once again reveals the low correlation between the two polarities. Dipping into Alhambra's opinions in which some of these two aspects appear, we have discovered that users usually rate their visit to this monument with a good score (4 and even 5 bubbles), but in their text they complain about the long queues at the time of entering or the bad management of the ticket system that the Alhambra has, which makes CoreNLP get a lower score for those set of opinions.

**Figure 5 (bottom)**: Although this monument has 10,062 opinions more than the Alhambra, the number of aspects extracted is very similar. *Pantheon* and *architecture* are the most frequent aspects. For the aspect *noise*, CoreNLP is 0.5 (neutral) while Users obtains a mean of 0.85 (positive). The aspect *queuing system* obtains a value of 0.23 (negative) for CoreNLP and 1 (positive) for Users. Analyzing text opinions we come to the same conclusion as in the previous case: users often complain about some aspect of the monument like the noise, but rank their visit positively. We also detect that aspect *selfies* has a very low score due to the fact that reviewers complain because there are many people taking self-portraits around the monument.

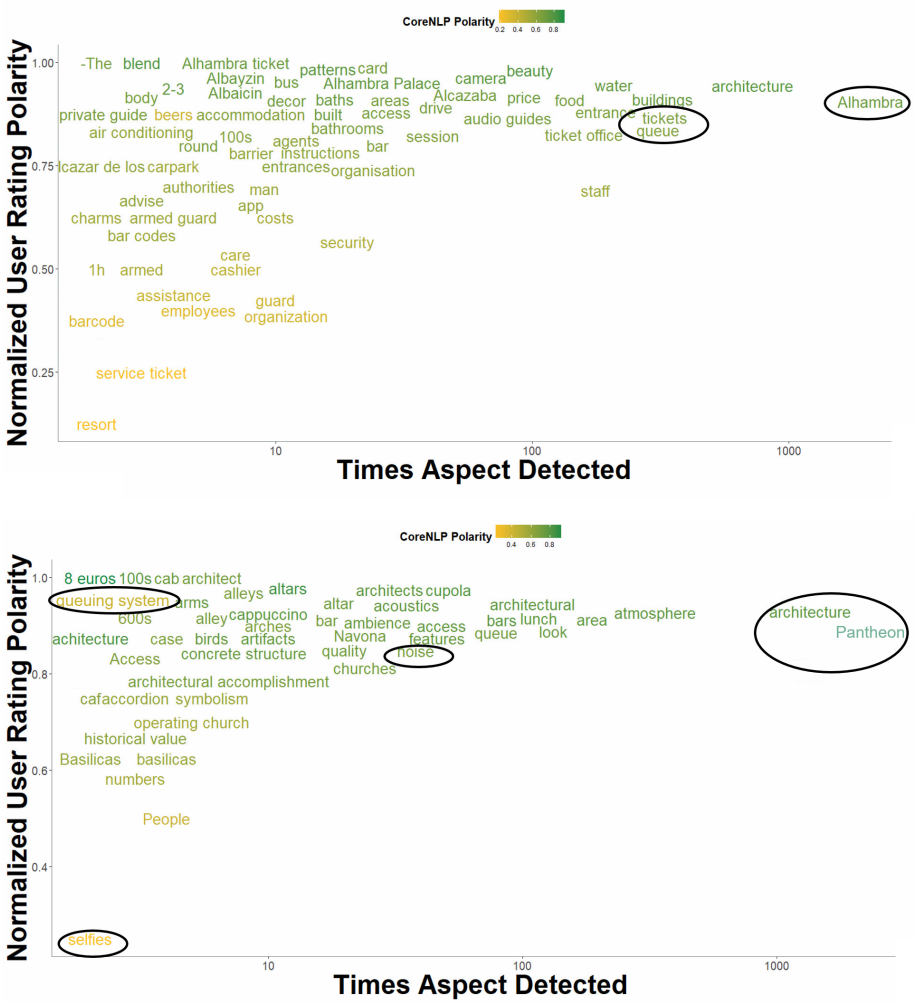



Figure 5: This aspect map represents times that an aspect is detected (x axis) taking User Polarity (y axis) and CoreNLP Polarity (color scale). Alhambra (top) and Pantheon (bottom).

We then aim at studying if the Polarity Aggregation Model fixes inconsistencies on the polarity of aspects. We analyze the polarity values of the three polarities for every aspect. For this, we set an experiment similar to the previous one.

However, in this case, Figure 6 shows the extracted aspects taking into account the three averaged polarities (User, CoreNLP and Polarity Aggregation Model).

We observe in those cases that the proposed Aggregation Model works well for detecting negative aspects in positive reviews. This is due to the property that we have previously mentioned of the geometric mean which penalizes very high values.

**Figure 6 (top)**: We note that the highest density of aspects are found on the right side of the image, i.e. when the Normalized User Polarity is positive (between 0.6 and 1). In this area, there are aspects which have a positive polarity with User, CoreNLP and so Polarity Aggregation Model: *Arabic design*, *forest*, *Alhambra Palace*, *architecture*. We also find other aspects in which CoreNLP detects a totally negative polarity, such as *sale* or *distances*. We have detected with *time frame* users complains about the time schedules of tickets for visiting the monument and with *distances* aspect that the reviewers warn of long distances to reach the Alhambra. In those aspects, CoreNLP gives 0.23 and 0.13 and Users 1 and 0.87, respectively which led the Polarity Aggregation Model obtains 0.37 and 0.26.

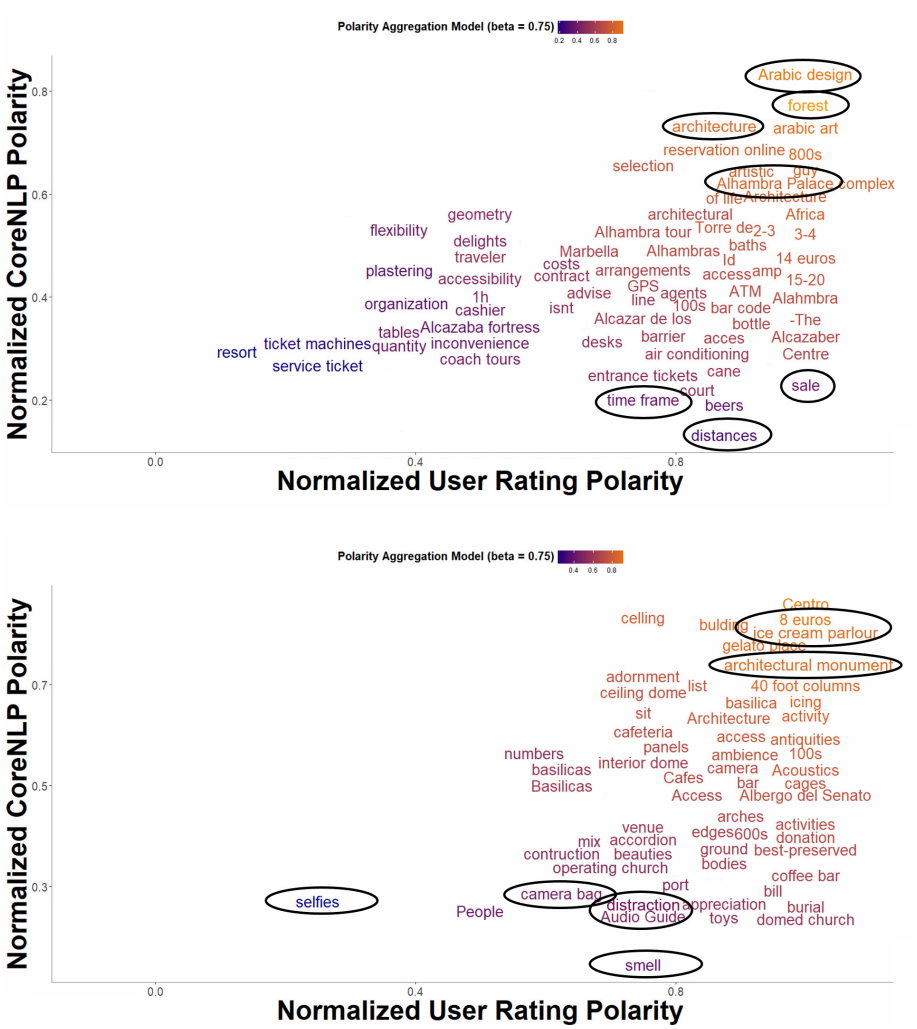



Figure 6: This aspects map represents the mean of each polarity of each aspect. From left to right it goes from negative to more positive depending on the User Polarity. From bottom to top goes from negative to more positive depending on CoreNLP Polarity. From more blue to more orange goes from more negative to more positive depending on the Polarity Aggregation Model, with $\beta = 0.75$. Alhambra (top) and Pantheon (bottom).

**Figure 6 (bottom)**: In this case, fewer negative aspects appear. We detect very positive aspects like: *8 euros*, *architectural monument*, *ice cream parlour*. The first aspect reflects the fact that visitor recommend the audio

guides. The second one refers to the Pantheon. Finally, reviewers highly recommend to rest next to the monument and buy an ice cream there. We detect other aspects (*smell*, *distraction*, *camera bag*) in which CoreNLP Polarity is very negative, User Polarity is very positive and so the Polarity Aggregation Model obtains a very negative score, penalizing the positive punctuation of the User Polarity. In those cases, users complain about unpleasant odors, distractions caused by clamor and thefts.

In view of the results, we conclude that:

**Inconsistencies.** In Figure 5 we detect, on both monuments, that there exist aspects with very different polarities between User and CoreNLP. This map of word reflects again inconsistencies and we show that wrong conclusions can be drawn on an aspect framework.

**Polarity Aggregation Model fixes inconsistencies.** Figure 6 depicts that those dismatchings between Users and SAMs are fixed with the Polarity Aggregation Model. Those aspects that obtain very different polarities end up getting averaging scores which led to obtain more reliable conclusions. We then show that our model is an effective approach to deal with the raised problem, taking the context of the overall sentiment, i.e, the User Polarity.

**Polarity Aggregation Model for discovering trustworthy insights.** In SA, aspects are analyzed for extracting knowledge. In this task, it is essential to define their relevant polarity. If we analyze TripAdvisor reviews and assign to their aspects the User Rating Polarity, we may be assigning wrong polarities to them. However, as it is depicted in this section with several aspects, the Polarity Aggregation Model solves this problem by taking into account both User and CoreNLP scores.

### 5.4 An example of the performance of our model within opinions

In this section we present a more detailed analysis the performance of our model by analyzing the whole text of the opinion, setting the parameter $\beta$ of our model equals to 0.75. To do so, we select for each monument (Alhambra and Pantheon) an aspect that appears in Figure 6 and study the accuracy of the three polarities (User, SAM and our model) regarding the text.

- ***time frame (Alhambra):*** As we presented in Section 5.3, the aspect *time frame* appears in reviews where users report positive polarities, but CoreNLP detects negativity (see Figure 6). If we analyze some opinions where this aspect appears (see Table 13), we observe that our proposed model gathers the overall and specific context of the aspect within an opinion. In the first one, the user reports a positive score (User Polarity = 1), but in the second one, the other user reports a neutral one (User Polarity = 0.5). On the other hand, CoreNLP detects that the second opinion is much more negative than the first one. Reading both opinions, we figure out that the first user uses the aspect *time frame* for warning other visitors, but the underlying sentiment is not completely negative. On the second opinion, the sentiment of the user is very negative, he or she expresses frustration towards that aspect of the visit. Therefore, if we analyze the scores obtained by our index, we observe that it gives 0.71 points to the first opinion and 0 to the second one. These scores represent both the context of the overall opinion, which in the first one is positivism and the second one is neutrality and frustration, and the specific context of the aspect, which in both cases in negative.

- ***audio guide (Pantheon):*** The aspect *audio guide* appears also in reviews where the sentiment of the user is positive, but CoreNLP detects negativity. As we can observe in Table 14, in both examples the user expresses a positive polarity (1 and 0.75 which corresponds to 5 and 4 bubbles in the TripAdvisor site), but CoreNLP detects in the first case a positive polarity (0.93) and in the second case a negative polarity (0.25). Reading the text of both reviews, we observe that the first user shows a positive polarity to the aspect, so our score obtains 0.97 points. On the second example, the user shows a negative review towards the aspect, but the overall context of the opinion, as we have explained, is positive. Therefore, our model obtains a score in between positivism and negativism, which clearly represents the situation of the aspect within this opinion.

## 6 Conclusions and Future Work

This work presented a problem related to the TripAdvisor Bubble Rating which, to the best of our knowledge, has never been raised before. We showed that users tend to evaluate positively the overall experience but there exist sentences with an opposite polarity. Hence, this rating cannot be representative for all sentences. In order to show this fact, we formulated our hypothesis and analyzed the polarity matching between User Polarity and eight SAMs. We showed that there exists a low correlation between them on detecting polarities. We also explained that the average of matching on

Table 13: Example of our model performance with the aspect *time frame* in two reviews of the Alhambra.

| Aspect | Text | User | CoreNLP | $\beta = 0.75$ |
|---|---|---|---|---|
| time frame | This place is amazing and should not be missed, no need to add to the thousands other good reviews written hear. I would like to write about my purchasing experience to possibly help someone out in getting this done the easiest way. Trying to get a ticket to see the Alhambra is a project you kind of have to study to know how to do so. I understand why many find it confusing and end up not getting it right. I can only recommend doing it the way I did, as it was simple as 1-2-3: 1. Go to Ticketmaster.es (the Spanish site) and search for tickets for the Alhambra. We got the cheapest best value ones- 15 euro for the general entrance, 2. Purchase tickets to either morning session (ends at 14) or afternoon session (starts at 14 ends at 18/20 depending on season). Know that you are allowed to be at the grounds within that **time frame** but that would be forced to exit, or not allowed in before/after your session. 3. Know that the specific time selected for your ticket indicates a 30 minute window for you to enter the Nasarid palace (but you can tour the rest of the grounds before or/and after visiting the palace) [...]. | 1 | 0.40 | 0.71 |
| time frame | I tried to book a ticket for this place month in advance and my credit card was declined all the time. Even called the local ticket office and they couldn't help, so in desperation asked the hotel I stayed to try to get tickets-well. I think what they try to do is to discourage you to buy the 'cheap' 14 euro ticket and pay 35 or 50 euros for a guided tour-since you have to book a **time frame**. We thought that it will give you space to move around-certainly. It's not-hundreds of people lining up at every corner and rooms, so it's grossly overcrowded. | 0.5 | 0 | 0 |

Table 14: Example of our model performance with the aspect *audio guide* in two reviews of the Pantheon.

| Aspect | Text | User | CoreNLP | $\beta = 0.75$ |
|---|---|---|---|---|
| audio guide | Well worth a visit! Definitely worth a visit!We got the **audio guide** which is worth doing especially to learn how they built the Pantheon it self! | 1 | 0.93 | 0.97 |
| audio guide | Literally just to see it!! The **audio guide** witch is 5 euros is not worth it. Unless you want to hear about the dome because everything else you can just read. I steped inside to see it and walked in and out in less than 30 min. | 0.75 | 0.25 | 0.51 |

detecting three polarities (positive, neutral and negative) is over 47%. This is because, as we explained, humans do not use the same sentiment in every sentence, but rather people tend to change, and SAMs are able to detect those changes.

In order to address this problem, we proposed the Polarity Aggregation Model. We presented this model as a unified index of two polarities. This model is guided by the geometric mean function of the polarity of the User and a SAM. The weight of the SAM polarity can be set by a parameter, $\beta$. This parameter can take positive values, although we showed that values above 1 get too negative aggregated polarities. The proposed model, with $\beta = 0.75$, obtained robust results and fixed the mismatch between humans and SAMs polarities. In an aspect analysis framework, the Polarity Aggregation Model helps drawing more accurate conclusions, since we observed how it helps to adjust polarities on extracted aspects.

The main advantage of our proposal is that the Polarity Aggregation Model obtains more trustworthy scores absorbing information from two sources: users and algorithms for automatic detection of sentiments. This averaging model fixes the inconsistencies presented when defining the polarity of a TripAdvisor review. It also detects and assigns different scores to negative aspects within positive reviews and vice versa. We showed in several aspects analysis that the insights extracted by this polarity are more corresponding to user's review.

There are several directions highlighted by our results. We studied the behavior of the model with only one parameter. We propose to carry out a study enriching our model by adding another parameter to the User Polarity. Our model has also shown an effective behavior by combining the value of users and SAMs into an ABSA scenario. However, the extraction of those aspects can be improved. We detect that different extracted aspects refers to the same object, so the output should be refined with pre processing methods and text mining techniques. These aspect representations can be also extended to bigrams or unigram+bigrams. Finally, we propose to extract more valuable insights through relational

models based on association rules or machine learning techniques within this framework. A concurrency analysis at aspect level on social network can be used to enrich the extraction of insights.

## Acknowledgment

This work is supported by the Spanish National Research Project TIN2017-89517-P. This work is partially supported by the Data Science and Artificial Intelligence Center (DSAIR) at the Nanyang Technological University.

## References

[1] Aciar, S. Mining context information from consumers reviews. In Proceedings of Workshop on Context-Aware Recommender System, ACM, 201(0), (2010).

[2] Ayeh, J. K., Au, N., and Law, R. Do we believe in TripAdvisor? Examining credibility perceptions and online travelers' attitude toward using user-generated content. Journal of Travel Research, 52(4), 437-452, (2013).

[3] Baccianella, S., Esuli, A., and Sebastiani, F. Multi-facet rating of product reviews. In European Conference on Information Retrieval in Springer Berlin Heidelberg, 461-472, (2009).

[4] Banic, L., Mihanovic, A., and Brakus, M. Using big data and sentiment analysis in product evaluation. In Information and Communication Technology Electronics and Microelectronics, 36th International Convention on IEEE, 1149-1154, (2013).

[5] Collobert, Ronan, and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th international conference on Machine learning. Association for Computing Machinery, 2008.

[6] Duan, W., Cao, Q., Yu, Y., and Levy, S. Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In System Sciences (HICSS), 2013 46th Hawaii International Conference on IEEE, 3119-3128, (2013).

[7] ElSahar, H., and El-Beltagy, S. R. Building large arabic multi-domain resources for sentiment analysis. In International Conference on Intelligent Text Processing and Computational Linguistics in Springer International Publishing, 23-34, (2015).

[8] Filieri, R., Alguezaui, S., and McLeay, F. Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. Tourism Management, 51, 174-185, (2015).

[9] García, A., Gaines, S., and Linaza, M. T. A lexicon based sentiment analysis retrieval system for tourism domain. Expert Syst Appl Int J, 39(10), 9166-9180, (2012).

[10] Hu, M., and Liu, B. Mining opinion features in customer reviews. In the American Association on Artificial Intelligence Conference on Artificial Intelligence, Vol. 4, No. 4, 755-760, (2004).

[11] Hu, M., and Liu, B. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177, (2004).

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

[13] Jockers, M. Syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text. R package version 1.0.0, (2016).

[14] Liu, B. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, (2015).

[15] Liu, B. Sentiment Analysis and Subjectivity. Handbook of natural language processing, 2, 627-666, (2010).

[16] Lu, B., Ott, M., Cardie, C., and Tsou, B. K. Multi-aspect sentiment analysis with topic models. In Data Mining Workshops (ICDMW), 11th International Conference on IEEE, 81-88, (2011).

[17] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In the Association for Computational Linguistics (System Demonstrations), 55-60, (2014).

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

[19] Popescu, A. M., and Etzioni, O. Extracting product features and opinions from reviews. In Natural language processing and text mining, Springer London, 9-28, (2007).

[20] Pozzi, F. A. and Fersini, E. and Messina, E. and Liu, B. Sentiment Analysis in Social Networks. Morgan Kaufmann, (2016).

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

[22] Titov, I. and McDonald, R. T. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In Association for Computational Linguistic, 8, 308-316, (2008).

[23] Yoo, K. H., Lee, Y., Gretzel, U., and Fesenmaier, D. R. Trust in travel-related consumer generated media. In Information and communication technologies in tourism, 49-59, (2009).

[24] Zhang, H. Y., Ji, P., Wang, J. Q., and Chen, X. H. A novel decision support model for satisfactory restaurants utilizing social information: a case study of TripAdvisor. com. Tourism Management, 59, 281-297, (2017).

[25] Pang, B., Lillian L., and Shivakumar V. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, (2002).

[26] Ku, L. and Liang, Y. and Chen, H. Opinion extraction, summarization and tracking in news and blog corpora. In Proceedings of American Association on Artificial Intelligence, 100-107, (2006)

[27] Sulis, E., Farias, D. I. H., Rosso, P., Patti, V., and Ruffo, G. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. Knowledge-Based Systems, 108, 132-143, (2016).

[28] Kågebäck, M. and Salomonsson, H. Word Sense Disambiguation using a Bidirectional LSTM. arXiv preprint arXiv:1606.03568, (2016).

[29] Ren, Y., and Ji, D. Neural networks for deceptive opinion spam detection: An empirical study. Information Sciences, 385, 213-224, (2017).

[30] Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1), 1-29, (2016).

[31] Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., and Herrera-Viedma, E. Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18-38, (2015).

[32] Hutto C.J., and Gilbert, E. VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International AAAI Conference on Weblogs and Social Media, (2014).

[33] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. Journal of the Association for Information Science and Technology, 61(12), 2544-2558, (2010).

[34] Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment strength detection for the social web. Journal of the Association for Information Science and Technology, 63(1), 163-173, (2012).

[35] Bradley, M. M., and Lang, P. J. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report C-1, the center for research in psychophysiology, University of Florida, 1-45, (1999).

[36] Nielsen, F. Å. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903, (2011).

[37] Poria, S., Cambria, E., Gelbukh, A., Bisio, F., and Hussain, A. Sentiment data flow analysis by means of dynamic linguistic patterns. IEEE Computational Intelligence Magazine, 10(4), 26-36, (2015).

[38] Chaturvedi, I., Ong, Y. S., Tsang, I. W., Welsch, R. E., and Cambria, E. Learning word dependencies in text by means of a deep recurrent belief network. Knowledge-Based Systems, 108, 144-154, (2016).

[39] Valdivia, A., Luzón, M. V., and Herrera, F. Neutrality in the sentiment analysis problem based on fuzzy majority. In proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1-6, (2017).

[40] Poria, S., Cambria, E., and Gelbukh, A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems, 108, 42-49, (2016).

[41] MeaningCloud – Opinion Mining API. `https://www.meaningcloud.com/products/sentiment-analysis`. Online; accessed Jan 2017, (2017).

[42] Valdivia, A., Luzón, M. V., and Herrera, F. Sentiment Analysis in TripAdvisor. IEEE Intelligent Systems, 32(4), 72-77, (2017).

[43] Balazs, J. A., and Velásquez, J. D. Opinion mining and information fusion: a survey. Information Fusion, 27, 95-110, (2016).

[44] Liu, B., and Zhang, L. A survey of opinion mining and sentiment analysis. In Mining text data. Springer US, (2012).

[45] Sun, S., Luo, C., and Chen, J. A review of natural language processing techniques for opinion mining systems. Information Fusion, 36, 10-25, (2017).

[46] O'Connor, P. User-generated content and travel: A case study on TripAdvisor. com. Information and communication technologies in tourism 2008, 47-58, (2008).

[47] Lu, B., and Tsou, B. K. Combining a large sentiment lexicon and machine learning for subjectivity classification. In 2010 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, 3311-3316, (2010).

[48] Schouten, K., and Frasincar, F. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3), 813-830, (2016).

[49] Marrese-Taylor, E., Velásquez, J. D., and Bravo-Marquez, F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications, 41(17), 7764-7775, (2014).

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

[51] Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. A Practical Guide to Sentiment Analysis. Cham, Switzerland: Springer, ISBN: 978-3-319-55394-8 (2017)

[52] Poria, S., Cambria, E., Bajpai, R., and Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37, pp. 98-125 (2017)

[53] Ma, Y., Peng, H., and Cambria, E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: AAAI (2018)

[54] Cambria, E., Poria, S., Hazarika, D., and Kwok, K. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings, pp. 2666-2677, AAAI (2018)

# 3 Consensus Vote Models for Detecting and Filtering Neutrality in Sentiment Analysis

- A. Valdivia, MV. Luzón, E. Cambria, F. Herrera. Consensus vote models for detecting and filtering neutrality in sentiment analysis. Information Fusion 44 126-135 (2018)
  - Status: **Published**.
  - Impact Factor (JCR 2017): **6.639**
  - Subject Category: Computer Science, Artificial Intelligence
  - Rank: **8/132**
  - Quartile: **Q1**

# CONSENSUS VOTE MODELS FOR DETECTING AND FILTERING NEUTRALITY IN SENTIMENT ANALYSIS

A PREPRINT

**Ana Valdivia***
Department of Computer Science
and Artificial Intelligence
Universidad de Granada
Granada, Spain 18014
avaldivia@ugr.es

**M. Victora Luzón**
Department of Software Engineering
Universidad de Granada
Granada, Spain 18014
luzon@ugr.es

**Erik Cambria**
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore 639798
cambria@ntu.edu.sg

**Francisco Herrera**
Department of Computer Science
and Artificial Intelligence
Universidad de Granada
Granada, Spain 18014
herrera@decsai.ugr.es

**ABSTRACT**

Recently, interest in sentiment analysis has grown exponentially. Many studies have developed a wide variety of algorithms capable of classifying texts according to the sentiment conveyed in them. Such sentiment is usually expressed as positive, neutral or negative. However, neutral reviews are often ignored in many sentiment analysis problems because of their ambiguity and lack of information. In this paper, we propose to empower neutrality by characterizing the boundary between positive and negative reviews, with the goal of improving the model's performance. We apply different sentiment analysis methods to different corpora extracting their sentiment and, hence, detecting neutral reviews by consensus to filter them, i.e., taking into account different models based on weighted aggregation. We finally compare classification performance on single and aggregated models. The results clearly show that aggregation methods outperform single models in most cases, which led us to conclude that neutrality is key for distinguishing between positive and negative and, then, for improving sentiment classification.

***K*eywords** Sentiment Analysis · Neutrality · Fuzzy Theory

## 1 Introduction

Over the past few decades, the amount of social media data (e.g., reviews, opinions or posts) stored in the Web 2.0 has grown exponentially. This type of website consists of social media platforms (e.g., Blogger and TripAdvisor), social networks (e.g., Facebook and Twitter) and photo, audio or video portal hosting (e.g., Instagram and YouTube). The essence of such tools is the possibility to interact with other users or provide content that enriches the browsing experience.

Sentiment analysis has emerged as a new tool for analyzing Web 2.0 information Cambria et al. (2017); Liu (2015); Balazs and Velásquez (2016); Sun et al. (2017); Bello-Orgaz et al. (2016). It is a branch of affective computing research Poria et al. (2017a) that aims to classify text (but sometimes also audio and video Poria et al. (2017b)) as either positive or negative.

*Corresponding author

The main aim of sentiment analysis is to systematically analyze people's opinion on a product, organization, or event Liu (2015). Hence, its most important goal is Sentiment Analysis Classification (SAC), i.e., determining whether an opinion, sentence, or aspect expresses a *positive*, *neutral*, or *negative* sentiment orientation. Because of the many possible applications and domains of sentiment analysis, different sentiment analysis methods (SAMs) have been developed to address SAC Ribeiro et al. (2016); Serrano-Guerrero et al. (2015).

Usually, in many SAC models, neutral reviews are not considered Koppel and Schler (2006, 2005). There are two main reasons for this: 1) most SAC models focus on binary classification, i.e., in the identification of positive versus negative opinions; 2) neutral reviews lack information due to their ambiguity. However, we consider that the neutral class is key for improving sentiment classification performance Koppel and Schler (2006). Since neutrality is considered somewhere between positivity and negativity, the idea is to deal with it as potential noise, i.e., from a noise filtering classification point of view Sáez et al. (2016). It is understood that detecting and removing noise can improve a model's performance.

In this paper, we assume that neutral opinions must be detected and filtered to improve binary polarity classification. We claim that there is a lack of agreement among SAMs for detecting neutral opinions. So, there is a need to develop a consensus model for improving the identification of neutrality.

Our proposal is to detect neutrality guided by consensus voting among SAMs, and to filter it before the opinion classification step. We first present a neutrality proximity function that assigns weights to polarities according to its proximity to the neutral point. We then propose two polarity aggregation models based on a Weighting Average using the proximity function and on Induced Ordered Weighted Averaging (IOWA) guided by linguistic quantifiers to represent the majority concept, respectively. The main idea is to obtain polarities from several SAMs and aggregate them based on those aggregation models designed using the neutrality proximity function.

We consider an experimental framework with 9 different context datasets and 6 off-the-shelf SAMS. We compute the aggregation polarities and filter out neutral reviews. After that, we develop a SAC task with positive and negative polarities, extracting unigram features and applying two machine learning algorithms. We finally compare and analyze the results and conclude that the polarity consensus voting models, together with neutrality filtering, outperform SAC results.

The paper is structured as follows: we first describe the sentiment analysis problem in Section 2; we define the model for detecting neutrality based on consensus voting in Section 3; we present the experiment setup and the results in Section 4; we draw conclusions and conclude the paper in Section 5.

## 2 Sentiment Analysis

In this section, we describe the main concepts of sentiment analysis (Section 2.1) and the SAMs that we apply (Section 2.2).

### 2.1 The Sentiment Analysis Problem

Due to the increasing number of online reviews, sentiment analysis has emerged as a new field for analyzing this amount of data Cambria (2016). It aims is to analyze sentiments in written text Liu (2015). The number of possible applications is very broad Pang et al. (2008): business intelligence (analyze customer's reviews towards a product) Mishne et al. (2006), politics (predict election results mining social opinions) Wang et al. (2012); Bermingham and Smeaton (2011); Ceron et al. (2014), tourism Marrese-Taylor et al. (2014); Valdivia et al. (2017), personality recognition Majumder et al. (2017) or social studies (evaluate the level of sexist messages in social networks or detect cyberbullying) Jha and Mamidi (2017); Xu et al. (2012). While most studies approach it as a simple categorization problem, sentiment analysis is actually a 'suitcase' research problem that requires tackling many NLP sub-tasks such as:

**Sentiment Analysis Classification**: This is the most popular task. The aim of SAC is to develop models capable of detecting sentiment in texts. The first step is to collect text or reviews to set our analysis. After that, the sentiment is detected. It can be computed by the reviewer or computed with SAMs. Then, features are selected to train the classification model. In this step, text mining techniques are commonly used to extract the most significant features.

**Subjectivity Detection**: This task is related to SAC in the sense that the objective is to classify subjective and objective opinions. The purpose is to filter subjective sentences because they are more opinionated and, hence, can improve classification models.

**Opinion Summarization**: Also known as aspect-based summary or feature-based summary. It consists of developing techniques to sum up large amounts of reviews written by people. The summarization should focus on entities or aspects and their sentiment and should be quantitative Poria et al. (2016); Hu and Liu (2004).

**Opinion Retrieval**: This is a retrieval process, which requires documents to be retrieved and ranked according to their relevance.

**Sarcasm and Irony**: This task aims to detect opinions with sarcastic or ironic expressions. As in subjectivity detection, the target is to delete these opinions from the sentiment analysis process Reyes et al. (2013, 2012).

**Others**: Due to the fact that sentiment analysis is a growing research branch, over recent years many new tasks have emerged, e.g, temporal tagging Zhong et al. (2017), word polarity disambiguation Xia et al. (2015).

As we have previously stated, SAC is a very important task in sentiment analysis. These models classify texts according to their sentiment. This sentiment can be identified in different ways: label polarity ({positive, neutral, negative}), numerical rating ({0, 2, ..., 4} or [0,1]) or emotions {anger, disgust, fear, happiness, sadness, surprise}.

There are three different levels of analysis in this problem:

- The *document level* extracts the sentiment of the whole opinion. This is considered to be the simplest task.
- The *sentence level* extracts sentiments in each sentence of the text. This level is highly related to classifying subjective and objective sentences.
- The *aspect level* is the fine-grained level. This is the most challenging analysis because it extracts sentiments with respect to each opinion target.

### 2.2 Sentiment Analysis Methods

The main task of sentiment analysis is to detect polarity within a text. Therefore, multiple SAMs have been developed to automatically address this challenge. These methods are considered to be varied due to the different properties of online reviews (short texts like tweets, long reviews of microblogs, texts with emoticons, etc.). There are different studies that analyze and compare a large variety of these tools Ribeiro et al. (2016); Serrano-Guerrero et al. (2015). These SAMs are mainly classified into three groups Medhat et al. (2014):

- **Lexicon-Dictionary Based Method (LD):** This method relies on a sentiment dictionary which contains words denoting a sentiment. This dictionary is built from seed words (contained in the corpus or not) and it is then extended with synonyms and antonyms from those seed words Cambria et al. (2016).
- **Machine Learning Based Method (ML):** The main idea is to develop classification models to evaluate new opinions. The classifier algorithm is trained and validated with labeled opinions Oneto et al. (2016).
- **Hybrid Based Method (LD & ML):** The hybrid method consists of a mixture of both methods, LD and ML Cambria and Hussain (2015).

Table 1 shows an overview of the main characteristics of those used in this study and they are introduced in the following subsections.

Table 1: Summary of 5 popular SAMs

| SAM | Type | Output | Reference |
|---|---|---|---|
| Bing | LD | {-1, 0, 1} | Hu and Liu (2004) |
| VADER | LD | [-1, 1] $\in \mathbb{R}$ | Hutto and Gilbert (2014) |
| CoreNLP | ML | {0, 1, 2, 3, 4} | Manning et al. (2014) |
| MeaningCloud | ML | [0, 1] $\in \mathbb{R}$ | bib (2016a) |
| Microsoft Azure | LD & ML | [0, 1] $\in \mathbb{R}$ | bib (2016b) |
| SentiStrength | LD & ML | {-1, 0, 1} | Thelwall (2017) |

#### 2.2.1 Bing

This method is considered one of the first LD methods. It was developed by Hu and Liu Hu and Liu (2004). They took a number of seed adjectives and then developed this dictionary with WordNet Miller et al. (1990). It contains around 6,800 words with its orientation. This method scores sentences with -1 (*negative*), 0 (*neutral*) or 1 (*positive*).

#### 2.2.2 CoreNLP

This method was developed by the Stanford NLP group. They introduce in Socher et al. (2013) a deep learning method, Recursive Neural Tensor Network (RNTN), trained with 215,154 labeled sentences. One of the main contributions of

this study is the introduction of a Sentiment Treebank capable of detecting the compositional effects of sentiments in language, such as negations. CoreNLP outperforms sentiment sentence classification improving by 80.7 This algorithm scores sentence sentiments with a discrete scale from 0 (*very negative*) to 4 (*very positive*).

#### 2.2.3 MeaningCloud

It is a ML method that performs a detailed multilingual sentiment analysis of texts from different sources bib (2016a). The text provided is analyzed to determine if it expresses a positive, negative or neutral sentiment. To this end, the local polarity of the different sentences in the text is identified and the relationship between them evaluated, resulting in a global polarity value for the whole text. Besides polarity at sentence and document level, MeaningCloud uses advanced NLP techniques to detect the polarity attributed to entities and concepts from the text. It provides a reference in the relevant sentence and a list of elements detected with the aggregated polarity derived from all their appearances, also taking into account the grammatical structures in which they are contained.

#### 2.2.4 Microsoft Azure

It is a NLP web service developed by Microsoft Corporation and integrated into Azure Machine Learning toolkit bib (2016b). This API analyzes unstructured text for many NLP tasks. The sentiment analysis task was built as a mix of LD and ML and it was trained for sentiment classification using Sentiment140 data Go et al. (2009). It scores close to 0 indicating negative sentiment and close to 1 indicating positive sentiment.

#### 2.2.5 SentiStrength

It estimates the strength of positive and negative sentiment in short texts, even for informal language. It has human-level accuracy for short social web texts in English, except political texts Thelwall et al. (2010); Thelwall (2017). It builds a lexicon dictionary annotated by humans and improved with the use of machine learning. SentiStrength reports two sentiment strengths, -1 (not negative) to -5 (extremely negative) and 1 (not positive) to 5 (extremely positive). It uses two set scores because psychological research has revealed that humans process simultaneously positive and negative sentiments.

#### 2.2.6 VADER

It is a human-validated SAM developed for twitter and social media contexts. VADER was created from a generalizable, balanced-based, human-curated gold standard sentiment lexicon, Hutto and Gilbert (2014). It combines a lexicon and the processing of the sentence characteristics to determine a sentence polarity. VADER's author identified five heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity that go beyond the bag-of-words model.

## 3 Neutrality Detection Based on Consensus Vote

SAMs are methods trained from different texts. But there are many types of texts: short, long, expressing opinions, objectives, etc. This makes the behavior of SAMs very diverse, and there is a lack of consensus when it comes to detecting polarities, in particular neutrality, as we will demonstrate later in this section. To address this problem, we propose ensembling the different polarities in order to reach a consensus.

Firstly, we explain the main facts that led us to propose an aggregation system for detecting neutrality based on, together with the polarities, aggregation and neutrality filtering for SAC (Section 3.1). We then present a proximity function for neutrality detection (Section 3.2). We describe the two main consensus voting models, the first considers on weighted aggregation based on a proximity function to the neutrality and the second one uses IOWA operators with weights based on a a fuzzy majority guided by linguistic quantifiers (Section 3.3).

### 3.1 Motivation: The global process of SAMs aggregation for SAC

There is a relation between *subjectivity* and *neutrality* which is not clear in the sentiment analysis literature. A *subjective* sentence is defined as the absence of factual material which implies a certain amount of opinion or sentiment that comes form the issuer Wiebe et al. (1999). Liu argues that there are *subjective* sentences that may express objective information Liu (2015).

*Neutrality* means the absence of sentiment or no sentiment Liu (2015). However, we think that opinions expressing mixed or conflicting sentiment may also be considered as *neutral*. *Neutral* reviews show an ambiguous weight of sentiment, i.e., contain an equitable burden of positive and negative polarity.

Due to this fact, this class has been considered as noisy and is broadly excluded in many sentiment models Pang et al. (2002); Wawre and Deshmukh (2016); Da Silva et al. (2014). However, some researchers have tackled the problem of classifying it: the authors in Koppel and Schler (2006) propose taking into account neutral reviews in order to improve classification results or authors in Pang and Lee (2005) propose a multi-sentiment scale, 1-5 stars, to solve the problem of a wider range of sentiment representation, including neutral reviews. In this direction, we propose detecting neutrality by obtaining the votes of several SAMs. We then filter these polarities to improve classification results. The SAC models can improve their polarity classification.

In this study, we consider that opinions can be labeled as *positive*, *negative* or *neutral*. *Neutral* opinions define a threshold between *positive* and *negative*.

We present the notation used in this paper:

- $M = \{m_1, \cdots, m_T\}$ the set of SAMs.
- $O = \{o_1, \cdots, o_N\}$ the set of opinions.
- $p_{ik}$ is the normalized value of the $i$th SAM on the $k$th opinion, i.e., $p_{ik} \in [0, 1]$.
- $I = (e, 1 - e)$ where $e \in [0, 0.5]$ as the *Neutral Interval*.

We thus define the *Sentiment Scale* (Figure 1) which is a numeric scale, from 0 to 1, divided into three chunks: the *negative*, the *neutral* and the *positive*.

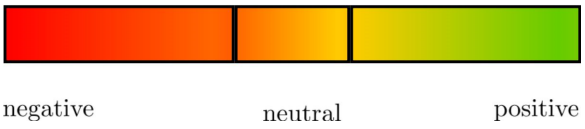



Figure 1: The Sentiment Scale. *Negative* opinions score from 0 to $e$, *neutral* from $e$ to $1 - e$ and *positive* from $1 - e$ to 1, $e \in [0, 0.5]$.

Following these considerations, we claim that there exists a very low agreement on detecting neutral opinions. To discuss this assumption, we apply 6 different SAMs (introduced in Section 2.2) on 9 datasets from different context and domain. Then, we count the total number of neutral opinions detected by each SAM on each corpus.

Table 2: Number of Neutrality Consensus per Corpus

| **Corpus (500 reviews per Corpus)** | **AllAgree** | **AtLeastOneAgree** |
|---|---|---|
| Amazon | 3 | 404 |
| ClintonTrump | 9 | 433 |
| Food | 0 | 422 |
| Cinema | 3 | 377 |
| Movies | 0 | 357 |
| RW | 0 | 441 |
| Ted | 0 | 388 |
| TA-Sagrada Familia | 0 | 348 |
| TA-Alhambra | 0 | 369 |

As we can observe in Table 2, there are only 0.33% of reviews where all SAMs agree on detecting neutrality in all the datasets. However, in a 78.64% of reviews, at least one SAM obtains neutral polarities. This fact has led us to conclude that our claim holds. There is a need for developing consensus models to detect neutrality, filter them and enhance sentiment classification.

To summarize our proposal graphically, the Figure 2 shows a flowchart that presents the global process of SAMs aggregation for SAC. The first step is to collect opinions, then we apply a total number of $T$ SAMs to these opinions. After that, we extract the consensual polarity applying our models. Filtering out neutral reviews and applying text mining techniques for aspect extraction are the next steps. Finally, we classify polarity labels.

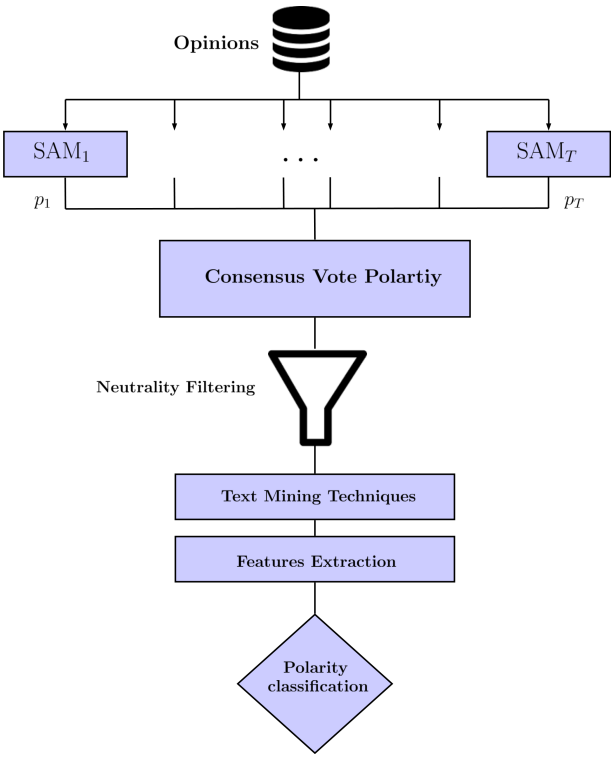



Figure 2: Flowchart of the global process of SAMs aggregation for SAC.

### 3.2 Neutral Proximity Function

In order to measure the proximity to the neutral point, we propose the Neutral Proximity Function (NPF).

**Definition 1** *Neutral Proximity Function.* The NPF is a function that measures the proximity of polarity $p_{ik}$ to the neutrality, rising its absolute maximum when $p_{ik} = 0.5$.

We propose to use the following parametric function of NPF with $\alpha \in (0, 2]$:

$$\begin{aligned} NPF_\alpha \colon [0,1] &\to [0,1] \\ p_{ik} &\mapsto 1 - \alpha|p_{ik} - 0.5|, \alpha \in (0,2]. \end{aligned}$$

The value $\alpha$ is used in NPF$_\alpha$ to scale the proximity values. Figure 3 shows two cases of NPF for $\alpha = 1$ and $\alpha = 2$. As we observe, if a polarity is very negative or very positive ($p_{ik} \approx 0$ or $p_{ik} \approx 1$, respectively) both functions obtain values close to 0. Otherwise, if a polarity is neutral ($p_{ik} \approx 0.5$), they get values close to 1. NPF$_\alpha$ always reaches the absolute minimum when $p_{ik} = 0$ or $p_{ik} = 1$ and the absolute maximum when $p_{ik} = 0.5$. So, it clearly models the proximity of polarities to the *Neutral Interval*, the closer the polarity is, the more weight it gets.

### 3.3 Neutrality Detection Weighting Aggregation

In this section we propose different aggregation models based on weights. The first ones are guided by the NPF (defined before) and the second ones by ordered weights averaging.

#### 3.3.1 Weighting Aggregation Based on a Proximity Function

We propose two average weighting models based on the proximity function (NPF$_\alpha$) to the neutral point, for detecting neutral reviews by consensus. The aggregated polarities are guided by this function. Thus, the aggregated polarity shows consensus on detecting neutrality if its value belongs to the neutrality interval.

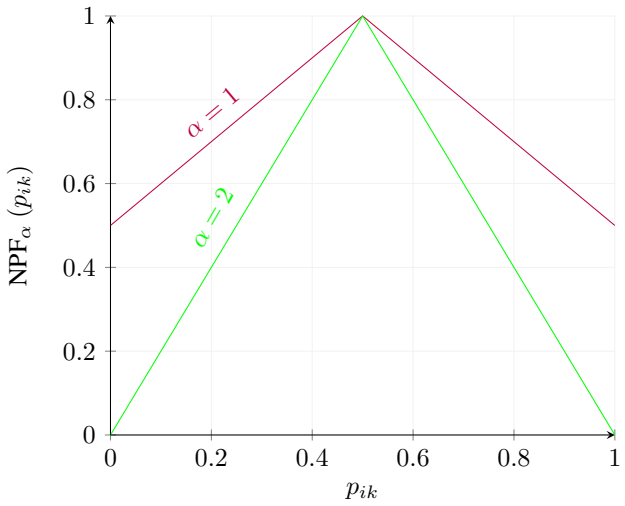



Figure 3: Representation of $\text{NPF}_1$ and $\text{NPF}_2$. If a polarity is very negative or very positive, the function gets the minimum value. If a polarity $p_{ik}$ is close to the *Neutral Interval*, the function rises the maximum value, which is always 1 ($\text{NPF}_\alpha(0.5)$ = max ($\text{NPF}_\alpha(p_{ik})) = 1$).

**Definition 2** *Pro-Neutrality Weight Based Model (ProN).* The *ProN*, $\Phi_{ProN}$, is an aggregation model for sentiment polarities which defines the weighting vector $W$ guided by the $\text{NPF}_{\alpha=1}$, i.e., $w_{ik} = \text{NPF}_{\alpha=1}(p_{ik}) = 1 - |p_{ik} - 0.5|$ and it is expressed such that:

$$\begin{aligned}\Phi_{ProN} &: [0,1]^T \to [0,1] \\ (p_{1k}, \cdots, p_{Tk}) &\mapsto \sum_{i=1}^{T} \frac{w_{ik}}{\sum_{i=1}^{T} w_{ik}} p_{ik}.\end{aligned}$$

Therefore, the ensembled polarity of an opinion $o_k$ is expressed by:

$$\begin{aligned}\Phi_{ProN}((p_{1k}, \cdots, p_{Tk})) &= \sum_{i=1}^{T} \frac{w_{ik}}{\sum_{h=1}^{T} w_{ik}} p_{ik} \\ &= \sum_{i=1}^{T} \frac{NPF_{\alpha=1}(p_{ik})}{\sum_{i=1}^{T} NPF_{\alpha=1}(p_{ik})} p_{ik} \\ &= \sum_{i=1}^{T} \frac{1 - |p_{ik} - 0.5|}{\sum_{i=1}^{T} 1 - |p_{ik} - 0.5|} p_{ik}.\end{aligned}$$

**Definition 3** *Pro-Neutrality Extreme Weight Based Model (ProNE).* The *ProNE*, $\Phi_{ProNE}$, is an aggregation model for sentiment polarities which defines the weighting vector $W$ guided by the $\text{NPF}_{\alpha=2}$, i.e., $w_{ik} = \text{NPF}_{\alpha=2}(p_{ik}) = 1 - 2|p_{ik} - 0.5|$ and it is expressed such that:

$$\begin{aligned}\Phi_{ProNE} &: [0,1]^k \to [0,1] \\ (p_{1k}, \cdots, p_{Tk}) &\mapsto \sum_{h=1}^{T} \frac{w_{ik}}{\sum_{h=1}^{T} w_{ik}} p_{ik}.\end{aligned}$$

Therefore, the ensembled polarity of an opinion $o_k$ is expressed by:

$$\Phi_{ProNE}((p_{1k},\cdots,p_{Tk})) = \sum_{h=1}^{T} \frac{w_{ik}}{\sum_{h=1}^{T} w_{ik}} p_{ik}$$
$$= \sum_{h=1}^{T} \frac{NPF_{\alpha=2}(p_{ik})}{\sum_{h=1}^{T} NPF_{\alpha=2}(p_{ik})} p_{ik}$$
$$= \sum_{h=1}^{T} \frac{1-2|p_{ik}-0.5|}{\sum_{h=1}^{T} 1-2|p_{ik}-0.5|} p_{ik}.$$

As reference for experimental analysis, we consider the basic model which averages polarities and give them an equal weight.

**Definition 4** *Average Based Model (AVG).* The *AVG* is an aggregation model for sentiment polarities which defines the weighting vector by $W = \frac{1}{T}$ and it is expressed such that:

$$\Phi_{AVG}\colon [0,1]^T \to [0,1]$$
$$(p_{1k},\cdots,p_{Tk}) \mapsto \frac{1}{T}\sum_{h=1}^{T} p_{ik}.$$

Note that this model is equivalent to the *arithmetic mean* over the $k$ polarities.

### 3.3.2 Aggregation Based on Majority Vote Guided by Linguistic Quantifiers

In many decision-making problems, the opinion of the majority of agents is the relevant output Jung and Jo (2007). Yager proposed the *Ordered Weighted Averaging (OWA) operator* modelling the *fuzzy majority*, i.e., the idea that a decision will be made if most of the agents agree Yager (1988, 1996). Soon after, the same author proposed an OWA operator but induced the order of the argument variable via an order-induced vector, the *Induced Ordered Weighted Averaging (IOWA) operator*. The IOWA operator is considered a generalization of OWA operators with a specific semantic in the aggregation process Yager (1988, 1996); Yager and Filev (1999). Recently, IOWA operators have been used for sentiment classification using the vote of majority for classifiers aggregation Appel et al. (2017).

**Definition 5** *OWA Yager (1988); Chiclana et al. (2007).* An *OWA operator* of dimension $n$ is a mapping $\phi : \mathbb{R}^n \to \mathbb{R}$ that has an associated weighting vector $W$ such that $w_i \in [0,1]$, $\sum_{i=1}^{n} w_i = 1$, and is defined to aggregate a list of values $\{p_1,\ldots,p_n\}$ following this expression:

$$\phi(p_1,\ldots,p_n) = \sum_{i=1}^{n} w_i p_{\sigma(i)},$$

being $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$ a permutation such that $p_{\sigma(i)} \geq p_{\sigma(i+1)}$, $\forall i = 1,\ldots,n-1$.

**Definition 6** *IOWA Yager and Filev (1999); Chiclana et al. (2007).* An *IOWA operator* of dimension $n$ is a mapping $\Psi : (\mathbb{R}\times\mathbb{R})^n \to \mathbb{R}$ that has an associated weighting vector W such that $w_i \in [0,1]$, $\sum_{i=1}^{n} w_i = 1$, and it is defined to aggregate the set of second arguments of a list of $n$ 2-tuples:

$$\Psi(\langle u_1,p_1\rangle,\ldots,\langle u_n,p_n\rangle) = \sum_{i=1}^{n} w_i p_{\sigma(i)},$$

being $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall i = 1,\ldots,n-1$.

The vector of values $U = (u_1,\ldots,u_n)$ is defined as the *order-inducing* vector and $(p_1,\ldots,p_n)$ as the *values of the argument variable*. In this way, the *order-inducing* reorders the *values of the argument variable* based on its magnitude.

*Linguistic quantifiers* are widely used for modeling the concept of quantification to represent the fuzzy majority Pasi and Yager (2006). *At least half*, *Most of* and *Many as possible* are some examples of these quantifiers (see Figure

4), which can be modeled explicitly as a fuzzy set by the following function proposed by Yager in Yager (1996). We propose to use these operators because they are aligned to the idea that we are considering for aggregating polarities Appel et al. (2017).

$$Q_{(a,b)}(x) = \begin{cases} 0 & \text{if } 0 \leq x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } b \leq x \leq 1 \end{cases}$$

The values that are used for the pair $(a, b)$ are Kacprzyk (1986):

$$Q_{\textit{At least half}}(x) = Q_{(0,0.5)}(x)$$

$$Q_{\textit{Most of}}(x) = Q_{(0.3,0.8)}(x)$$

$$Q_{\textit{Many as possible}}(x) = Q_{(0.5,1)}(x)$$

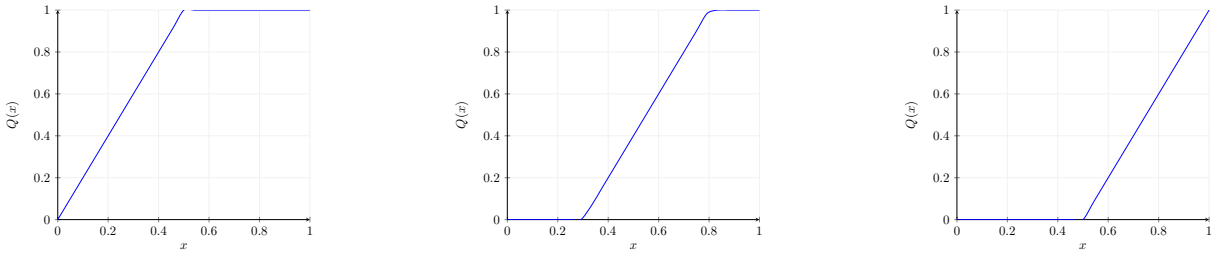



Figure 4: Linguistic Quantifiers Represented as Fuzzy Sets: *At least half*, *Most of* and *Many as possible*, respectively.

Then, the weights are calculated as follows:

$$w_i^{(a,b)} = Q_{(a,b)}\left(\frac{i}{T}\right) - Q_{(a,b)}\left(\frac{i-1}{T}\right)$$

We define the following IOWA based models taking into account linguistic quantifiers and the $\text{NPF}_\alpha$ to tackle the consensus voting among SAMs based on majority:

**Definition 7** *IOWA At Least Half Pro-Neutrality System Based (ALH-ProN).* The *IOWA ALH-ProN operator* of dimension $T$ is a mapping $\Psi_{ALH-ProN} : ([0,1] \times [0,1])^T \rightarrow [0,1]$ that has an associated weighting vector W such that $w_i^{(0,0.5)}$ and it is defined to aggregate the set of second arguments of a list of $T$ 2-tuples:

$$\Psi_{ALH-ProN}(\langle u_1, p_{1k}\rangle, \ldots, \langle u_T, p_{Tk}\rangle) = \sum_{i=1}^{T} w_i^{(0,0.5)} p_{\sigma(i)k},$$

being $\sigma : \{1, \ldots, T\} \rightarrow \{1, \ldots, T\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall i = 1, \ldots, T-1$, and $u_i = NPF_{\alpha=1}(p_{ik}) = 1 - |p_{ik} - 0.5|$.

**Definition 8** *IOWA Most Of Pro-Neutrality System Based (MO-ProN).* The *IOWA MO-ProN operator* of dimension $T$ is a mapping $\Psi_{MO-ProN} : ([0,1] \times [0,1])^T \rightarrow [0,1]$ that has an associated weighting vector W such that $w_i^{(0.3,0.8)}$ and it is defined to aggregate the set of second arguments of a list of $T$ 2-tuples:

$$\Psi_{MO-ProN}(\langle u_1, p_{1k}\rangle, \ldots, \langle u_T, p_{Tk}\rangle) = \sum_{i=1}^{T} w_i^{(0.3,0.8)} p_{\sigma(i)k},$$

being $\sigma : \{1, \ldots, T\} \rightarrow \{1, \ldots, T\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall i = 1, \ldots, T-1$, and $u_i = NPF_{\alpha=1}(p_{ik}) = 1 - |p_{ik} - 0.5|$.

**Definition 9** *IOWA Many As Possible Pro-Neutrality System Based (MAP-ProN).* The *IOWA MAP-ProN operator* of dimension $T$ is a mapping $\Psi_{MAP-ProN} : ([0,1] \times [0,1])^T \rightarrow [0,1]$ that has an associated weighting vector W such that $w_i^{(0.5,1)}$ and it is defined to aggregate the set of second arguments of a list of $T$ 2-tuples:

$$\Psi_{MAP-ProN}(\langle u_1, p_{1k}\rangle, \ldots, \langle u_T, p_{Tk}\rangle) = \sum_{i=1}^{T} w_i^{(0.5,1)} p_{\sigma(i)k},$$

being $\sigma : \{1,\ldots,T\} \rightarrow \{1,\ldots,T\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall i = 1,\ldots,T-1$, and $u_i = NPF_{\alpha=1}(p_{ik}) = 1 - |p_{ik} - 0.5|$.

Note that in these operators, the neutrality proximity function (see Figure 3) sorts polarities and linguistic quantifiers (see Figure 4) provide weights.

Finally, we introduce two particular cases of IOWA. They induce weights taking into account the minimum and maximum extreme polarity. More precisely:

**Definition 10** *IOWA Minimum Neutrality (MinN).* The *IOWA MinN* of dimension $T$ is a mapping $\Psi_{MinN} : ([0,1] \times [0,1])^T \rightarrow [0,1]$ that has an associated weighting vector W and it is defined to aggregate the set of second arguments of a list of $T$ 2-tuples:

$$\Psi_{MinN}(\langle u_1, p_{1k}\rangle, \ldots, \langle u_T, p_{Tk}\rangle) = \sum_{i=1}^{T} w_i p_{\sigma(i)k},$$

being $\sigma : \{1,\ldots,T\} \rightarrow \{1,\ldots,T\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall h = 1,\ldots,T-1$, and $u_i = NPF_{\alpha=1}(p_{ik}) = 1 - |p_{ik} - 0.5|$ with $w_T = 1$ and $w_i = 0$ for $\forall i = 1,\ldots,T-1$.

**Definition 11** *IOWA Maximum Neutrality (MaxN).* The *IOWA MaxN* of dimension $T$ is a mapping $\Psi_{MaxN} : ([0,1] \times [0,1])^T \rightarrow [0,1]$ that has an associated weighting vector W and it is defined to aggregate the set of second arguments of a list of $T$ 2-tuples:

$$\Psi_{MaxN}(\langle u_1, p_{1k}\rangle, \ldots, \langle u_T, p_{Tk}\rangle) = \sum_{h=1}^{T} w_i p_{\sigma(i)k},$$

being $\sigma : \{1,\ldots,T\} \rightarrow \{1,\ldots,T\}$ a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$, $\forall h = 1,\ldots,T-1$, and $u_i = NPF_{\alpha=1}(p_{ik}) = 1 - |p_{ik} - 0.5|$ with $w_1 = 1$ and $w_i = 0$ for $\forall h = 2,\ldots,T$.

The *IOWA MinN* operator ($\Psi_{MinN}$) simply selects the polarity with the more extreme value (very positive or negative polarities) and the *IOWA MaxN* ($\Psi_{MaxN}$) with the more central value (neutral polarities).

A very interesting property of *IOWA MinN* operator ($\Psi_{MinN}$) is that it can detect whether all SAM agree on detecting neutrality. Note that if all SAM polarities are close to the neutral point, the aggregated polarity of this operator is also close to this point (Maximum Consensus on Neutrality). On the other hand, *IOWA MaxN* ($\Psi_{MaxN}$) detects when at least one SAM detects a neutral review. Table 2 shows their associated neutralities for our cases of study, where *All agree* refers to *MinN* and *AtLeastOneAgree* to *MaxN*.

## 4 Experimented Study

In this section, we present an experimented analysis to validate the consensus vote for neutrality detection. We describe the datasets that we use for our study (Section 4.1). After we explain the process of our experiment (Section 4.2) and finally show the results (Section 4.3).

### 4.1 Datasets

This study is based on nine datasets. We have collected text data from different sources. In order to develop robust analysis, we get data with different properties (short texts like tweets, long reviews like Trip Advisor data) and from different domains (politics, tourism, movies...).

- Amazon[5]: Sentiment ratings from a minimum of 20 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability).
- ClintonTrump[2]: Tweets from the major party candidates for the 2016 US Presidential Election.

[2] `https://www.kaggle.com/benhamner/clinton-trump-tweets`

- Food[3]: Food reviews from Amazon McAuley and Leskovec (2013).
- Cinema Reviews[5]: It includes 10,605 sentence-level snippets. The snippets were derived from an original set of 2,000 movie reviews (1,000 positive and 1,000 negative).
- Movies[4]: Single sentences extracted from movie reviews Pang and Lee (2005).
- Runner's World (RW)[5]: Comments from Runner's World Forum.
- TED Talks[5]: Influential videos from expert speakers on education, business, science, tech and creativity, with subtitles in more than 100 languages.
- TA-Sagrada Familia: TripAdvisor reviews from the most popular monument in Barcelona, the Sagrada Familia.
- TA-Alhambra: TripAdvisor reviews from the most popular monument in Granada, the Alhambra.

Table 3 shows the number of words and sentences and its average by corpus. From these numbers, we can infer that Amazon, ClintonTrump, Cinema, Movies and Ted are short reviews. This is because these corpora are texts from Twitter (140 character limit) or single sentences. Food, RW, and TA are corpora with larger reviews.

Table 3: Summary of Quantitive Text Analysis of Datasets (Words and Sentences)

| Corpus | NumWords | AVGNumWords | NumSentences | AVGNumSentences |
|---|---|---|---|---|
| Amazon | 7,787 | 15.57 | 500 | 1.00 |
| ClintonTrump | 8,674 | 17.35 | 881 | 1.76 |
| Food | 40,775 | 81.55 | 2,512 | 5.02 |
| Cinema | 10,433 | 20.87 | 564 | 1.13 |
| Movies | 9,623 | 19.25 | 523 | 1.05 |
| RW | 34,871 | 69.74 | 2,337 | 4.67 |
| Ted | 8,971 | 17.94 | 502 | 1.00 |
| TA-Sagrada Familia | 30,520 | 61.04 | 2,033 | 4.07 |
| TA-Alhambra | 45,665 | 91.33 | 2,800 | 5.60 |

From each dataset, we randomly select 500 reviews which sum up a total of 4,500 opinions.

### 4.2 Experimental Setup

The main target of our experiments is to study the behaviour of polarity classification algorithms in different scenarios. The idea is to evaluate if these algorithms considering neutral reviews as class noise can improve their performance. The experiment setup is described as follows (considering the flowchart of Figure 2).

We apply the six described SAMs on the datasets. For the Bing and CoreNLP methods, we split the text into sentences and extract sentiment for each. The overall sentiment is defined by majority vote. For MeaningCloud, Microsoft Azure, SentiStrength and VADER the whole text is evaluated.

Once we obtain the sentiment for each review, we normalize the polarities taking into account the SAM and the corpus at the $[0, 1]$ interval.

We then compute the proposed aggregation approaches over the 6 SAM normalized outputs. Finally, we label the polarities as follows: $[0, 0.4]$ are negative reviews, $(0.4, 0.6)$ are neutral reviews and $[0.6, 1]$ are positive reviews (taking notation of Section 3.1, $e = 0.4$ and $I = (0.4, 0.6)$).

We preprocess the text removing stop words, punctuation and numbers. We stem all words and extract the 10 more relevant features in positive and negative reviews with the *tf-idf* metric. We then build the document-term matrix dummy which element $a_{ij} = 1$ if in the $i$-document/review the $j$-word is present.

The models are validated with a 5-fold cross validation (datasets are split in 80 % for training and 20 % for testing). The classification algorithms selected for this study are: SVM and XGBOOST. We select SVM algorithm because it has been broadly used in the sentiment analysis literature Poursepanj et al. (2013). On the other hand, XGBOOST has been widely deployed in many data science competitions Chen and Guestrin (2016). The parameters of these algorithms are tuned by the `train function` of the `caret` package of R Studio. Finally, we analyze the AUC measure in the test set.

[3] `https://www.kaggle.com/snap/amazon-fine-food-reviews`

[4] `https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data`

[5] `https://bitbucket.org/matheusaraujo/ifeel-benchmarking-datasets/src`

### 4.3 Results

In this section we first evaluate the consensus among SAMs in detecting neutral opinions (Section 4.3.1). Afterwards, we present the classification results of the proposed models (Section 4.3.2).

Table 4: Number of Neutral Instances of **Individual SAMs**, I=(0.4, 0.6)

| Corpus | Bing | CoreNLP | MC | Microsoft | SentiStr | VADER |
|---|---|---|---|---|---|---|
| Amazon | 115 | 105 | 186 | 144 | 251 | 193 |
| ClintonTrump | 277 | 221 | 115 | 93 | 228 | 130 |
| Food | 239 | 209 | 61 | 40 | 102 | 22 |
| Cinema | 283 | 34 | 92 | 52 | 152 | 110 |
| Movies | 124 | 42 | 124 | 86 | 174 | 150 |
| RW | 236 | 211 | 114 | 50 | 156 | 87 |
| Ted | 147 | 55 | 107 | 98 | 194 | 155 |
| TA-Sagrada Familia | 80 | 222 | 66 | 36 | 99 | 34 |
| TA-Alhambra | 99 | 239 | 54 | 22 | 84 | 16 |
| Average | 177.778 | 148.667 | 102.111 | 69.000 | 160.000 | 99.667 |

Table 5: Number of Neutral Instances of **Aggregation Models**, I=(0.4, 0.6)

| Corpus | MinN | MaxN | AvgN | ProN | ProNE | MAP-ProN | ALH-ProN | MO-ProN |
|---|---|---|---|---|---|---|---|---|
| Amazon | 3 | 404 | 221 | 245 | 287 | 183 | 182 | 174 |
| ClintonTrump | 9 | 433 | 207 | 228 | 298 | 163 | 225 | 149 |
| Food | 0 | 422 | 115 | 151 | 329 | 60 | 258 | 77 |
| Cinema | 3 | 377 | 143 | 167 | 239 | 103 | 156 | 106 |
| Movies | 0 | 357 | 199 | 220 | 289 | 175 | 168 | 157 |
| RW | 0 | 441 | 175 | 214 | 333 | 123 | 270 | 128 |
| Ted | 0 | 388 | 144 | 172 | 247 | 119 | 134 | 98 |
| TA-Sagrada Familia | 0 | 348 | 119 | 158 | 331 | 61 | 260 | 76 |
| TA-Alhambra | 0 | 369 | 96 | 132 | 307 | 41 | 281 | 61 |
| **Average** | 1.667 | 393.222 | 157.667 | 187.444 | 295.556 | 114.222 | 214.889 | 114.000 |

Table 6: Test AUC for SVM models, **Individual SAMs**

| Corpus | Bing | CoreNLP | MC | MSAzure | SentiStr | VADER | Average |
|---|---|---|---|---|---|---|---|
| Amazon | 0.407 | 0.800 | 0.576 | 0.663 | **0.851** | 0.734 | 0.672 |
| ClintonTrump | **0.878** | 0.721 | 0.816 | 0.694 | 0.723 | 0.755 | 0.764 |
| Food | **0.763** | 0.647 | 0.500 | 0.613 | 0.469 | 0.413 | 0.567 |
| Cinema | 0.387 | **0.628** | 0.507 | 0.343 | 0.368 | 0.381 | 0.436 |
| Movies | 0.360 | 0.431 | 0.398 | **0.546** | 0.362 | 0.509 | 0.434 |
| RW | 0.697 | **0.751** | 0.505 | 0.559 | 0.682 | 0.596 | 0.632 |
| Ted | **0.804** | 0.260 | 0.263 | 0.341 | 0.215 | 0.346 | 0.371 |
| TA-Sagrada Familia | **0.850** | 0.656 | 0.714 | 0.651 | 0.690 | 0.724 | 0.714 |
| TA-Alhambra | 0.647 | **0.895** | 0.763 | 0.751 | 0.592 | 0.565 | 0.702 |
| **Average** | **0.644** | 0.643 | 0.560 | 0.573 | 0.550 | 0.558 | 0.588 |

#### 4.3.1 Model Analysis: Neutrality Consensus Among SAMs

We first study the consensus rate when it comes to detect neutrality. For this, we present Tables 4 and 5 which show the number of neutral instances by the six individual SAMs and the aggregation models, respectively.

As we observe in Table 4 of individual SAMs, there are significant differences between the number of neutral reviews for each dataset. For instance, we observe that VADER detects 16 neutral reviews in TA-Alhambra, while CoreNLP detects 239 instances. But this same SAM only detects 34 neutral reviews in the Cinema's dataset, while Bing detects 283. This fact confirms that our claim holds, there is a need for a consensus voting model to detect neutral reviews.

In Table 5 we present the total number of neutral reviews detected for each proposed aggregation model. MinN obtains a very low number of neutral reviews per corpus but, on the other hand, MaxN obtains a high number. As we have explained before, that means that low agreement exists when detecting neutrality in reviews. There are significant differences in the consensus voting guided by the proximity function, averaging and linguistic quantifiers.

#### 4.3.2 Model Analysis: Classification Performance

We study the classification performance after filtering neutral polarities. We discuss the results of the individual SAMs and the consensus models. We present Tables 6 and 7 which contain the test AUC scores of the SVM and XGBOOST models. Polarities are obtained by each SAM. We then introduce Tables 8 and 9 which show the test AUC scores of the two classifiers. In this case, the polarities correspond to the consensus vote models.

Table 7: Test AUC for XGBOOST models, **Individual SAMs**

| Corpus | Bing | CoreNLP | MC | MSAzure | SentiStr | VADER | Average |
|---|---|---|---|---|---|---|---|
| Amazon | 0.400 | 0.265 | 0.434 | 0.613 | **0.853** | 0.690 | 0.542 |
| ClintonTrump | **0.795** | 0.735 | 0.767 | 0.719 | 0.711 | 0.731 | 0.743 |
| Food | 0.507 | 0.648 | **0.692** | 0.564 | 0.566 | 0.629 | 0.601 |
| Cinema | 0.485 | 0.387 | 0.518 | 0.367 | 0.438 | **0.595** | 0.465 |
| Movies | 0.529 | 0.471 | 0.437 | **0.574** | 0.415 | 0.409 | 0.472 |
| RW | 0.291 | **0.710** | 0.579 | 0.498 | 0.498 | 0.418 | 0.499 |
| Ted | **0.366** | 0.318 | 0.302 | 0.296 | 0.252 | 0.358 | 0.315 |
| TA-Sagrada Familia | 0.529 | 0.643 | 0.439 | 0.622 | 0.653 | **0.746** | 0.605 |
| TA-Alhambra | 0.678 | 0.870 | **0.916** | 0.592 | 0.501 | 0.541 | 0.683 |
| **Average** | 0.509 | 0.561 | 0.565 | 0.538 | 0.543 | **0.569** | 0.547 |

Table 8: Test AUC for SVM models, **Aggregation Models**

| Corpus | MinN | AvgN | ProN | ProNE | MAP-ProN | ALH-ProN | MO-ProN |
|---|---|---|---|---|---|---|---|
| Amazon | 0.302 | 0.795 | **0.853** | 0.752 | 0.633 | 0.730 | 0.757 |
| ClintonTrump | 0.461 | 0.733 | **0.870** | 0.620 | 0.683 | 0.790 | 0.824 |
| Food | 0.456 | 0.407 | 0.302 | 0.500 | **0.613** | 0.609 | 0.609 |
| Cinema | 0.440 | 0.473 | 0.218 | **0.740** | 0.612 | 0.719 | 0.558 |
| Movies | 0.638 | 0.396 | 0.429 | 0.645 | **0.645** | 0.532 | 0.524 |
| RW | 0.349 | 0.375 | 0.631 | 0.652 | 0.713 | **0.737** | 0.530 |
| Ted | 0.305 | 0.215 | 0.230 | 0.286 | 0.713 | 0.708 | **0.757** |
| TA-Sagrada Familia | 0.535 | **0.891** | 0.693 | 0.693 | 0.875 | 0.776 | 0.629 |
| TA-Alhambra | 0.819 | 0.824 | **0.941** | 0.941 | 0.507 | 0.845 | 0.767 |
| **Average** | 0.478 | 0.568 | 0.574 | 0.648 | 0.666 | **0.716** | 0.662 |

We discuss the attained results summarized in the following items:

- **SVM and XGBOOST for Individual SAMs (Tables 6 and 7).** As we observe in Table 6, there is no method that stands out from the rest. The classification results with the SVM algorithm varies widely on each column, which means that SAMs strongly depends on the corpus where they are evaluated. The results of the XGBOOST classifier presented in Table 7 shows a very similar behaviour. We also detect overfitted models due to the fact that the test AUC is much lower than train (CoreNLP and Ted dataset, SentiStrength and Ted dataset, etc.)
- **SVM and XGBOOST for Aggregation Models (Tables 8 and 9).** We observe that the Aggregation Models also present widely results over the datasets. The results for the SVM and XGBOOST classifiers are similar. We also detect some overfitted models, but to a lesser extent. Note that MaxN is not reported because of the low number of positive and negative instances.
- **Weighting Aggregation vs. Linguistic Quantifiers (Tables 8 and 9).** Studying the average of the polarity classification results of both aggregation models, we observe that Linguistic Quantifiers shows a better performance except for the MinN.
- **ALH-ProN, the best aggregation model (Tables 8 and 9).** This model obtains the best average classification results (0.716 and 0.669 for SVM and XGBOOST algorithms). The main idea behind this Linguistic Quantifier is to obtain *at least half* of the consensus among the different SAMs. If we analyze the number of detected neutral instances of this model (see Table 5), we observe that ALH-ProN obtains an average level of neutrality detection. The weights obtained by this model are: (0.3, 0.3, 0.3, 0, 0, 0) which led us to conclude that it does not take into account the 3 SAMs with extreme maximum polarities and gives more weight to those with more conservative behaviour. Therefore, ALH-ProN is a conservative model in terms of neutrality.
- **ALH-ProN vs. Single Models (Tables 6, 7, 8 and 9).** Finally, we compare the performance of ALH-ProN and the SAMs. Analyzing the results for SVM (see Tables 6 and 8), we observe that ALH-ProN obtains better results on average (ALH-ProN gets 0.072 more points than Bing). Analyzing the results for XGBOOST (see Tables 7 and 9), we observe that ALH-ProN also obtains better results on average (ALH-ProN gets 0.1 more points than VADER). Therefore, ALH-ProN outperforms single models.

## 5 Concluding Remarks

In this study we have shown that there is a low consensus among SAMs in detecting neutrality. This may be due to different reasons, such as that some tools are trained for one type of text, making it difficult for them to find the polarity in another. As we know, humans write in very different ways and even more so if we have space constraints, as in the case of Twitter. Therefore, a tool trained with tweets will not behave well when analyzing opinions in TripAdvisor, where the text is longer and emoticons are not usually used.

Table 9: Test AUC for XGBOOST models, **Aggregation Models**

| Corpus | MinN | AvgN | ProN | ProNE | MAP-ProN | ALH-ProN | MO-ProN |
|---|---|---|---|---|---|---|---|
| Amazon | 0.336 | 0.776 | **0.819** | 0.730 | 0.600 | 0.767 | 0.737 |
| ClintonTrump | 0.372 | 0.671 | **0.827** | 0.415 | 0.648 | 0.724 | 0.656 |
| Food | 0.671 | 0.234 | 0.344 | 0.324 | **0.687** | 0.637 | 0.538 |
| Cinema | 0.444 | 0.425 | 0.371 | 0.500 | **0.657** | 0.448 | 0.514 |
| Movies | 0.435 | **0.598** | 0.471 | 0.427 | 0.568 | 0.434 | 0.440 |
| RW | 0.315 | 0.375 | 0.640 | 0.708 | 0.691 | 0.704 | **0.717** |
| Ted | 0.346 | 0.215 | 0.264 | 0.323 | 0.701 | **0.810** | 0.686 |
| TA-Sagrada Familia | 0.499 | **0.859** | 0.667 | 0.667 | 0.721 | 0.756 | 0.551 |
| TA-Alhambra | 0.768 | 0.792 | **0.932** | 0.932 | 0.606 | 0.743 | 0.652 |
| **Average** | 0.465 | 0.549 | 0.593 | 0.559 | 0.653 | **0.669** | 0.610 |

This led us to propose two models of consensus via polarity aggregation. The idea is to detect neutrality based on these consensus models and then filter it out. Then, we study their performance on positive and negative polarities. The results obtained in this study have shown that detecting neutrality based on a consensus improves classification precision. In fact, the ALH-ProN model gets the best results on average. It weighs the polarity of the 3 out of 6 less extreme SAMs.

In fact, there is a wide analysis of classification aggregation Kuncheva and Rodríguez (2014); Rokach (2016). There are studies showing that ensembles have proven to outperform single models for SAC Appel et al. (2017); Da Silva et al. (2014). The aggregation is also positive for neutrality detection via polarity aggregation.

For future work, we will consider studying different methods for feature or aspects extraction in order to evaluate the robustness of the models Schouten and Frasincar (2016); Poria et al. (2016). We propose to compare the aggregated polarities with the ground truth. In this sense, we propose to label opinions by different experts and then to build aggregation models taking into account experts' sentiment and learning how to aggregate SAM based polarities.

## Acknowledgments

We thank our anonymous reviewers for their valuable feedback, which helped to improve the paper. This research has been supported by FEDER and by the Spanish National Research Project TIN2014-57251-P.

## References

E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, A Practical Guide to Sentiment Analysis, Springer, 2017.

B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.

J. A. Balazs, J. D. Velásquez, Opinion Mining and Information Fusion. A survey, Information Fusion 27 (2016) 95–110.

S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Information Fusion 36 (2017) 10–25.

G. Bello-Orgaz, J. J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Information Fusion 28 (2016) 45–59.

S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017a) 98–125.

S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, 873–883, 2017b.

F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, F. Benevenuto, SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Science 5 (1) (2016) 1–29.

J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, E. Herrera-Viedma, Sentiment analysis: a review and comparative analysis of web services, Information Sciences 311 (2015) 18–38.

M. Koppel, J. Schler, The importance of neutral examples for learning sentiment, Computational Intelligence 22 (2) (2006) 100–109.

M. Koppel, J. Schler, Using neutral examples for learning polarity, in: International Joint Conference on Artificial Intelligence, vol. 19, Morgan Kaufmann Publishers Inc., 1616 – 1617, 2005.

J. A. Sáez, M. Galar, J. Luengo, F. Herrera, INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control, Information Fusion 27 (2016) 19–32.

E. Cambria, Affective Computing and Sentiment Analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107.

B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Foundations and Trends® in Information Retrieval 2 (1–2) (2008) 1–135.

G. Mishne, N. S. Glance, et al., Predicting Movie Sales from Blogger Sentiment., in: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 155–158, 2006.

H. Wang, D. Can, A. Kazemzadeh, F. Bar, S. Narayanan, A system for real-time twitter sentiment analysis of 2012 us presidential election cycle, in: Proceedings of the ACL 2012 System Demonstrations, 115–120, 2012.

A. Bermingham, A. Smeaton, On using Twitter to monitor political sentiment and predict election results, in: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology), 2–10, 2011.

A. Ceron, L. Curini, S. M. Iacus, G. Porro, Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France, New Media & Society 16 (2) (2014) 340–358.

E. Marrese-Taylor, J. D. Velásquez, F. Bravo-Marquez, A novel deterministic approach for aspect-based opinion mining in tourism products reviews, Expert Systems with Applications 41 (17) (2014) 7764–7775.

A. Valdivia, M. V. Luzón, F. Herrera, Sentiment Analysis in TripAdvisor, IEEE Intelligent Systems 32 (4) (2017) 72–77.

N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning-based document modeling for personality detection from text, IEEE Intelligent Systems 32 (2) (2017) 74–79.

A. Jha, R. Mamidi, When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data, in: Proceedings of the Second Workshop on NLP and Computational Social Science, 7–16, 2017.

J.-M. Xu, X. Zhu, A. Bellmore, Fast learning for sentiment analysis on bullying, in: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, 10, 2012.

S. Poria, E. Cambria, A. Gelbukh, Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network, Knowledge-Based Systems 108 (2016) 42–49.

M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 168–177, 2004.

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

A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: The figurative language of social media, Data & Knowledge Engineering 74 (2012) 1–12.

X. Zhong, A. Sun, E. Cambria, Time expression analysis and recognition using syntactic token types and general heuristic rules, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 420 – 429, 2017.

Y. Xia, E. Cambria, A. Hussain, H. Zhao, Word Polarity Disambiguation Using Bayesian Model and Opinion-Level Features, Cognitive Computation 7 (3) (2015) 369–380.

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

E. Cambria, S. Poria, R. Bajpai, B. Schuller, SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives, in: International Conference on Computational Linguistics, 2666–2677, 2016.

L. Oneto, F. Bisio, E. Cambria, D. Anguita, Statistical learning theory and ELM for big social data analysis, IEEE Computational Intelligence Magazine 11 (3) (2016) 45–55.

E. Cambria, A. Hussain, Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis, Springer, 2015.

C. J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55–60, 2014.

MeaningCloud – Opinion Mining API, `https://www.meaningcloud.com/products/sentiment-analysis`, online; accessed Jan 2017, 2016a.

Microsoft Cognitive Services – Text Analytics API, `https://www.microsoft.com/cognitive-services/en-us/text-analytics-api`, online; accessed Jul 2016, 2016b.

M. Thelwall, The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength, in: Cyberemotions: Collective Emotions in Cyberspace, Springer International Publishing, 119–134, 2017.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to WordNet: An on-line lexical database, International journal of lexicography 3 (4) (1990) 235–244.

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

A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford 1 (2009) 12.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, Journal of the American Society for Information Science and Technology 61 (12) (2010) 2544–2558.

J. M. Wiebe, R. F. Bruce, T. P. O'Hara, Development and use of a gold-standard data set for subjectivity classifications, in: Proceedings of the 37th annual meeting of the Association for Computational Linguistics, 246–253, 1999.

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

S. V. Wawre, S. N. Deshmukh, Sentiment classification using machine learning techniques, Int. J. Sci. Res 5 (4) (2016) 1–3.

N. F. Da Silva, E. R. Hruschka, E. R. Hruschka, Tweet sentiment analysis with classifier ensembles, Decision Support Systems 66 (2014) 170–179.

B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 115–124, 2005.

J. J. Jung, G.-S. Jo, Consensus-based evaluation framework for cooperative information retrieval systems, in: KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, Springer, 169–178, 2007.

R. R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, IEEE Transactions on systems, Man, and Cybernetics 18 (1) (1988) 183–190.

R. R. Yager, Quantifier guided aggregation using OWA operators, International Journal of Intelligent Systems 11 (1) (1996) 49–73.

R. R. Yager, D. P. Filev, Induced ordered weighted averaging operators, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29 (2) (1999) 141–150.

O. Appel, F. Chiclana, J. Carter, H. Fujita, A Consensus Approach to the Sentiment Analysis Problem Driven by Support-Based IOWA Majority, International Journal of Intelligent Systems 32 (9) (2017) 947–965.

F. Chiclana, E. Herrera-Viedma, F. Herrera, S. Alonso, Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations, European Journal of Operational Research 182 (1) (2007) 383–399.

G. Pasi, R. R. Yager, Modeling the concept of majority opinion in group decision making, Information Sciences 176 (4) (2006) 390–414.

J. Kacprzyk, Group decision making with a fuzzy linguistic majority, Fuzzy sets and systems 18 (2) (1986) 105–118.

J. J. McAuley, J. Leskovec, From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews, in: Proceedings of the 22nd international conference on World Wide Web, ACM, 897–908, 2013.

H. Poursepanj, J. Weissbock, D. Inkpen, uOttawa: System description for SemEval 2013 Task 2 Sentiment Analysis in Twitter, in: SemEval@ NAACL-HLT, 380–383, 2013.

T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, ACM, 785–794, 2016.

L. I. Kuncheva, J. J. Rodríguez, A weighted voting framework for classifiers ensembles, Knowledge and Information Systems 38 (2) (2014) 259–275.

L. Rokach, Decision forest: Twenty years of research, Information Fusion 27 (2016) 111–125.

K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Transactions on Knowledge and Data Engineering 28 (3) (2016) 813–830.

# 4 What do People Think about this Monument? Understanding Negative Reviews via Deep Learning and Descriptive Rules.

- A. Valdivia, E. Martínez-Cámara, I. Chaturvedi, MV. Luzón, E. Cambria, YS. Ong, F. Herrera. What do people think about this monument? Understanding negative reviews via deep learning and descriptive rules. Journal of Ambient Intelligence & Humanized Computing.
  - Status: **Accepted**.
  - Impact Factor (JCR 2017): **1.423**
  - Subject Category: **Computer Science, Artificial Intelligence**
  - Rank: **78/132**
  - Quartile: **Q3**

# *What do people think about this monument?* UNDERSTANDING NEGATIVE REVIEWS VIA DEEP LEARNING, CLUSTERING AND DESCRIPTIVE RULES

A PREPRINT

**Ana Valdivia**
Andalusian Research Institute on
Data Science and Computational Intelligence
Universidad de Granada
Granada, Spain 18014
avaldivia@ugr.es

**Eugenio Martínez-Cámara**
Andalusian Research Institute on
Data Science and Computational Intelligence
Universidad de Granada
Granada, Spain 18014
emcamara@decsai.ugr.es

**Iti Chaturvedi**
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore 639798
iti@ntu.edu.sg

**M. Victora Luzón**
Andalusian Research Institute on
Data Science and Computational Intelligence
Universidad de Granada
Granada, Spain 18014
luzon@ugr.es

**Erik Cambria**
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore 639798
cambria@ntu.edu.sg

**Yew-Soon Ong**
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore 639798
asysong@ntu.edu.sg

**Francisco Herrera**
Andalusian Research Institute on
Data Science and Computational Intelligence
Universidad de Granada
Granada, Spain 18014
herrera@decsai.ugr.es

## ABSTRACT

Aspect-based sentiment analysis enables the extraction of fine-grained information, as it connects specific aspects that appear in reviews with a sentiment. Although we detect that the information from these algorithms is very accurate at a local level, it does not contribute to obtain an overall understanding of reviews. To fill this gap, we propose a methodology to portray opinions through the most relevant connections between aspects and polarities. Our methodology combines three off-the-shelf algorithms: (1) deep learning for extracting aspects, (2) clustering for joining together similar aspects and (3) subgroup discovery for obtaining descriptive rules that summarize the sentiment information of set of reviews. Concretely, we aim at depicting negative opinions from three cultural monuments in order to detect those features that need to be improved. Experimental results show that our approach clearly gives an overview of negative aspects, therefore it will be able to attain a better comprehension of opinions.

***K*eywords** Sentiment Analysis · Deep Learning · Aspect Clustering · Subgroup Discovery

# 1 Introduction

In recent years, Sentiment Analysis (SA) has become increasingly popular for processing social media data on online communities, blogs, wikis, microblogging platforms, and other online collaborative media (Cambria, 2016). It is defined as a branch of affective computing research that aims to classify text into either positive or negative, and sometimes also neutral polarities.

While most works approach it as a simple categorization problem, SA is actually a suitcase research problem that requires tackling many NLP tasks, including subjectivity detection (Chaturvedi et al., 2018), concept extraction (Rajagopal et al., 2013), and aspect extraction (Schouten and Frasincar, 2016). Aspect extraction, in particular, is an important subtask of Aspect-Based Sentiment Analysis (ABSA), which focuses on detecting polarities on entities and aspects mentioned within the opinion. ABSA has become very popular due to the fact that it permits obtaining fine-grained information about product features such as product components and attributes.

However, we observe that ABSA methods work at entity level which implies that the information obtained is very specific for each opinion. These approaches permit in depth analysis of opinions pertaining to the features and attributes, but they do not contribute to obtaining an overview of the state of the opinion. We detect the necessity of developing a method that concisely summarizes the content of a set of documents, taking into account the sentiment expressed in it. In this sense, we show that Descriptive Rules (DR) methods can enhance the quality of the information provided by ABSA algorithms through obtaining the most relevant connections between aspects and polarities, in a cross-document scenario.

In this work, we present a novel methodology based on ABSA and DR methods for depicting the content of reviews. We show that DR techniques can describe the content of the text and also providing insights on the negative polarities. Our methodology is based on the following work-flow:

1. **Aspect Extraction.** We extract aspects using a deep learning approach. These models have been known to outperform the state of the art of ABSA task, presenting significantly better accuracy. For this reason, we propose to use the algorithm presented in (Poria et al., 2016).
2. **Aspect Clustering.** Since the same aspect or entity may be referred with different words, we cluster those words that refer to the same aspect.
3. **Subgroup Discovery (SD).** In order to summarize the sentiment information, we propose to apply a DR approach based on the use of a subgroup discovery method (Kavšek and Lavrač, 2006) in order to obtain DRs, which will provide useful insights about negative reviews on the extracted aspects.

We set our experiment on TripAdvisor English reviews of the most popular monuments in Spain: the Alhambra, the Mezquita, and the Sagrada Familia (Valdivia et al., 2017). We focus on the negative reviews because they contain aspects that allow us to identify those features of the monuments that need to be improved. The results clearly show that our approach provides an cross-document overview of the content of all negative reviews. It also detects distinctive aspects of negative polarities which cultural managers have to take into account for improving the visitor's experience within its monument.

The remainder of this paper is organized as follows: Section 2 encompasses a brief introduction to the main concepts for a better understanding of the current work and a succinct review of related works; Section 3 presents the proposed methodology; Section 4 shows the results obtained on the three monuments dataset; We discuss these results on Section 5; finally, Section 6 presents concluding remarks and suggests future research lines.

# 2 Background

This section presents the theoretical concepts necessary to properly comprehend this work. We define sentiment analysis and ABSA in Section 2.1 and Section 2.2 respectively. We describe the use of deep learning for extracting aspects in Section 2.3. We then introduce different methods for extracting DRs, among them Subgroup Discovery methods in Section 2.4. Finally, we present some related works (Section 2.5).

## 2.1 Sentiment Analysis

SA is an area which aims at identifying sentiments in opinions towards a specific entities. More formally, an opinion can be defined as a quintuple `(e, a, s, h, t)`, in which `e` refers to the `entity` of the opinion, `a` to the `aspects` and components of this entity, `s` is the `sentiment` of the opinion, `h` the author or *opinion holder* of the review, and `t` the date when it was expressed (Liu, 2015). Hence, the main target of sentiment analysis is to discover the underlying

sentiment s. For example, in restaurant reviews the entity is the restaurant and the aspects are typical characteristics of the restoration field as: the service, the food, the price, etc.

Since the range of human emotions is wide (Plutchik, 1984), three main categories are considered in sentiment analysis: positive, neutral and negative. There exist some studies that present a binary classification problem, i.e., considering positive and negative polarities. There are other studies that perform a multi-class classification, working at different levels of sentiment intensity: very positive, positive, neutral, negative, very negative. There are also other studies that try to detect figurative expressions in text (irony, sarcasm, etc.) (Nguyen and Jung, 2017).

Sentiment analysis can be divided into three levels of analysis. First, the document level, whose aim is to obtain the sentiment of the whole text. Second, sentence level, whose goal is to detect the sentiment of each sentence. Finally, the most in-depth level is the aspect or entity. This level studies the sentiment of the target of the opinion and obtain very fine-grained sentiment information about the reviews (Schouten and Frasincar, 2016). Taking the cues, in this work we propose to understand monument reviews by analyzing the aspect information.

### 2.2 Aspect-Based Sentiment Analysis

ABSA focuses on extracting aspects and entities that have been evaluated in the reviews and gives a more detailed information about the purpose of the opinion (Schouten and Frasincar, 2016). People tend to review different aspects of the same entity rather than give an opinion of the whole object. For example, if we analyze the following statement about the Alhambra monument:

*"The Alhambra itself was fabulous just such a shame about some of the ticket staff. It is also very crowded with 10,000 visitors per day, so rivers of people moving with you is to be expected."*

We observe that the holder first says that the monument is wonderful, but he then starts to criticize some related aspects like the staff and the high density of people inside the monument. Therefore, the overall sentiment is not clear, but if we evaluate the review at a aspect level, the author shows a positive sentiment towards the Alhambra monument but a negative sentiment towards the Alhambra's staff and its operation.

This task has experienced a constant evolution of its techniques (Schouten and Frasincar, 2016). The first methods were based on setting the most frequent nouns and compound nouns as aspects (Frequency-Based Methods). Hu and Liu (2004) identified product features from customers opinions through rule association algorithms and produced an opinion summarization with the discovered information. This approach was applied in the tourism domain by Marrese-Taylor et al. (2013) where they aimed at extracting aspects from restaurants and hotel reviews. However, these methods do not detect low-frequency aspects, which can also be a key for opinion summarization. Syntax-Based Methods focus on analyzing syntactical relations (Zhao et al., 2010). These methods need to describe a high number of syntactical rules for detecting as many aspects as possible.

### 2.3 Deep Learning for ABSA

Most of the previous works in aspect term extraction have either used supervised learning like conditional random fields (CRFs) (Jakob and Gurevych, 2010; Toh and Wang, 2014) or linguistic patterns (Hu and Liu, 2004). Both of these approaches have their own limitations. Supervised learning strongly depends on manually selected features. Linguistic patterns need to be handcrafted, and they crucially depend on the grammatical accuracy of the sentences. Moreover, language evolves and are rich, making very hard its modeling with rules. In this work, we apply an ensemble of deep learning and linguistics to tackle the problem of aspect extraction in raw text.

In recent years, deep learning has revolutionized a large part of the computer science field. They provide the versatility of supervised learning and do not need to design and previously select a set of features. Furthermore, deep learning models are non-linear supervised classifiers which can fit the data in a more accurately way. Collobert et al. (2011) was the first to introduce the use of Convolutional Neural Networks (CNN) in Natural Language Processing (NLP) tasks. Poria et al. (2016) presented a deep learning-based approach to ABSA, which is built upon two CNN layers combined with a set of linguistic patterns.

### 2.4 Descriptive Rules and Subgroup Discovery

Supervised learning are all those data mining methods that learn a function that maps instances to a set of labeled classes. They are used when the objective is to predict the class of new instances. Unsupervised learning are methods that aim at inferring hidden structures from unlabelled data. In this case, they are conceived as techniques for describing data. These methods analyze the inherent structure of the data, so they are very useful for extracting knowledge.

One of the most popular techniques of unsupervised learning is DR. It is defined as the set of techniques that aim at discovering descriptive knowledge guided by a class variable (Novak et al., 2009). The main objective of DR is to understand the patterns that are conveyed in the data rather than classify instances regarding a class variable.

Although there is a wide range of DR methods (García-Vico et al., 2017; Mihelčić et al., 2017), they can mainly divided in three groups:

- **Contrast Set Mining (CSM):** It was defined by Bay and Pazzani (2001) as the "conjunctions of attributes and values that differ meaningfully in their distributions across groups". The algorithms based on CSM are usually applied for finding robust contrasts on variables that characterize groups in data.
- **Emerging Pattern Recognition (EPR):** It was proposed by Dong and Li (1999) as a technique "to capture emerging trends in time-stamped data, or useful contrasts between data classes". It was lately proposed as a Bayesian approach by Fan and Ramamohanarao (2003). The idea is to discover trends in data with respect to a specific time or class variable.
- **Subgroup Discovery:** It was proposed by Klösgen (1996) and Wrobel (1997), and it was defined as: given a population of individuals and a property of those individuals that we are interested in, find population subgroups that are statistically *most interesting*, for example, are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest. It aims at discovering interesting rules fixing a class label.

Since SD algorithms obtain the best trade-off between generality of rules and precision compared to CSM and ERP (Carmona et al., 2011), we propose to use SD for extracting insights from negative reviews.

More formally, SD is an unsupervised data mining technique that discovers interesting rules regarding the class label (Herrera et al., 2011). This task does not focus on finding complex relations in the data, but attempts to cover instances from data in a comprehensive way. It can be described as conjunctions of features that are characteristic for a selected class. Therefore, it can be defined as condition rule (Novak et al., 2009; Herrera et al., 2011):

$$\texttt{R: } \{\texttt{Subgroup Conditions}\} \longrightarrow \{\texttt{Class}\},$$

where the antecedent is the set of features (`Subgroup Conditions`) that describe the consequent, i.e., the value of the class variable (`Class`). For instance, let `SC` be the set of three monument aspects: `Ticket System` := {0, 1}, `Staff` := {0, 1}, `Wheel Chair Accessible` := {0, 1}. Let `C` be the variable class: the `Sentiment` := {`positive`, `negative`}. As a possible rules we can find:

$$\begin{aligned} \texttt{R}_1\texttt{:}\ \ \{\texttt{Staff = 1}\} &\longrightarrow \{\texttt{Sentiment = negative}\},\\ \texttt{R}_2\texttt{:}\ \ \{\texttt{Wheel Chair Accessible = 1, Staff = 0}\} &\longrightarrow \{\texttt{Sentiment = positive}\},\\ \texttt{R}_3\texttt{:}\ \ \{\texttt{Ticket = 1, Staff = 1}\} &\longrightarrow \{\texttt{Sentiment = negative}\}. \end{aligned}$$

One of the most important facts about SD is the choice of the quality measure for evaluating the rules. The most popular measures in the literature are (Lavrač et al., 2004):

- **Coverage:** Number of instances covered on average. This can be computed as:

$$\mathit{Cov}(\texttt{R}) = \frac{|\texttt{Covered Instances}|}{N},$$

  where `Covered Instances` is the total number of instances that satisfies the subgroup conditions, and $N$ is the total number of instances in the dataset.
- **Support:** Number of instances in the dataset that satisfies the `Subgroup Conditions` and the value of the `Class`. This can be computed as:

$$\mathit{Sup}(\texttt{R}) = \frac{|\texttt{Covered Instances} \cap \texttt{Class}|}{N}.$$

- **Confidence:** Measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. This can be computed as:

$$\mathit{Conf}(\texttt{R}) = \frac{|\texttt{Covered Instances} \cap \texttt{Class}|}{|\texttt{Covered Instances}|}.$$

- **Weighted Relative Accuracy:** This measure is defined as the Weighted Relative Accuracy of a rule and it measures the unusualness of a rule (Lavrač et al., 1999). It can be computed as:

$$WRAcc(\texttt{R}) = \frac{|\texttt{Covered Instances}|}{N}\left(\frac{|\texttt{Covered Instances} \cap \texttt{Class}|}{|\texttt{Covered Instances}|} - \frac{|\texttt{Class}|}{N}\right).$$

More precisely, a rule R has coverage $cov$ if the $cov \cdot 100\%$ of rows in the dataset support Subgroup Conditions. A rule R has support $s$ if the $s \cdot 100\%$ of rows in the dataset contains Subgroup Conditions $\cap$ Class. The rule R holds in the dataset with confidence $c$ if $c \cdot 100\%$ rows in the datasets support Subgroup Conditions also support Class. Therefore, the support is considered as a measure of generality and the confidence as a measure of precision.

Further details about SD methods and applications are in (Carmona et al., 2014; Atzmueller, 2015; García-Vico et al., 2017; Mihelčić et al., 2017; Carmona et al., 2018).

### 2.5 Related Work

This work is presented as an approach for improving cultural experiences. We aim at describing patterns in cultural monument reviews through a methodology that combines SD and ABSA techniques, descriptive and deep learning models. In the literature, we find studies that also combine both areas of knowledge. Li et al. (2015) presented a system for identifying hotel features of interest and understanding customers behavior. They developed an approach based on EPR to detect important changes or trends in travelers' concerns. Hai et al. (2011) proposed an association rule mining approach for identifying implicit features in products on-line reviews based on co-occurrences. They build a set of rules with opinion words and explicit features and then given an opinion with an implicit feature, they assign it the feature of the rule that best fits. Li et al. (2010) developed a rule association method on tourist data of Hong Kong, which gave useful insights about tourist patterns in that city. Poria et al. (2014) proposed a rule-based model for extracting explicit and implicit aspects for product review. The rules were based on parsing dependences like: sentences having subject, or having auxiliary verbs, etc. Their model was fully unsupervised and it outperformed the state of the art. As we can observe, rules models have been always used in sentiment analysis for analyzing features or aspects relations. As far as we know, this work is the first that presents a methodology that combines aspects and sentiments for describing patterns on reviews.

## 3 Methodology for Describing Negative Reviews based on Deep Learning, Clustering and Subgroup Discovery

ABSA is a subtask of sentiment analysis that aims at obtaining fine-grained information about the target of the review. It is able to relate the aspects mentioned in an opinion with a polarity. However, we detect that ABSA approaches are not able to summarize reviews for a better comprehension of the content of text. For this reason, in this work we propose to combine an ABSA algorithm with a SD technique in order to detect and present the most relevant connections between aspects and polarities. In this sense, we propose to combine two powerful tools: an algorithm built upon a deep learning method based on a CNN, and SD method for aggregating information. We propose to build rules that associate a set of aspects with the negative polarity, for example:

$$\texttt{R: } \{\texttt{aspect\_a = 1, aspect\_b = 1}\} \longrightarrow \{\texttt{sentiment = negative}\}.$$

To do so, we propose a work-flow base on three steps (see Figure 1):

1. **Section 3.1**: The first one aims at extracting aspects using a deep learning technique.
2. **Section 3.2**: We then cluster similar aspects in order to represent the same idea within the same feature. To do this, we represent aspects with word embeddings and apply a cluster algorithm over them.
3. **Section 3.3**: Finally, we apply a SD method for extracting aggregated information. We aim at discovering the most relevant aspects of negative reviews through rule-based representations.

### 3.1 Deep Learning: CNN for Extracting Aspects

Deep learning models are non-linear classifiers that are the state of the art in most of NLP tasks. Thus, we use the deep learning method presented in (Poria et al., 2016) for aspect extraction, which is grounded in the use of convolutional layers. The features of an aspect term depend on its surrounding words. Thus, we used a window of five words around

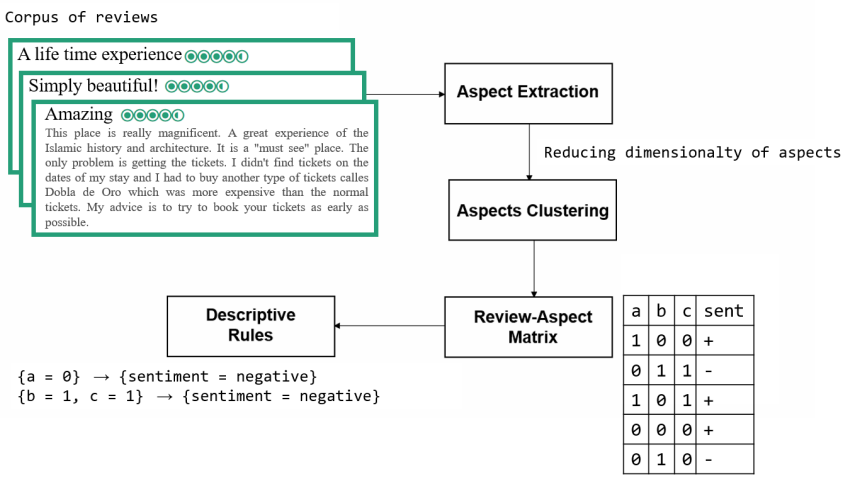



Figure 1: Work-flow of the proposed methodology.

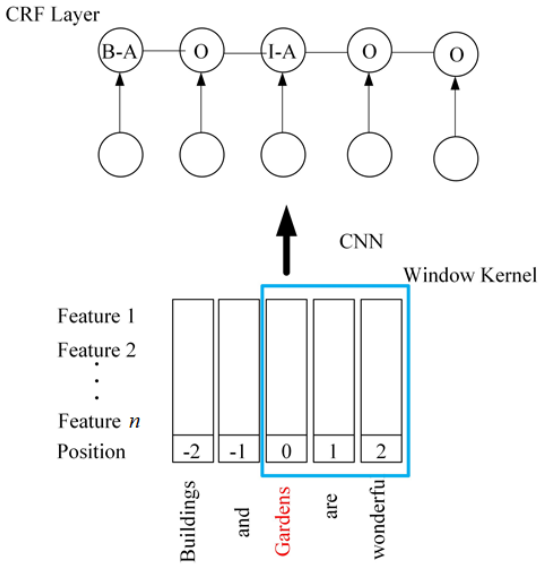



Figure 2: Aspect Extraction using Convolutional Neural Networks.

each targeted word in a sentence, i.e., two words. We formed the local features of that window and considered them to be features of the middle word. Then, the feature vector was fed to a CNN. The network contained one input layer, two convolution layers, two max-pool layers, and a fully connected layer with softmax output. The first convolution layer consisted of 100 feature maps with filter size 2. The second convolution layer had 50 feature maps with filter size 3. The stride in each convolution layer is 1 as we wanted to tag each word. A max-pooling layer followed each convolution layer. The pool size we use in the max-pool layers was 2. We used regularization with dropout on the penultimate layer with a constraint on L2-norms of the weight vectors, with 30 epochs. The output of each convolution layer was computed using a non-linear function; in our case we used $tanh$. The network architecture is formed by one input layer, two convolution layers, two max-pool layers, and a fully connected layer with softmax output. The output of each convolution layer was computed using the $tanh$ function. A set of heuristic linguistic patterns, which leverage on SenticNet (Cambria et al., 2018) and its extensions (Poria et al., 2012a,b), are run on the output of the deep learning model, which enhances the performance of the aspect extraction method.

Figure 2 describes the process of extracting aspects from each sentence. Here we consider three types of labels for each word B-A (Begin Aspect), I-A (Internal Aspect) and O(Non-aspect). CNN use a sliding window of 3 or more words to look for features in the training data. For example here a convolutional kernel of 3 words is shown in blue. Each word is converted to feature representation using pre-trained word vectors. In this diagram we consider $n$ features. Next, to encode the position information of the *aspect* word we include a position feature. For example, when training CNN for the aspect word *Gardens*, we set its position to 0, the word in front is set to 1 and the word before it is set to -1 and so on. The output layer of CNN predicts the aspect category that is B-A, I-A or O for each word. To predict the final aspect category for a word we make use of the predicted labels of all the words in the sentence. This can be done by using a Conditional Random Field (CRF) where the label at each position is predicted using the previous two or three position words. The traditional CRF is unable to model long-range dependencies between words far apart in a sentence. However, CNN is able to capture such dependencies. Hence, the combined model is ideal for aspect term extraction.

### 3.2 Clustering: `K-means` for Clustering Aspects

When people write they do not usually use the same word or expression to convey the same idea. Therefore, the variety of aspects extracted by Poria et al. (2016) method is very large and many of them may refer to the same aspect. For example, we observe that when tourists have an opinion about the *ticket* of a monument they usually refer to it in many different ways:

**ticket** $\longrightarrow$ {*onsite ticket office, senior ticket, ticket area, garden ticket, ticket check points, ticket office, entry ticket, ticket seller, service ticket, machine ticket, ticket tip, ticket staff, ticket box, ticket master, ticket price, ticket process, ticket desks, ...*}.

The great diversity of language implies: (1) we have to face the high dimensionality of aspects, because there are many words that express the same idea; (2) aspects with similar meanings have different representations. To address these problems, we propose to cluster those aspects into the same group to decrease the dimensionality of features and produce a more descriptive summary by using a distributional representation of the aspects.

#### 3.2.1 From words to vectors

We first look up aspects in a set of pre-trained word embeddings. Word embeddings are representations of words as numerical vectors. Mikolov et al. (2013) presented one of the most used set of pre-trained word embeddings, which are widely known as `word2vec`. Levy and Goldberg (2014) generalized this model taking into account the syntactical relations of words within the text. They demonstrated that syntactic contexts capture different information than bag-of-word contexts, so their embeddings (Levy embeddings) show more functional similarities. These models have been widely used as features for NLP and machine learning applications.

We use Levy embeddings[1] as our set of pre-trained word embeddings. For those aspects that are represented as $n$-grams, we compute the mean of the $n$ word embeddings vectors that represents each word (De Boom et al., 2016).

#### 3.2.2 From vectors to clusters

Clustering is defined as the task of joining together a set of objects in such a way that objects in the same group or cluster are more similar to each other than to those in other clusters. There exists a rich variety of cluster analysis algorithms. The main difference between them is their notion of what constitutes a cluster and how to efficiently find them. One of the most popular algorithms is `k-means`, which is conceptually simple and it often shows a well performance in practical applications. This is an iterative clustering algorithm that aims to partition instances into $k$ clusters in which each observation belongs to the cluster with the nearest mean. More formally, it can be express as follow:

*Given a set of elements* $\{w_1, \ldots, w_n\}$*, k-means aims to cluster* $n$ *observations in* $k$ *clusters (*$\{C_1, \ldots, C_k\}$*), minimizing the function:*

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{w \in C_i} \|w - \mu_i\|^2, \tag{1}$$

[1] `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`

algorithm ***APRIORI*** $-$ ***SD*** (*Examples*, *Classes*, *minSup*, *minConf*, *k*)
*Ruleset* = ***APRIORI*** $-$ ***C***(*Examples*, *Classes*, *minSup*, *minConf*) set all example weights of *Examples* to 1
*Majority* = the majority class in *Examples*
*Resultset* = {}
**repeat**
  *BestRule* = rule with the highest weighted relative accuracy value in *Ruleset* (computed using Equation 4)
  *Resultset* = *Resultset* ∪ *BestRule*
  *Ruleset* = *Ruleset* \ *BestRule* decrease the weights of examples covered by *BestRule* (using the example weighting scheme) remove from *Examples* the examples covered more than *k*-times
**until** *Examples* = {} **or** *Ruleset* = {}
**return** *Resultset* = *Resultset* ∪ "*true* → *Majority*"

Figure 3: Pseudocode for the Apriori-SD algorithm (Kavšek and Lavrač, 2006).

*where $\mu_i$ is the mean of points in $C_i$.*

Once we have clustered similar aspects, we build the *review-aspect matrix*. This matrix has the same structure of the well-known document-term matrix: the element $a_{ij}$ is equal to 1 if the $i$th-review contains the $j$th-clustered aspect, otherwise it is equal to 0. We then add to this matrix the polarity of the TripAdvisor user of each review.

### 3.3 Subgroup Discovery: Apriori-SD for Descriptive Rules

Association rules algorithms aim to obtain relations between the variables of the dataset. In this case, variables can appear both in the antecedent and the consequent. However, in SD algorithms the structure of the rules are similar although in this case the consequent is prefixed. That means that association rules algorithms can be geared for SD tasks.

Apriori algorithm was proposed by Agrawal et al. (1996). It is designed for operating in a *transaction* dataset where each elements is defined as an *item*. The aim of this algorithm is to mine frequent *itemsets*, i.e., sets of items that have a minimum support. The strategy that follows can be summarized in two steps: (1) minimum support is applied to find all frequent itemsets and (2) these frequent itemsets and the minimum confidence constraint are used to form rules. Apriori-SD is the SD version of the Apriori algorithm (see Figure 3). It was developed adding several modifications of Apriori C (Jovanoski and Lavrač, 2001) like: implementation of an example weighting scheme in rule post-processing, a modified rule quality function incorporating example weights into the weighted relative accuracy heuristic, etc.

In our case, we apply the Apriori-SD taking into account that:

- *items* are aspects,
- the *transaction* dataset is the review-aspects matrix,
- the *antecedent* is a set of aspects that occur together, and
- the *consequent* is the prefixed sentiment polarity.

Therefore, the idea is to characterize negative opinions by the most frequent aspects. We evaluate the quality of the rules guided by the support and confidence measures.

## 4 Experiments

In this section we evaluate the effectiveness of our proposal. First, we describe the corpora employed (Section 4.1), we analyze the performance of aspects clustering (Section 4.2), and the results of aspect rules (Section 4.3).

### 4.1 Datasets

TripAdvisor is a travel website company providing reviews from traveler experiences about hotels, restaurants and monuments. This website has made up the largest travel community, reaching 630 million unique monthly visitors, and 350 million reviews and opinions covering more than 7.5 million accommodations, restaurants and attractions over 49 markets worldwide.[2] The most interesting feature of this website is the large amount of opinions of million of everyday tourists that it contains. In fact, its opinions have been used as a source of data for many sentiment analysis studies, such as (Valdivia et al., 2017; Lu et al., 2011; Kasper and Vela, 2011; Marrese-Taylor et al., 2013).

[2]Source:`https://tripadvisor.mediaroom.com/uk-about-us`

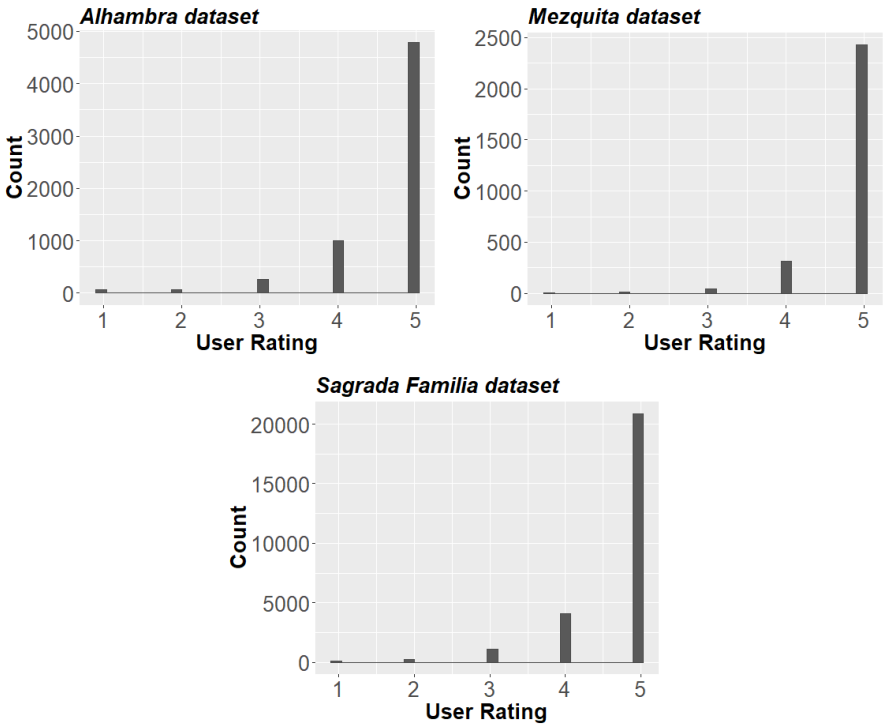



Figure 4: Distribution of TripAdvisor rates.

We based our analysis on three of the main cultural monuments of Spain: the Alhambra (Granada), the Sagrada Familia (Barcelona) and the Mezquita (Córdoba). We gathered 45,301 reviews from July 2012 to June 2016. Table 1 shows the number of reviews per monument, the number of reviews with detected aspects, and the number of extracted aspects by the method described in Section 3.1. As Table 1 shows, Sagrada Familia is the monument that contains more reviews and because of that, more aspects. We removed those reviews without any detected aspect.

We also study the distribution of sentiments on each dataset (see Figure 4). As we observe, the most common score are the 5 points in all the three datasets. Low punctuations are a minority. Therefore, we set the user ratings from 1 to 3 as negative, and from 4 to 5 as positive.

| **Monument** | **Reviews** | **Reviews with Aspects** | **Aspects** |
|---|---|---|---|
| Alhambra | 7,217 | 6,186 | 9,284 |
| Mezquita | 3,526 | 2,802 | 3,688 |
| Sagrada Familia | 34,558 | 26,386 | 18,553 |

Table 1: Summary of Reviews, Reviews with Aspects and Total number of unique Aspects per monument.

If we analyze the distribution of polarities after the aggregation in Table 2, we observe that polarities are highly unbalanced. Positive opinions are much higher than negative ones which means that users tend to evaluate positively their visit to those monuments.

| **Monument** | **Positive** | **Negative** |
|---|---|---|
| Alhambra | 6,781 (93.96 %) | 436 (6.04 %) |
| Mezquita | 3,454 (97.96 %) | 72 (2.04 %) |
| Sagrada Familia | 32,664 (94.52 %) | 1,894 (5.48 %) |

Table 2: Distributions of positive and negative polarities per monument.

| Monument | Total Aspects | Aspects with embeddings | Aspects out-of-vocabulary |
|---|---|---|---|
| Alhambra | 9,284 | 5,430 | 3,854 |
| Mezquita | 3,688 | 2,291 | 1,397 |
| Sagrada Familia | 18,553 | 10,247 | 8,306 |

Table 3: Total number of aspects with embeddings and out-of-vocabulary.

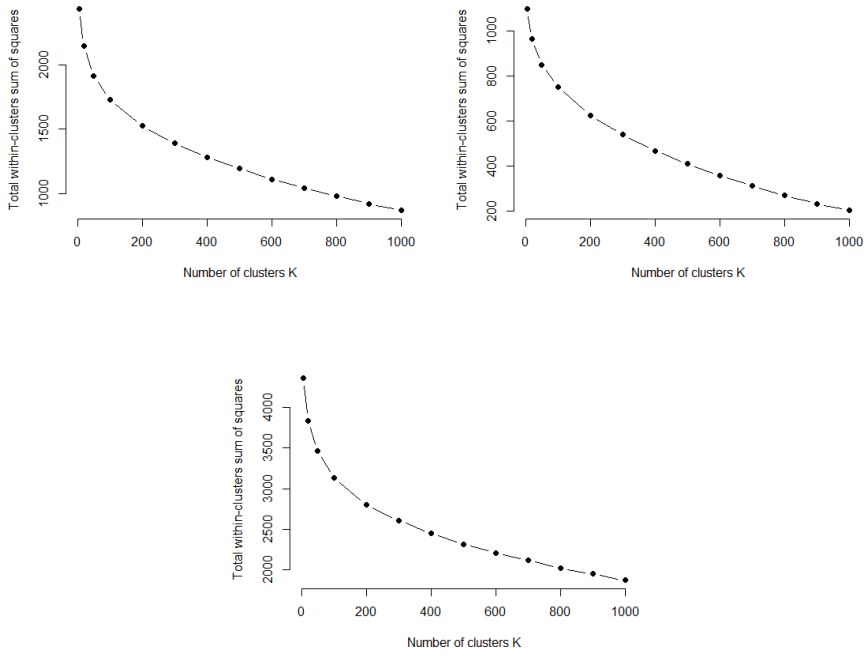



Figure 5: Elbow plots for different $k$ clusters.

### 4.2 When Thousand Words Represents a Common Idea

In this section, we describe the results of the clustering approach. We first used Levi embeddings to represent the extracted aspects. Those aspects that are not in the set of pre-trained word embeddings are not considered. Since $n$-grams aspects do not have a word embedding representation, we built its embedding as the mean of the $n$ vectors of their words. Table 3 shows the total number of aspects with a word embedding representation.

We study the categorization of clusters and select as initial number of $k$ the values 5, 20, 50, 100, 200, 500 and 1000. We observe that setting $k$ with very small values, clusters are formed by a large amount of aspects which may not represent the same concept. In these cases, the clusters are not representative of a common idea.

In order to select the optimal number of clusters, we run the *Elbow method* (Thorndike, 1953). This method analyze the percentage of variance explained as a function of the number of clusters. The percentage of variance explained by the clusters is plotted against the number of clusters. The first clusters add more information, but at some point the marginal gain drops dramatically and draws an angle in the plot. The point on this angles is the correct $k$. Doing this, we find that the best $k$ is 200 (see Figure 5). Table 4 shows some clusters of aspects when $k$ = 200.

Other important advantage of aspect clustering is the feature reduction. As we can observe in Table 1 we extract 9,284, 3,688 and 18,553 of aspects for each monument review, respectively. After the clustering process for $k$ = 200, these aspects are reduced to 4,041, 1,589, 8,229 features, respectively. Note that aspects without embeddings representation are considered as a cluster of just one element.

| Monument | Cluster Label | Cluster Content |
|---|---|---|
| Alhambra | BDA | staff, staff member, local staff, hotel staff, map staff, ground staff, male staff |
| Alhambra | GG | tickets, individual tickets, garden tickets, access tickets, internet tickets |
| Alhambra | BFD | gardens, garden, generalife gardens, gardens water features, garden ticket, beauty of the gardens, general life gardens, main garden, generalife garden |
| Mezquita | BJB | ceiling, floors, marble ceiling, walls, vaulted ceiling, marble floor, roof |
| Mezquita | EG | guard, security guard, security guard berating |
| Mezquita | GG | audio guide, audio guide available, audio guides, map audio guide, auto lingual guide, audio guide facility |
| Sagrada Familia | GI | lift, lift up amp, lift amp lift ride 65m, lift elevator, tower lift, lift down, towers lift up, lift down wait, lift service |
| Sagrada Familia | BAI | ticket online, tickets on-line online tickets, tickets online, entrance tickets online, online ticket, entrance ticket online, prepurchased online tickets, book your ticket online |
| Sagrada Familia | BFJ | shop, souvenir shop, gauds shop, bookshop, citys souvenir shops |

Table 4: Examples of aspects grouped into clusters, with $k = 200$.

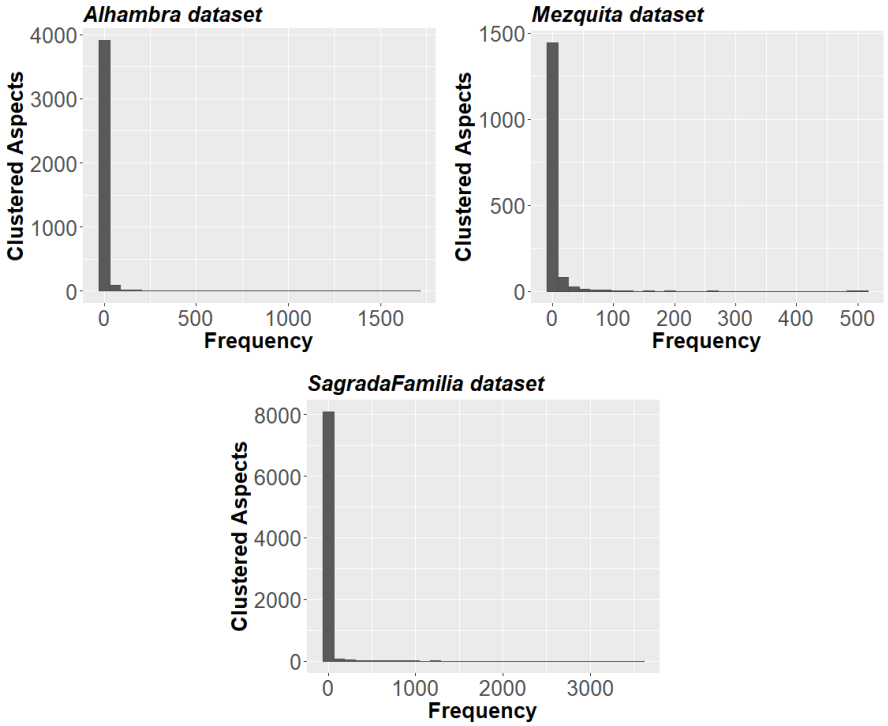



Figure 6: Histograms of clustered aspects of the three monuments.

### 4.3 Depicting Negative Reviews of Cultural Monuments

Before applying SD algorithms, we aim at studying the frequencies of the clustered aspects in all the three datasets. As we observe in Figure 6, the vast majority of aspects occur less than 5 times. Most of these aspects correspond to those words without word vector representations (see Table 3). On the other hand, clustered aspects obtain high frequency values, which makes sense because they represent several aspects.

We also analyze the most frequent clustered aspects in the three datasets. As we observe in Table 5, over the three monuments the most popular words are related with architectural topics. Therefore, we conclude that users tend to describe the monument while they are reviewing their visit in TripAdvisor.

| Monument | Cluster Label | Total | Cluster Content (List of Aspects) |
|---|---|---|---|
| Alhambra | BGB | 1,687 | alhambra, alhambra monument, alhambra ticketing, alhambra opens, alhambra museum, alhambra bookshop, alhambra tours, alhambra walking back |
| Alhambra | BFD | 1,162 | gardens, garden, generalife gardens, generalife garden, rose gardens, garden benches, garden landscaping, walk-in garden |
| Alhambra | BIA | 735 | building, palaces, monuments, palaces fortresses, arab palaces, castles |
| Mezquita | BHF | 509 | mosque, mosque complex, double-arched mosque, mezquita mosque, cordoban mosque |
| Mezquita | BEJ | 489 | architecure, architecture design, architecture buff, cathedrals architecture, cultures architecture |
| Mezquita | FI | 256 | building, views, view, perspectives, viewpoints |
| Sagrada Familia | ID | 3,555 | architecture, skeletal architecture, colors architecture, architecture engineering, designed architecture, catalan architecture |
| Sagrada Familia | HE | 2,798 | church, basilica, cathedral, church building, church glass, working church, gaudis unfinished church |
| Sagrada Familia | BIA | 2,128 | construction, continuous construction, construction noise, construction cranes, construction plaster, construction stones |

Table 5: Top 3 of the most frequent clustered aspects per monument.

Finally, we use the Apriori algorithm version for SD for identifying aspects with negative connotation. We set the consequence of rules as `negative` and let the algorithm discover the clustered aspects of the antecedent side. The Apriori parameters that we set are: minimum length = 2, maximum length = 10, maximum time = 15, minimum support = 0.001 and minimum confidence = 0.01.

Table 6, 7 and 8 present the most relevant rules for the Alhambra, Mezquita and Sagrada Familia, respectively. As we can observe, we obtain very low values for the support, the confidence and the weighted relative accuracy measures. Low support values are due to data sparsity. We observe that zeros, i.e. aspects that do not appear in the document, are predominant in all datasets. Although aspects are grouped into clusters, the ratio of 0 and 1 (not occurring and occurring clustered aspects) is highly unbalanced. In fact, if we compute the percentage of 0 values of the Alhambra, Mezquita and Sagrada Familia datasets, we obtain 99.88%, 99.75%, and 99.96%, respectively. Low support values are driven by data sparsity. We observe that zeros, i.e., aspects that do not appear in the review, are predominant in all datasets. The frequency of a clustered aspect is generally higher in positive instances than in negative instances, hence the confidence gets low rates. Finally, we also observed that the weighted relative accuracy is also close to 0 for all significant rules, which are guided by the low values of the coverage, the confidence and the ratio of negative reviews in the whole dataset.

Analyzing the content of the rules, we detect some interesting patterns in the data. In the Alhambra dataset, the clustered aspects related with *staff, guard* and *cashier* are the most significant rules with length equals to 2. The rule that is highly distinctive of the negative polarity, i.e., obtains the higher confidence is the one with length 3. In this rule, the clusters contains the words *staff* and *price*. That means that TripAdvisor users tend to complain about these two aspects together in negative reviews.

The metrics of the Mezquita and Sagrada Familia datasets are lower than the Alhambra. In those two datasets, all relevant rules obtain a length of 2, which means that they have only one element in the antecedent. We also observe that in these datasets there exist other type of DRs more related with the type and characteristics of monument itself {*garden, architecture, ceiling, arches,* ... }. Consequently, we conclude that TripAdvisor users usually give an objective description of the building which they have visited.

## 5 Discussion

When people give their opinion towards a hotel or a restaurant they usually complain about the service, cleanliness, price, etc. This implies that the type of opinion according to their sentiment can be certainly balanced. We can find either positive or negative reviews. However, when TripAdvisor users review his/her experience visiting a cultural monument, the sentiment is generally positive. We discover that in the three monumental datasets, the positive sentiment represents a vast majority, more than the 90% of total instances (see Figure 4).

| Aspect | Rule | Cov | Sup | Conf | WRAcc |
|---|---|---|---|---|---|
| staff | { BDA = 1 } ⟶ { negative } | 0.03 | $< 0.01$ | 0.28 | $< 0.01$ |
| guard | { BDG = 1 } ⟶ { negative } | $< 0.01$ | $< 0.01$ | 0.38 | $< 0.01$ |
| cashier | { HH = 1 } ⟶ { negative } | $< 0.01$ | $< 0.01$ | 0.27 | $< 0.01$ |
| queue | { BDI = 1 } ⟶ { negative } | $< 0.01$ | $< 0.01$ | 0.15 | $< 0.01$ |
| staff, price | { BDA = 1, BGI = 1 } ⟶ { negative } | $< 0.01$ | $< 0.01$ | 0.48 | $< 0.01$ |

Table 6: Most relevant rules of the Alhambra monument.

| Aspect | Rule | Cov | Sup | Conf | WRAcc |
|---|---|---|---|---|---|
| mosque | { BHF = 1 } ⟶ { negative } | 0.18 | $< 0.01$ | 0.02 | 0 |
| garden | { BHB = 1 } ⟶ { negative } | 0.06 | $< 0.01$ | 0.03 | $< 0.01$ |
| architecture | { BEJ = 1 } ⟶ { negative } | 0.17 | $< 0.01$ | 0.02 | 0 |
| place | { DJ = 1 } ⟶ { negative } | 0.06 | $< 0.01$ | 0.03 | 0 |
| arches | { BIC = 1 } ⟶ { negative } | 0.07 | $< 0.01$ | 0.02 | 0 |

Table 7: Most relevant rules of the Mezquita monument.

As we show in previous section, the measures of precision, confidence, and weighted relative accuracy obtain very low values for discovering negative patterns. The fact is that aspects have very low frequency, which led to datasets with a highly sparsity and those affects to the metrics. However, as we observe in Alhambra's rules (see Table 6, the confidence of these rules are higher than in the others datasets, which means that those aspects are more representative of negative reviews.

Our results also highlight that we obtain a lot of rules satisfying that condition for the Mezquita and Sagrada Familia datasets, but there were not relevant for depicting negative reviews. For instance, we obtain the following rule from the Mezquita:

$$\{\texttt{BIA = 0, ID = 0}\} \longrightarrow \{\texttt{negative}\}.$$

We consider that these type of rules do not contribute to describe negative reviews, since it does not give information about its inner content.

Diving into the data we find that some aspects are labeled as positive because the user scored the overall of the review as positive, but the sentiment related to this aspect does not correspond to the overall of the review. The overall polarity represents a global evaluation of users towards the tourist attraction, but they usually write negative sentences despite reporting 4 or 5 score. For example:

*The **Nasrid palaces** are quite wonderful with intricate plasterwork and tiling and wonderful use of cooling water.The **Generalife gardens** are equally as pleasing. There were two things I thought could be improved about **the site** generally. First **refreshments** are limited to a small kiosk and vending machines. Second it is not geared for **disabled visitors**.*

This review was scored with a 4, so we set it as positive. However, we can find some negativity in the last two sentences. The user is complaining about the lack of refreshments and the adaptability of the monument to people with disabilities. Therefore, these aspects are labelled with the overall sentiment of the review (positive) while they must be labelled as negative. This fact results in low confidence scores.

| Aspect | Rule | Cov | Sup | Conf | WRAcc |
|---|---|---|---|---|---|
| ceiling | { CJ = 1 } ⟶ { negative } | $< 0.01$ | $< 0.01$ | 0.1 | $< 0.01$ |
| natural | { HG = 1 } ⟶ { negative } | 0.04 | $< 0.01$ | 0.07 | $< 0.01$ |
| entry | { CI = 1 } ⟶ { negative } | 0.01 | $< 0.01$ | 0.07 | $< 0.01$ |
| queue | { BCD = 1 } ⟶ { negative } | $< 0.01$ | $< 0.01$ | 0.05 | 0 |
| sagrada familia | { BEJ = 1 } ⟶ { negative } | 0.01 | $< 0.01$ | 0.07 | $< 0.01$ |

Table 8: Most relevant rules of the Sagrada Familia monument.

We also conclude that data sparsity affects measures of generality. The occurrences of aspects or clustered aspects in the datasets are very low, while the number of instances is high. This implies that coverage and support obtain values $< 0.3$, i.e., they appear in less than 30% of instances of the dataset. However, this is a typical issue when dealing with text data.

In spite of these facts, we assess that our methodology is sound for describing the content of reviews. Although we get low values of rules measures, the extracted rules can be used to discover patterns and insights.

## 6 Conclusion

This work presented a novel and effective methodology to describe review data. ABSA algorithms extract information of reviews through aspects, but they do not provide an overview of what the text contains. Consequently, we proposed to combine ABSA methods with DR techniques to represent the content of a text tying aspects to polarities. Our method is based on three steps: (1) Aspect Extraction, (2) Aspects Clustering and (3) Descriptive Rules. We focused on understanding negative reviews from cultural monuments, as they give the most important insights to help cultural managers to enhance visitors' experiences.

The results show that the proposed methodology is effective for describing review data. The main advantage is that it gives a straightforward representation of the content of the text. We were able to describe the content of cultural reviews via DRs. We also concluded that our methodology is able to find out useful information which strengthens the understanding of negative opinions. For instance, Alhambra visitors usually complain about the *staff*, the *ticket* system, and long *queues*. We also discovered through our approach that users tend to describe the elements of the monument visited, which is considered as objective information. We found this fact very interesting because it is not observed when restaurants or hotels reviews are analyzed. However, we detected that the measure of rules are very low, mainly because of the sparsity in text. We also identified that in some cases, using the polarity of the review for all the aspects may lead to a misinterpretation of the text.

There are several directions highlighted by our results. The first one is motivated by the fact that our rules get very low confidence. We propose to create a new corpus with more detailed information about aspects and its polarity. Due to the positive outcome of our methodology, we propose to set this as a baseline, and then compare it with different adaptations using other techniques of aspect extraction, clustering and subgroup discovery. We also propose to extend this methodology to different contexts like: restaurants, hotels or products reviews.

## Acknowledgments

We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript. This research work was supported by the TIN2017-89517-P project from the Spanish Government. Eugenio Martínez-Cámara was supported by the Juan de la Cierva Formación Programme (FJCI-2016-28353) also from the Spanish Government.

## References

E. Cambria, Affective Computing and Sentiment Analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107.

I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, Bayesian network based extreme learning machine for subjectivity detection, Journal of The Franklin Institute 355 (4) (2018) 1780–1797.

D. Rajagopal, E. Cambria, D. Olsher, K. Kwok, A graph-based approach to commonsense concept extraction and semantic similarity detection, in: WWW, 565–570, 2013.

K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Transactions on Knowledge and Data Engineering 28 (3) (2016) 813–830.

S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowledge-Based Systems 108 (2016) 42–49.

B. Kavšek, N. Lavrač, APRIORI-SD: Adapting association rule learning to subgroup discovery, Applied Artificial Intelligence 20 (7) (2006) 543–583.

A. Valdivia, M. V. Luzón, F. Herrera, Sentiment analysis in tripadvisor, IEEE Intelligent Systems 32 (4) (2017) 72–77.

B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.

R. Plutchik, Emotions: A general psychoevolutionary theory, Approaches to emotion 1984 (1984) 197–219.

H. L. Nguyen, J. E. Jung, Statistical approach for figurative sentiment analysis on social networking services: a case study on twitter, Multimedia Tools and Applications 76 (6) (2017) 8901–8914.

M. Hu, B. Liu, Mining and summarizing customer reviews, in: KDD, ACM, 168–177, 2004.

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

Y. Zhao, B. Qin, S. Hu, T. Liu, Generalizing syntactic structures for product attribute candidate extraction, in: HLT-NAACL, Association for Computational Linguistics, 377–380, 2010.

N. Jakob, I. Gurevych, Extracting opinion targets in a single- and cross-domain setting with conditional random fields, in: EMNLP, ACL, 1035–1045, 2010.

Z. Toh, W. Wang, Dlirec: Aspect term extraction and term polarity classification system, in: SemEval, 235–240, 2014.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (Aug) (2011) 2493–2537.

P. K. Novak, N. Lavrač, G. I. Webb, Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining, Journal of Machine Learning Research 10 (Feb) (2009) 377–403.

A. García-Vico, C. Carmona, D. Martín, M. García-Borroto, M. del Jesus, An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (1) (2017) e1231.

M. Mihelčić, S. Džeroski, N. Lavrač, T. Šmuc, A framework for redescription set construction, Expert Systems with Applications 68 (2017) 196 – 215.

S. D. Bay, M. J. Pazzani, Detecting group differences: Mining contrast sets, Data mining and knowledge discovery 5 (3) (2001) 213–246.

G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, in: KDD, ACM, 43–52, 1999.

H. Fan, K. Ramamohanarao, A bayesian approach to use emerging patterns for classification, in: ADC, Australian Computer Society, Inc., 39–48, 2003.

W. Klösgen, Explora: A multipattern and multistrategy discovery assistant, in: Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, 249–271, 1996.

S. Wrobel, An algorithm for multi-relational discovery of subgroups, in: European Symposium on Principles of Data Mining and Knowledge Discovery, Springer, 78–87, 1997.

C. J. Carmona, P. González, M. Del Jesus, M. Navío-Acosta, L. Jiménez-Trevino, Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department, Soft Computing 15 (12) (2011) 2435–2448.

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

N. Lavrač, B. Kavšek, P. Flach, L. Todorovski, Subgroup discovery with CN2-SD, Journal of Machine Learning Research 5 (Feb) (2004) 153–188.

N. Lavrač, P. Flach, B. Zupan, Rule evaluation measures: A unifying view, in: ILP, Springer, 174–185, 1999.

C. J. Carmona, P. González, M. J. del Jesus, F. Herrera, Overview on Evolutionary Subgroup Discovery: Analysis of the Suitability and Potential of the Search Performed by Evolutionary Algorithms, Wiley Int. Rev. Data Min. and Knowl. Disc. 4 (2) (2014) 87–103.

M. Atzmueller, Subgroup discovery, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5 (1) (2015) 35–49, doi:\let\@tempa\bibinfo@X@doi10.1002/widm.1144.

C. Carmona, M. del Jesus, F. Herrera, A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy, Knowledge-Based Systems 139 (2018) 89 – 100.

G. Li, R. Law, H. Q. Vu, J. Rong, X. R. Zhao, Identifying emerging hotel preferences using Emerging Pattern Mining technique, Tourism management 46 (2015) 311–321.

Z. Hai, K. Chang, J.-j. Kim, Implicit feature identification via co-occurrence association rule mining, in: CICLING, Springer, 393–404, 2011.

G. Li, R. Law, J. Rong, H. Q. Vu, Incorporating both positive and negative association rules into the analysis of outbound tourism in Hong Kong, Journal of travel & tourism marketing 27 (8) (2010) 812–828.

S. Poria, E. Cambria, L.-W. Ku, C. Gui, A. Gelbukh, A rule-based approach to aspect extraction from product reviews, SocialNLP 2014 (2014) 28.

E. Cambria, S. Poria, D. Hazarika, K. Kwok, SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings, in: AAAI, 1795–1802, 2018.

S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, T. Durrani, Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis, in: ICSP, vol. 2, 1251–1255, 2012a.

S. Poria, A. Gelbukh, E. Cambria, D. Das, S. Bandyopadhyay, Enriching SenticNet Polarity Scores through Semi-Supervised Fuzzy Clustering, in: ICDMW, 709–716, 2012b.

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 .

O. Levy, Y. Goldberg, Dependency-based word embeddings, in: ACL, vol. 2, 302–308, 2014.

C. De Boom, S. Van Canneyt, T. Demeester, B. Dhoedt, Representation learning for very short texts using weighted word embedding aggregation, Pattern Recognition Letters 80 (2016) 150–156.

R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, et al., Fast discovery of association rules., Advances in knowledge discovery and data mining 12 (1) (1996) 307–328.

V. Jovanoski, N. Lavrač, Classification rule learning with APRIORI-C, in: EPIA, Springer, 44–51, 2001.

B. Lu, M. Ott, C. Cardie, B. K. Tsou, Multi-aspect sentiment analysis with topic models, in: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, IEEE, 81–88, 2011.

W. Kasper, M. Vela, Sentiment analysis for hotel reviews, in: Computational linguistics-applications conference, vol. 231527, 45–52, 2011.

R. L. Thorndike, Who belongs in the family?, Psychometrika 18 (4) (1953) 267–276.

# Bibliography


[ACCF17] Appel O., Chiclana F., Carter J., and Fujita H. (2017) A consensus approach to the sentiment analysis problem driven by support-based iowa majority. *International Journal of Intelligent Systems* 32(9): 947–965.

[AHYL18] Al-Hasan A., Yim D., and Lucas H. C. (2018) A tale of two movements: Egypt during the arab spring and occupy wall street. *IEEE Transactions on Engineering Management* .

[BBL07] Brown J., Broderick A. J., and Lee N. (2007) Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing* 21(3): 2–20.

[BDVJ03] Bengio Y., Ducharme R., Vincent P., and Jauvin C. (2003) A neural probabilistic language model. *Journal of machine learning research* 3(Feb): 1137–1155.

[BKW18] Barnes J., Klinger R., and Walde S. S. i. (2018) Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. *arXiv preprint arXiv:1806.04381* .

[CH12] Cambria E. and Hussain A. (2012) *Sentic computing: Techniques, tools, and applications*, volumen 2. Springer Science & Business Media.

[CHVHA07] Chiclana F., Herrera-Viedma E., Herrera F., and Alonso S. (2007) Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations. *European Journal of Operational Research* 182(1): 383–399.

[DR18] Díaz M. R. and Rodríguez T. F. E. (2018) Determining the reliability and validity of online reputation databases for lodging: Booking. com, tripadvisor, and holidaycheck. *Journal of Vacation Marketing* 24(3): 261–274.

[EF03] Ekman P. and Friesen W. V. (2003) *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

[JC11] Jeacle I. and Carter C. (2011) In tripadvisor we trust: Rankings, calculative regimes and abstract systems. *Accounting, Organizations and Society* 36(4): 293–309.

[Kac86] Kacprzyk J. (1986) Group decision making with a fuzzy linguistic majority. *Fuzzy sets and systems* 18(2): 105–118.

[KFN92] Kacprzyk J., Fedrizzi M., and Nurmi H. (1992) Group decision making and consensus under fuzzy preferences and fuzzy majority. *Fuzzy Sets and Systems* 49(1): 21–31.

[KV11] Kasper W. and Vela M. (2011) Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference*, volumen 231527, pp. 45–52.

[LG13] Lindquist K. A. and Gendron M. (2013) What's in a word? language constructs emotion perception. *Emotion Review* 5(1): 66–71.

[LG14] Levy O. and Goldberg Y. (2014) Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volumen 2, pp. 302–308.

[Liu12] Liu B. (2012) Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1): 1–167.

[Liu15] Liu B. (2015) *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

[M+97] Mitchell T. M. *et al.* (1997) Machine learning. wcb.

[MHK14] Medhat W., Hassan A., and Korashy H. (2014) Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5(4): 1093–1113.

[MSB+14] Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., and McClosky D. (2014) The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60.

[MSC+13] Mikolov T., Sutskever I., Chen K., Corrado G. S., and Dean J. (2013) Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

[MTVBM14] Marrese-Taylor E., Velásquez J. D., and Bravo-Marquez F. (2014) A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications* 41(17): 7764–7775.

[O'C08] O'Connor P. (2008) User-generated content and travel: A case study on tripadvisor. com. *Information and communication technologies in tourism 2008* pp. 47–58.

[PCG16] Poria S., Cambria E., and Gelbukh A. (2016) Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108: 42–49.

[PFML16] Pozzi F. A., Fersini E., Messina E., and Liu B. (2016) *Sentiment analysis in social networks*. Morgan Kaufmann.

[PLV02] 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*, pp. 79–86. Association for Computational Linguistics.

[QGLS85] Quirk R., Greenbaum S., Leech G., and Svartvik J. (1985) A comprehensive grammar of the english language.

[RAG+16] Ribeiro F. N., Araújo M., Gonçalves P., Gonçalves M. A., and Benevenuto F. (2016) Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5(1): 1–29.

[RB16] Rojas-Barahona L. M. (2016) Deep learning for sentiment analysis. *Language and Linguistics Compass* 10(12): 701–719.

[RFN17] Rosenthal S., Farra N., and Nakov P. (2017) Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518.

[ST11] Shariff A. F. and Tracy J. L. (2011) What are emotion expressions for? *Current Directions in Psychological Science* 20(6): 395–399.

[TBT+11] Taboada M., Brooke J., Tofiloski M., Voll K., and Stede M. (2011) Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2): 267–307.

[TZ18] Tang D. and Zhang M. (2018) Deep learning in sentiment analysis. In *Deep Learning in Natural Language Processing*, pp. 219–253. Springer.

[WBCB17] Wehrmann J., Becker W., Cagnini H. E., and Barros R. C. (2017) A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pp. 2384–2391. IEEE.

[Yag88] Yager R. R. (1988) On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics* 18(1): 183–190.

[YF99] Yager R. R. and Filev D. P. (1999) Induced ordered weighted averaging operators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29(2): 141–150.

[YHPC18] Young T., Hazarika D., Poria S., and Cambria E. (2018) Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* 13(3): 55–75.

[Zad83] Zadeh L. A. (1983) A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with applications* 9(1): 149–184.

[ZWL18] Zhang L., Wang S., and Liu B. (2018) Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* page e1253.