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 β 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., ±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.

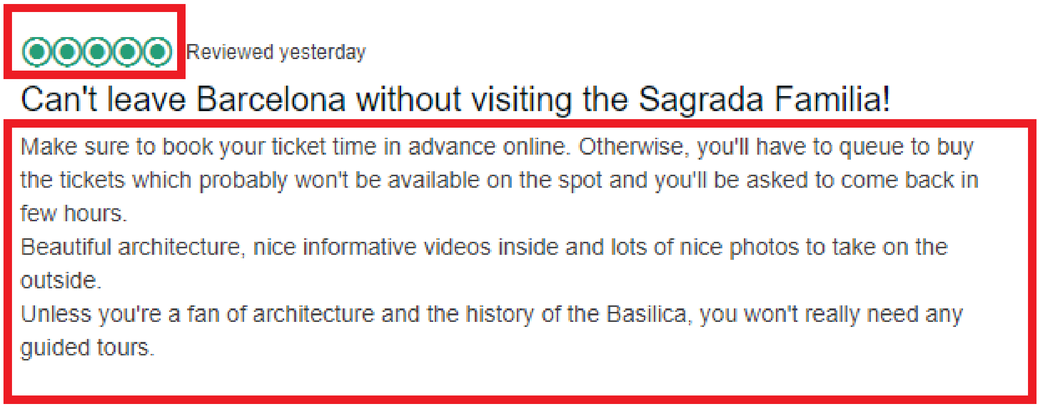
Review Title: A main title of the text.

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).



Reviewed yesterday

Can't leave Barcelona without visiting the Sagrada Familia!

Make sure to book your ticket time in advance online. Otherwise, you'll have to queue to buy the tickets which probably won't be available on the spot and you'll be asked to come back in few hours.
Beautiful architecture, nice informative videos inside and lots of nice photos to take on the outside.
Unless you're a fan of architecture and the history of the Basilica, you won't really need any guided tours.

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 × 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).

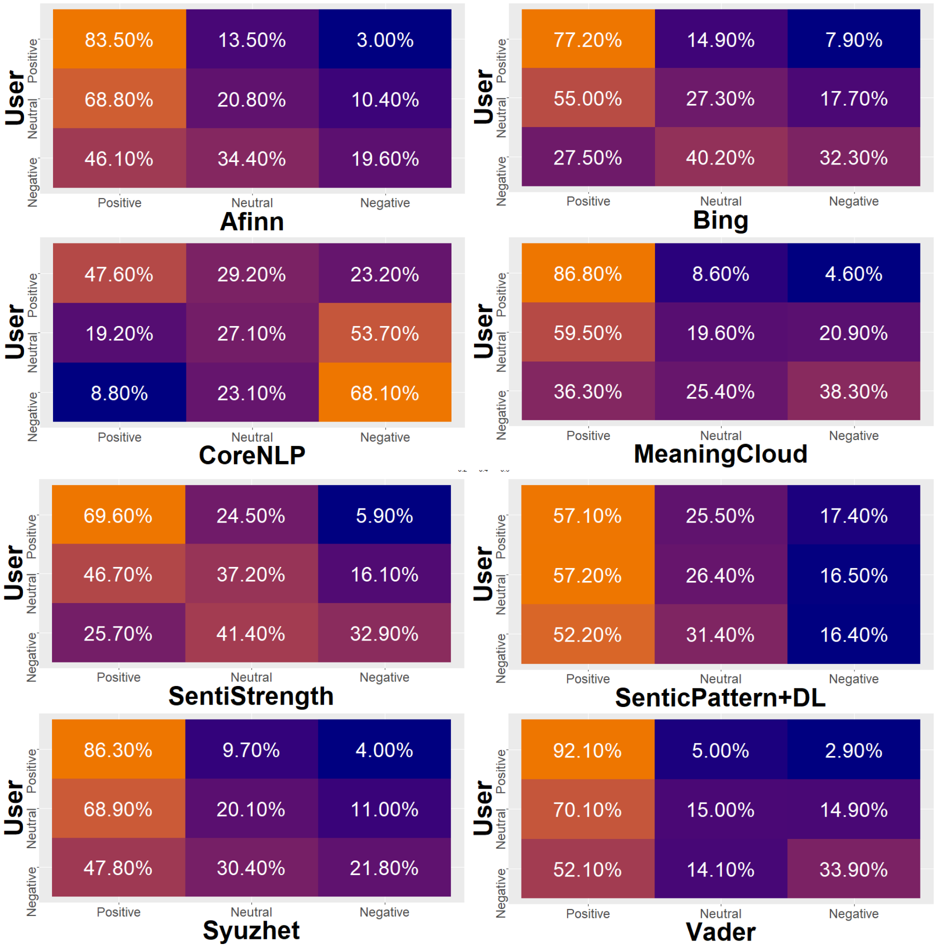


Figure 2: Percentage of matching between Users (rows) and SAMs (columns) polarities. The values are the average over the six monuments. A more orangeade color on cells indicates higher correlation, bluer lower correlation.

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 nth 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 ith-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 kth-*Normalized SAM Polarity* of the ith-opinion and $y \in [0,1]$.
- β, is the parameter to control the SAMs polarity influence and $\beta \in \mathbb{R}^+$.
- p_i^{USER} is the User Polarity of the ith-opinion.
- $p_i^{SAM_k}$ is the kth-SAM Polarity of the ith-opinion.

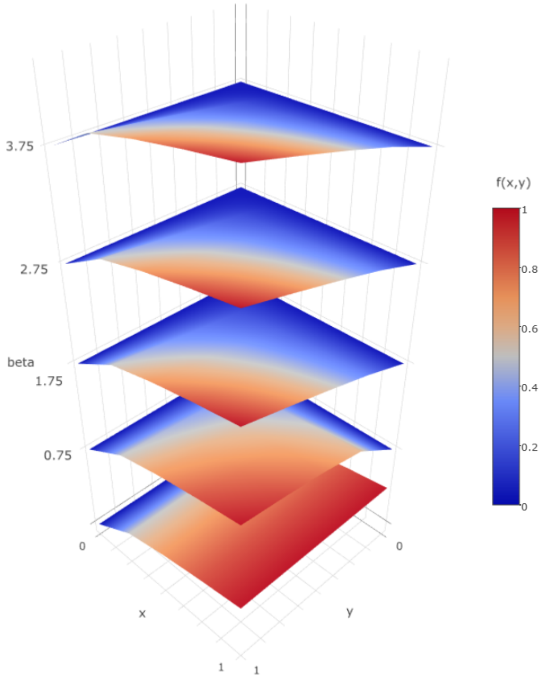


Figure 3: Distribution of the Polarity Aggregation Model for different β values (0, 0.75, 1.75, 2.75 and 3.75). Bluer colors represent more negative aggregated polarities, more orange colors more positive aggregated polarities.

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 β 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 β contained more red, which means that it gets more positive scores. As we increase the value of β, 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 β 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 β, 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 β

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 β 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 β also in Table 12. For low β 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.

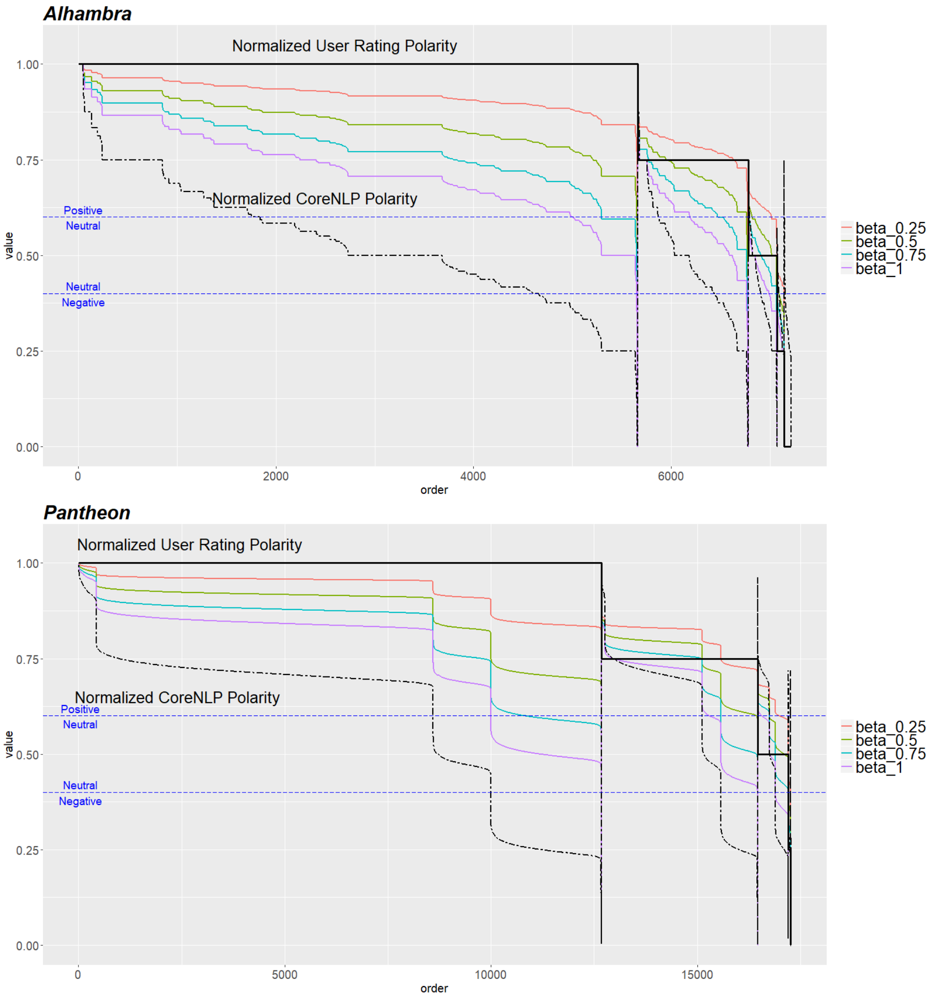


Figure 4: Different Polarity Aggregation Models taking account beta's values. Reviews are sorted on the x label in ascending order, from most positive (left) to most negative (right). The thick line represents the Normalized User Polarity. The two-dash line represents the Normalized CoreNLP Polarities.

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.

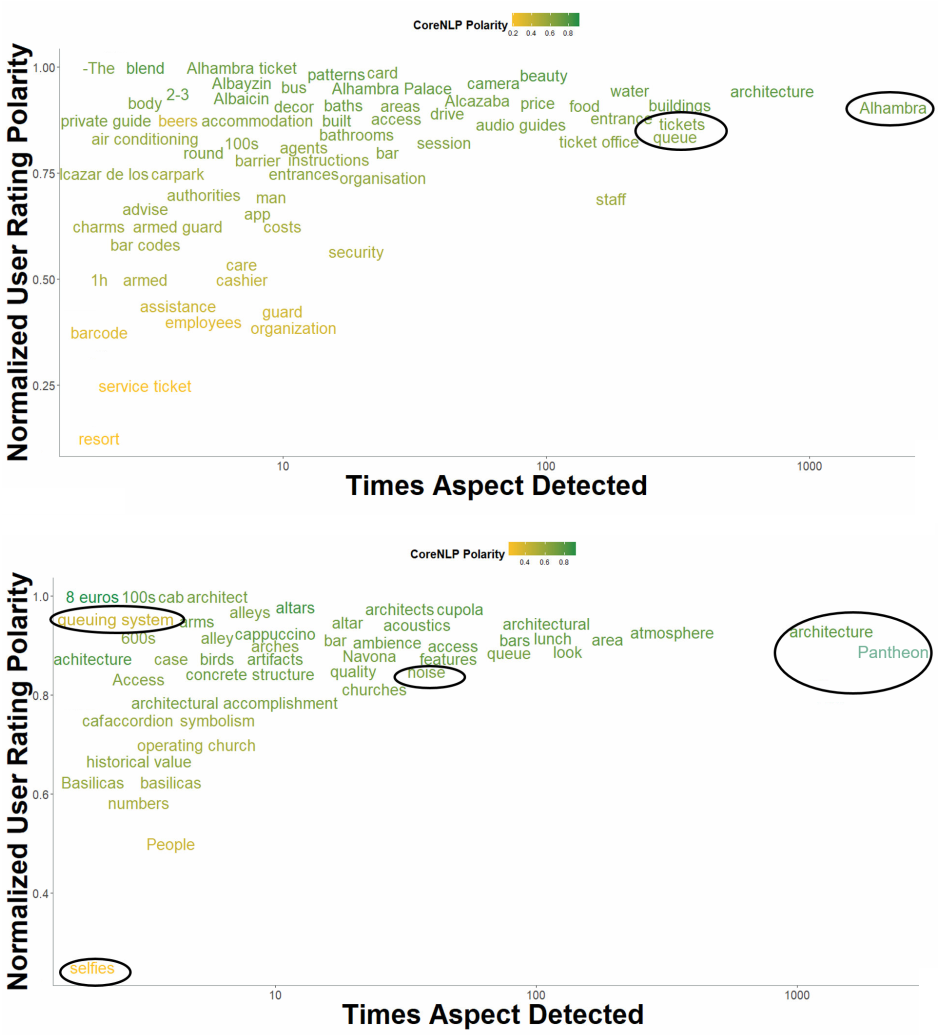


Figure 5: This aspect map represents times that an aspect is detected (x axis) taking User Polarity (y axis) and CoreNLP Polarity (color scale). Alhambra (top) and Pantheon (bottom).

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.

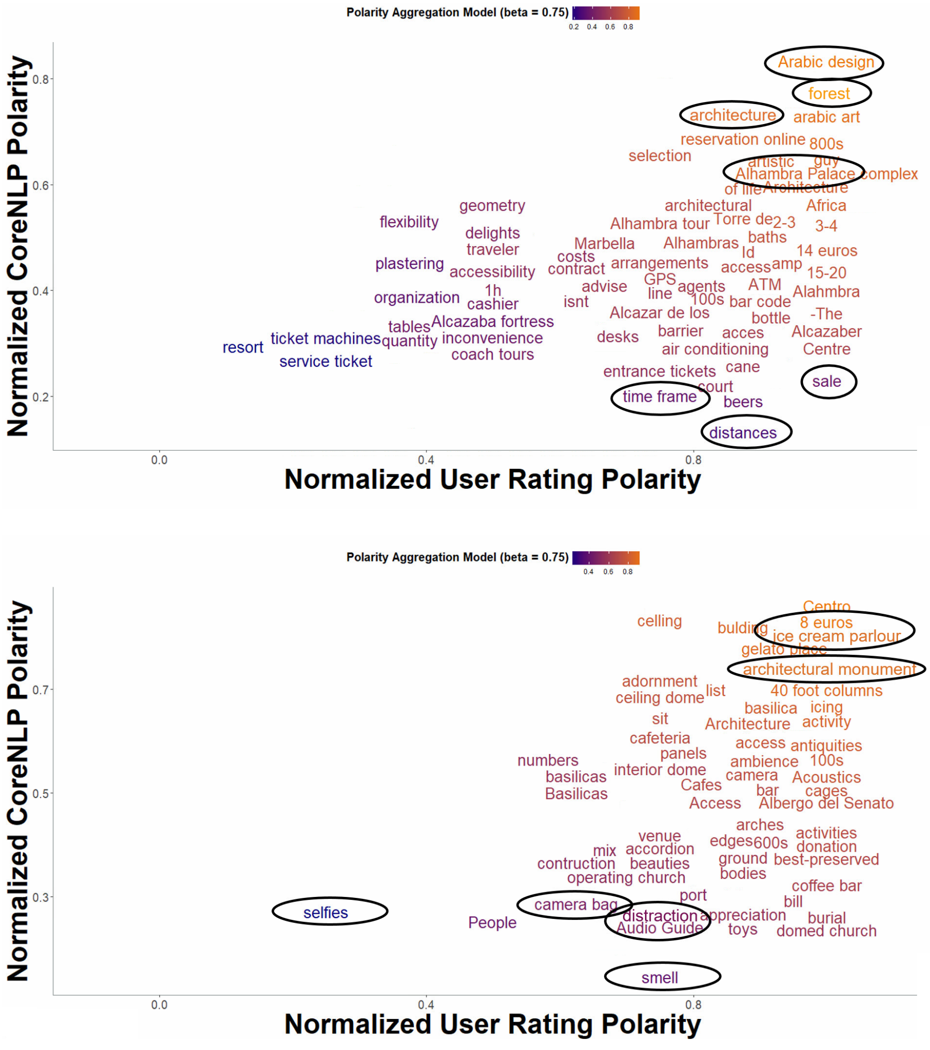


Figure 6: This aspects map represents the mean of each polarity of each aspect. From left to right it goes from negative to more positive depending on the User Polarity. From bottom to top goes from negative to more positive depending on CoreNLP Polarity. From more blue to more orange goes from more 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 β 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**	$\beta = \mathbf{0.75}$
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**	$\beta = \mathbf{0.75}$
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, β. 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)