Neurocomputing 353 (2019) 3–16

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Inconsistencies on TripAdvisor reviews: A unified index between users and Sentiment Analysis Methods

Ana Valdivia[a,*], Emiliya Hrabova[b], Iti Chaturvedi[c], M. Victoria Luzón[a], Luigi Troiano[b], Erik Cambria[c], Francisco Herrera[a]

[a] Andalusian Research Institute on Data Science and Computational Intelligence (DaSCI), University of Granada, Granada 18071, Spain
[b] Department of Engineering, University of Sannio, Italy
[c] School of Computer Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 9 January 2018
Revised 20 August 2018
Accepted 17 September 2018
Available online 13 March 2019

Keywords:
Sentiment analysis
Cultural monuments
E-tourism
Polarity aggregation
Aspect based sentiment analysis

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.

1. Introduction

Sentiment Analysis (SA), also referred to as Opinion Mining, is a branch of affective computing research [48] 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 [14,40,41,47]. Several teams have developed algorithms, Sentiment Analysis Methods (SAMs), capable of automatically detecting the underlying sentiment of a written review [28,29,42]. 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 [6,46].

In past works, we shown the problem of using the TripAdvisor Bubble Rating, which we refer to as the *User Polarity* [39]. 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?"*

[*] Corresponding author.
E-mail addresses: avaldivia@ugr.es (A. Valdivia), ITI@ntu.edu.sg (I. Chaturvedi), luzon@ugr.es (M.V. Luzón), cambria@ntu.edu.sg (E. Cambria), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.neucom.2018.09.096
0925-2312/© 2019 Elsevier B.V. All rights reserved.

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 ith opinion *entity*, a_{ij} is the jth *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 kth *opinion holder* or reviewer and t_l is lth *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 [44], polarity classification [23], opinion summarization [24], sarcasm detection [25], word sense disambiguation [26], opinion spam detection [27], 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 [28,29].

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.

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

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

SAM	Group	Numerical output	Reference
Afinn	LD	$\{-5, \ldots, 5\}$	[33]
Bing	LD	$\{-1, 0, 1\}$	[11]
CoreNLP	DL	$\{0, 1, 2, 3, 4\}$	[17,19]
MeaningCloud	ML	$[0, 1] \in \mathbb{R}$	[38]
SentiStrength	LD & ML	$\{-1, 0, 1\}$	[31,32]
SenticPattern+DL	DL	$\{0, 1, 2\}$	[34,35]
Syuzhet	LD & ML	$[0, 1] \in \mathbb{R}$	[13]
VADER	LD	$[-1, 1] \in \mathbb{R}$	[30]

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 [45], 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 [37]. 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 tth tag having ith 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 Sections 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[1] is one of the most popular travel social network websites [43] founded in 2000. This Web 2.0 contains 570 million reviews about 7.3 million restaurants, hotels and attractions over the world[2]. 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 [39]. Examples of works analyzing hotels reviews are [1,3,4,6,7,16,18,20]. Restaurant reviews are analyzed in [7,9,22]. Monument reviews are analyzed in [36,39].

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 [2,8,12,21].

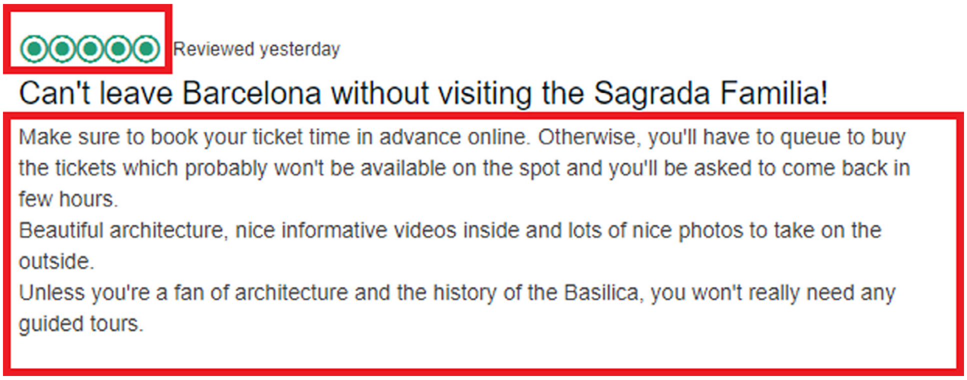
3.2. Web scraping

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.

[1] https://www.TripAdvisor.com

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



●●●●● 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.

Fig. 1. Information of a review in TripAdvisor. For this study, we analyze the bubble scale and the text of the review.

Table 2
Summary of text properties of the six datasets.

	Reviews	Words	Sentences	Avg. # words	Avg. # sentences	Avg. User Polarity
Alhambra	7217	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	3526	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

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 Fig. 1).

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.

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.

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

User Polarity	Positive	Neutral	Negative
Alhambra	6781	293	143
Grand Canal	13,832	548	104
Mezquita de Córdoba	3454	55	17
Pantheon	23,635	1087	107
Sagrada Familia	32,664	1443	451
Trevi Fountain	19,515	3363	2513

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

In this section, we present an extended study of [39]. 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),

Table 5
Distribution of polarities of monuments reviews. Bing.

Bing Polarity	Positive	Neutral	Negative
Alhambra	3310	1252	2655
Grand Canal	12,531	1505	448
Mezquita de Córdoba	1918	642	966
Pantheon	22,235	2085	509
Sagrada Familia	16,541	6644	11,373
Trevi Fountain	18,320	4806	2265

Table 6
Distribution of polarities of monuments reviews. CoreNLP.

CoreNLP Polarity	Positive	Neutral	Negative
Alhambra	3154	1143	2920
Grand Canal	7283	4483	2718
Mezquita de Córdoba	1992	577	957
Pantheon	14,491	7168	3170
Sagrada Familia	17,561	6007	10,990
Trevi Fountain	10,281	8134	6976

Table 7
Distribution of polarities of monuments reviews. MeaningCloud.

MeaningCloud Polarity	Positive	Neutral	Negative
Alhambra	6050	730	437
Grand Canal	12,458	1284	742
Mezquita de Córdoba	3062	290	174
Pantheon	22,487	1572	770
Sagrada Familia	28,124	3998	2436
Trevi Fountain	19,379	3139	2873

Table 8
Distribution of polarities of monuments reviews. SentiStrength.

SentiStrength Polarity	Positive	Neutral	Negative
Alhambra	5277	1341	599
Grand Canal	8777	5153	554
Mezquita de Córdoba	2674	585	267
Pantheon	17,476	6584	769
Sagrada Familia	23,964	6880	3714
Trevi Fountain	14,490	8715	2186

Table 9
Distribution of polarities of monuments reviews. Syuzhet.

Syuzhet Polarity	Positive	Neutral	Negative
Alhambra	5423	1252	2655
Grand Canal	13,000	1176	308
Mezquita de Córdoba	2704	466	356
Pantheon	22,925	1601	303
Sagrada Familia	25,379	4805	4374
Trevi Fountain	19,722	4211	1458

Table 10
Distribution of polarities of monuments reviews. VADER.

VADER Polarity	Positive	Neutral	Negative
Alhambra	6505	362	350
Grand Canal	13,368	753	363
Mezquita de Córdoba	3206	200	120
Pantheon	23,319	1042	468
Sagrada Familia	30,485	2450	1623
Trevi Fountain	20,979	2093	2319

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 Fig. 1 where user rates Sagrada Familia with 5 bubbles (positive opinion) but there are sentences with a negative polarity within the same opinion.

Fig. 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).

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 [36].

As can be hinted from Fig. 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 indi-

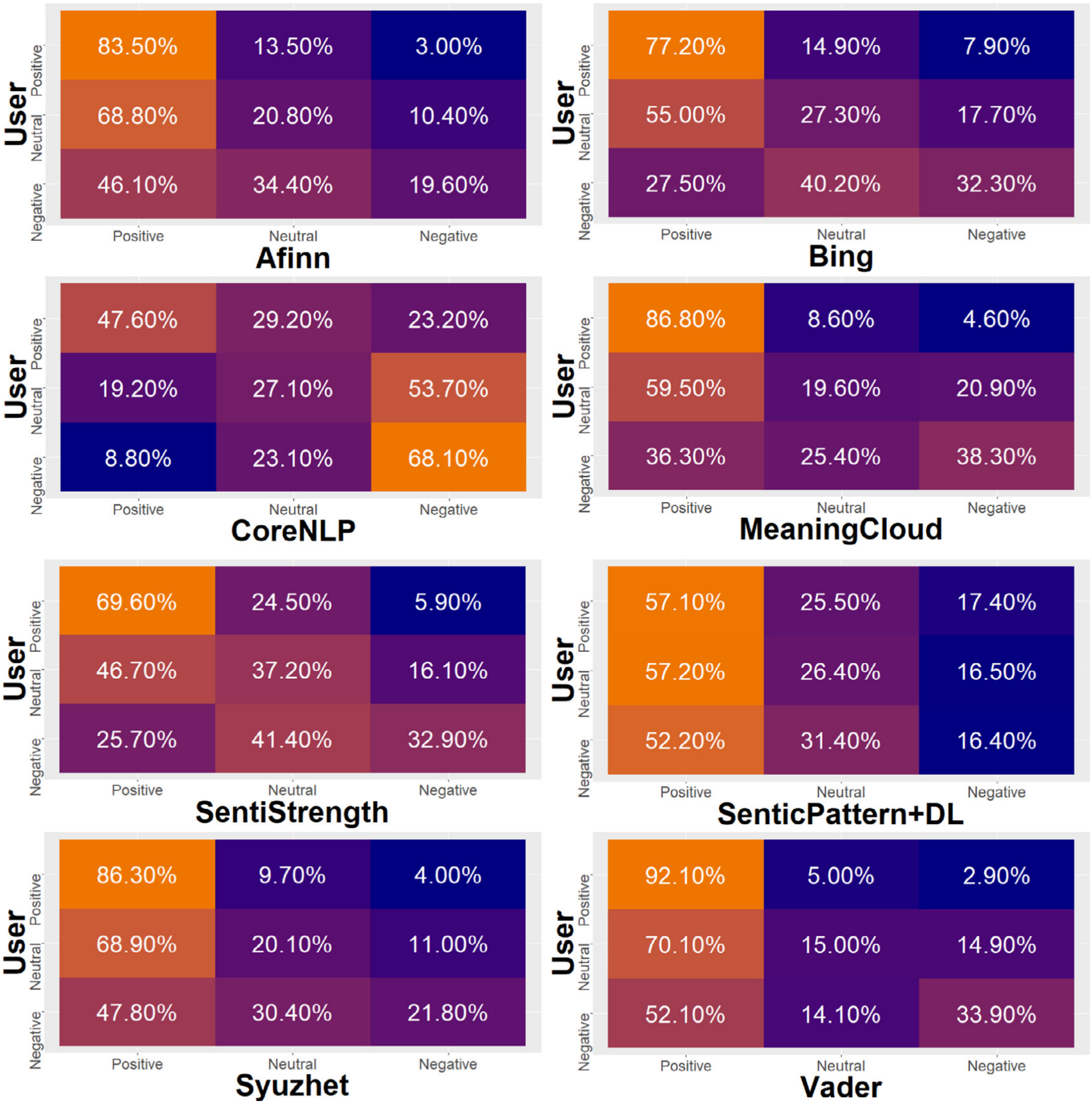


Fig. 2. Percentage of matching between Users (rows) and SAMs (columns) polarities. The values are the average over the six monuments. A more orangeade color on cells indicates higher correlation, bluer lower correlation.

cates the central tendency by using the product of their values and it is defined as the nth root of the product n numbers.[3] 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.

In Fig. 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 $\beta < 1 \Rightarrow f(x,y) > \sqrt{xy}$. 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.

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

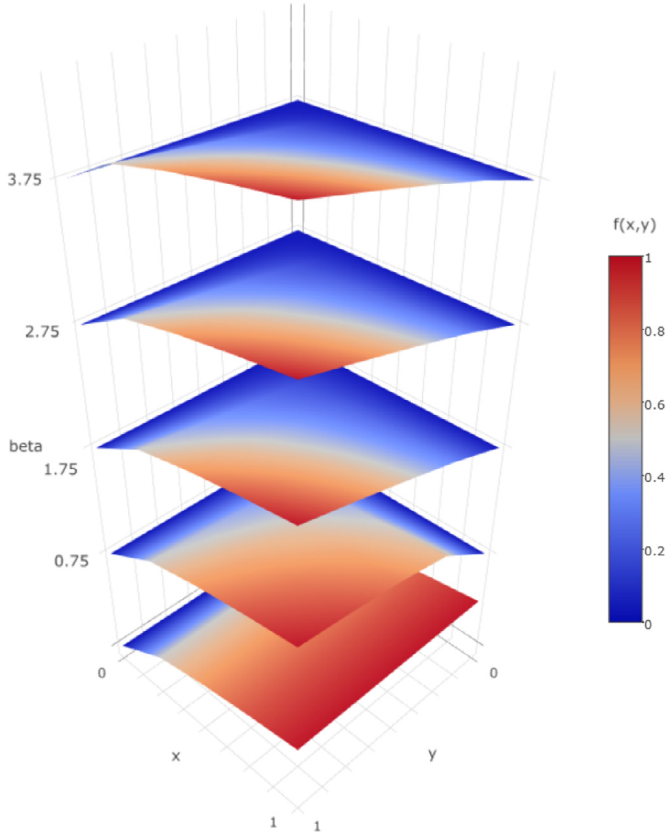


Fig. 3. Distribution of the Polarity Aggregation Model for different β values (0, 0.75, 1.75, 2.75 and 3.75). Bluer colors represent more negative aggregated polarities, more orange colors more positive aggregated polarities.

- If $\beta \geq 1 \Rightarrow f(x, y) \leq \sqrt{x}y$. In that case, the value of y gains relevance in the final score. If we observe the top surfaces on Fig. 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. Fig. 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.

Fig. 4(top): In the Alhambra's dataset, from the 1st to the 5660-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.

Fig. 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 2000-th row, in this case, after the 7500-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

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

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 Fig. 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.

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 [37]. We then

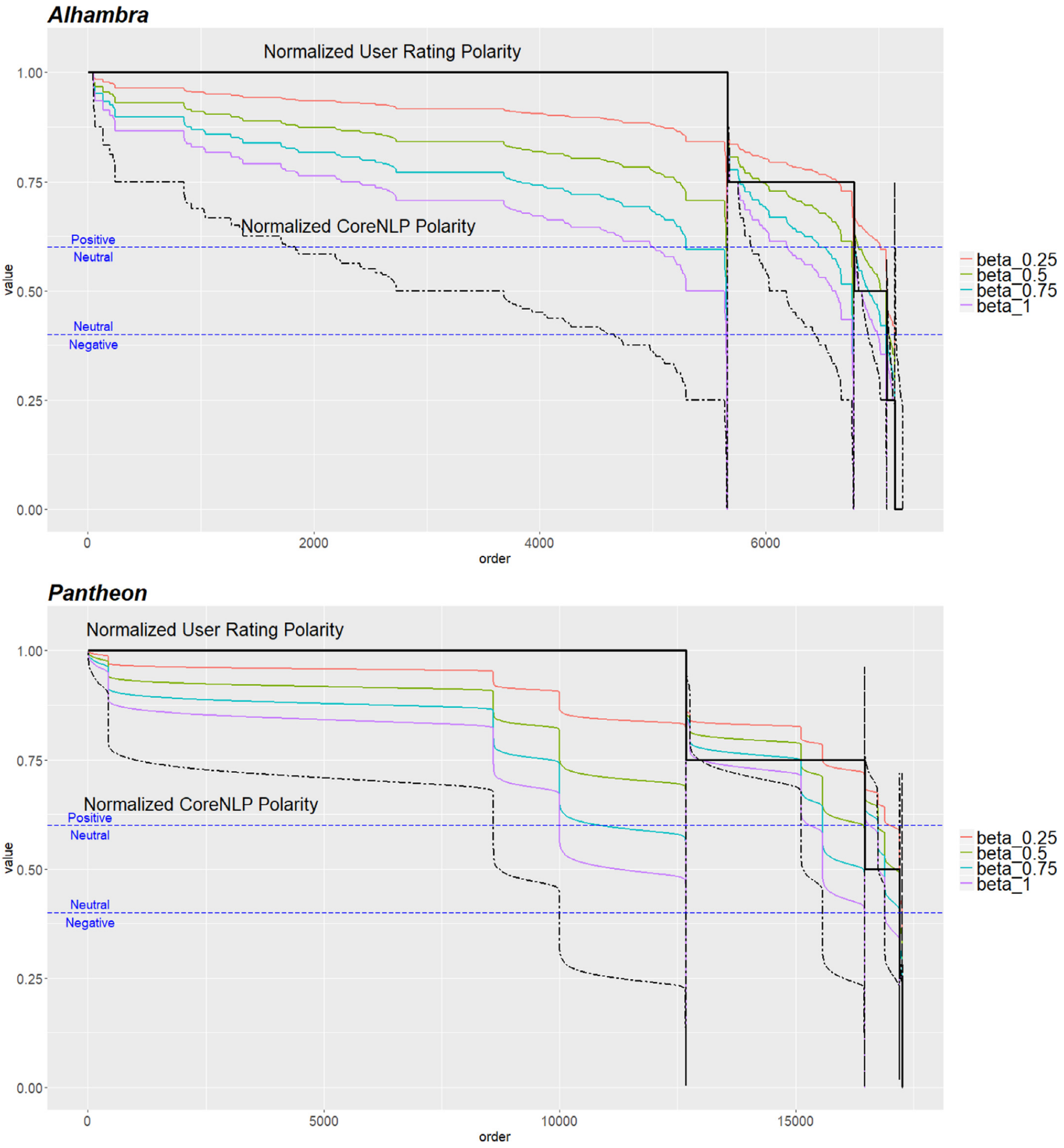


Fig. 4. Different Polarity Aggregation Models taking account beta's values. Reviews are sorted on the x label in ascending order, from most positive (left) to most negative (right). The thick line represents the Normalized User Polarity. The two-dash line represents the Normalized CoreNLP Polarities.

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 Fig. 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. Fig. 5 shows the relationship between User and CoreNLP Polarity on Alhambra's and Pantheon's aspects appearing at least twice.

Fig. 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*

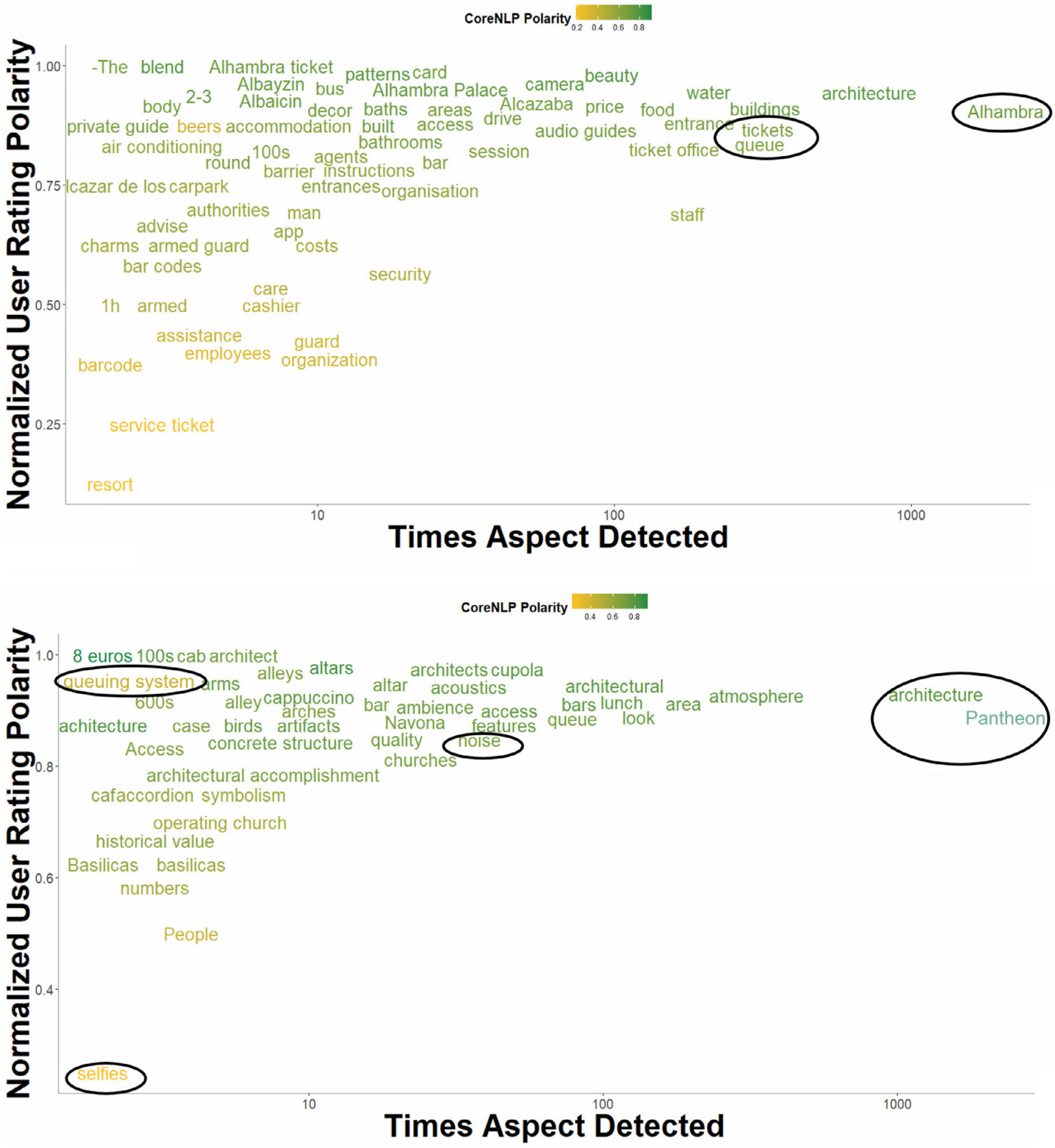


Fig. 5. This aspect map represents times that an aspect is detected (*x* axis) taking User Polarity (y axis) and CoreNLP Polarity (color scale). Alhambra (top) and Pantheon (bottom).

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.

Fig. 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*

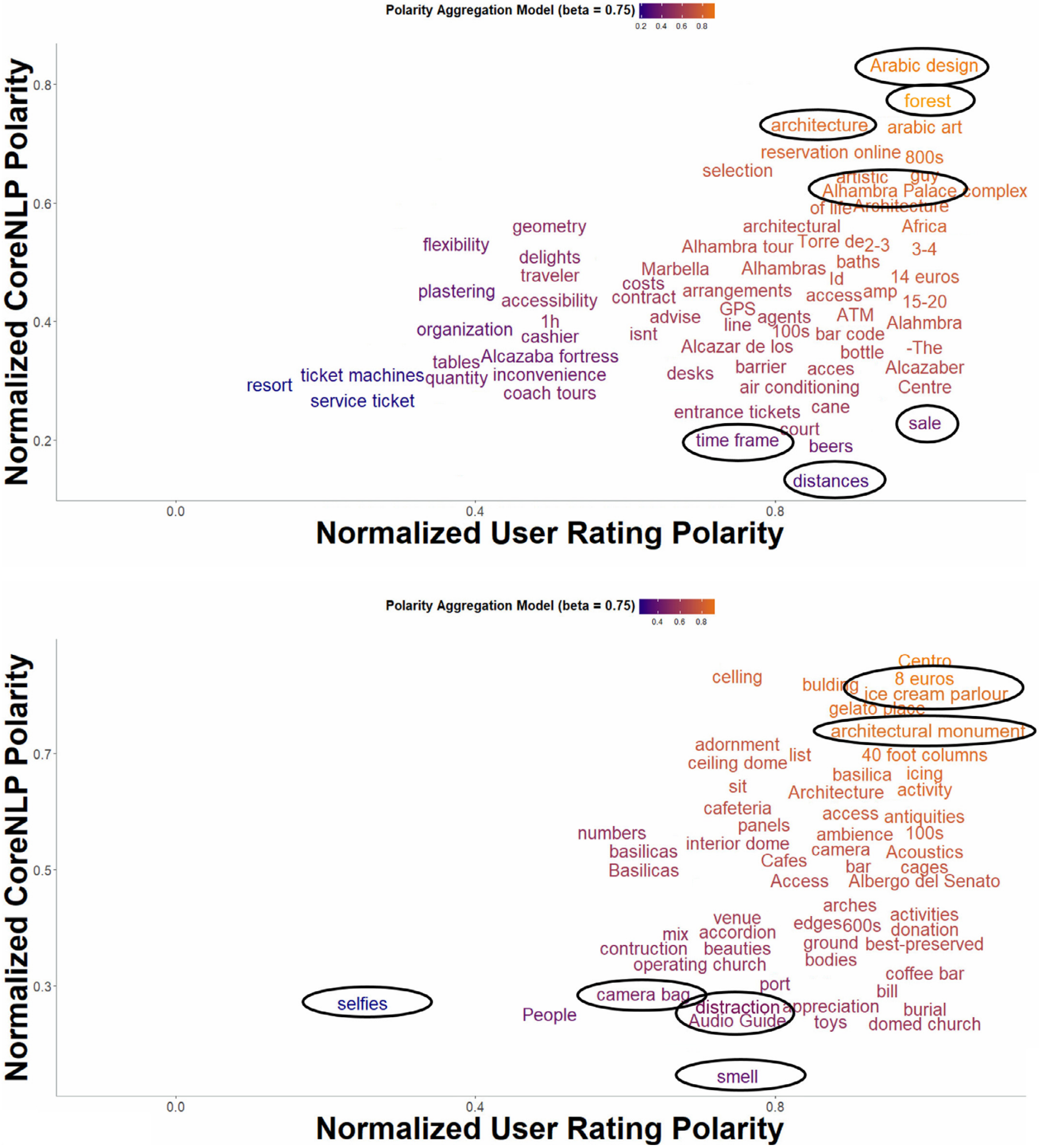


Fig. 6. This aspects map represents the mean of each polarity of each aspect. From left to right it goes from negative to more positive depending on the User Polarity. From bottom to top goes from negative to more positive depending on CoreNLP Polarity. From more blue to more orange goes from more negative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

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.

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

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 min 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

case, Fig. 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.

Fig. 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.

Fig. 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 Fig. 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. Fig. 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 Fig. 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 Fig. 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.

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

- *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 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.

Acknowledgments

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

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

[3] S. Baccianella, A. Esuli, F. Sebastiani, Multi-facet rating of product reviews, in: Proceedings of the European Conference on Information Retrieval, Springer Berlin Heidelberg, 2009, pp. 461–472.

[4] L. Banic, A. Mihanovic, M. Brakus, Using big data and sentiment analysis in product evaluation, in: Proceedings of the Thirty-sixth International Convention on Information and Communication Technology Electronics and Microelectronics, IEEE, 2013, pp. 1149–1154.

[5] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the Twenty-fifth International Conference on Machine Learning, Association for Computing Machinery, 2008.

[6] W. Duan, Q. Cao, Y. Yu, S. Levy, Mining online user-generated content: using sentiment analysis technique to study hotel service quality, in: Proceedings of the Forty-sixth Hawaii International Conference on System Sciences (HICSS), IEEE, 2013, pp. 3119–3128.

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

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

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

[10] M. Hu, B. Liu, Mining opinion features in customer reviews, in: Proceedings of the Conference on American Association for Artificial Intelligence, 4, 2004, pp. 755–760. 4

[11] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.

[12] I. Jeacle, C. Carter, In TripAdvisor we trust: rankings, calculative regimes and abstract systems, Account. Organ. Soc. 36 (4) (2011) 293–309.

[13] M. Jockers, Syuzhet: extracts sentiment and sentiment-derived plot arcs from text, R package version 1.0.0, 2016.

[14] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015.

[15] B. Liu, Sentiment analysis and subjectivity, in: Handbook of Natural Language Processing, 2, Cambridge University Press, 2010, pp. 627–666.

[16] B. Lu, M. Ott, C. Cardie, B.K. Tsou, Multi-aspect sentiment analysis with topic models, in: Proceedings of the Eleventh International Conference Data Mining Workshops (ICDMW), IEEE, 2011, pp. 81–88.

[17] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, D. McClosky, The stanford coreNLP natural language processing toolkit, in: Proceedings of

the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.

[18] A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Natural Language Processing and Text Mining, Springer, London, 2007, pp. 9–28.

[19] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the Conference on Empirical methods in Natural Language Processing (EMNLP), 2013. 1631, 1642

[20] I. Titov, R.T. McDonald, A joint model of text and aspect ratings for sentiment summarization, in: Proceedings of the Association for Computational Linguistic, 8, 2008, pp. 308–316.

[21] K.H. Yoo, Y. Lee, U. Gretzel, D.R. Fesenmaier, Trust in travel-related consumer generated media, in: Information and Communication Technologies in Tourism, 2009, pp. 49–59.

[22] H.Y. Zhang, P. Ji, J.Q. Wang, X.H. Chen, A novel decision support model for satisfactory restaurants utilizing social information: a case study of tripadvisor.com, Tour. Manag. 59 (2017) 281–297.

[23] B. Pang, L. Lillian, V. Shivakumar, Thumbs up?: Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10, Association for Computational Linguistics, 2002.

[24] L. Ku, Y. Liang, H. Chen, Opinion extraction, summarization and tracking in news and blog corpora, in: Proceedings of the American Association on Artificial Intelligence, 2006, pp. 100–107.

[25] E. Sulis, D.I.H. Farias, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in twitter: differences between #irony, #sarcasm and #not, Knowl. Based Syst. 108 (2016) 132–143.

[26] M. Kågebäck, H. Salomonsson, Word sense disambiguation using a bidirectional lstm, 2016. arXiv preprint arXiv:1606.03568.

[27] Y. Ren, D. Ji, Neural networks for deceptive opinion spam detection: an empirical study, Inf. Sci. 385 (2017) 213–224.

[28] F.N. Ribeiro, M. Arajo, P. Gonalves, M.A. Gonalves, F. Benevenuto, Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods, EPJ Data Sci. 5 (1) (2016) 1–29.

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

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

[31] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Assoc. Inf. Sci. Technol. 61 (12) (2010) 2544–2558.

[32] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Assoc. Inf. Sci. Technol. 63 (1) (2012) 163–173.

[33] F. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, 2011. arXiv:1103.2903.

[34] S. Poria, E. Cambria, A. Gelbukh, F. Bisio, A. Hussain, Sentiment data flow analysis by means of dynamic linguistic patterns, IEEE Comput. Intell. Mag. 10 (4) (2015) 26–36.

[35] I. Chaturvedi, Y.S. Ong, I.W. Tsang, R.E. Welsch, E. Cambria, Learning word dependencies in text by means of a deep recurrent belief network, Knowl. Based Syst. 108 (2016) 144–154.

[36] A. Valdivia, M.V. Luzón, F. Herrera, Neutrality in the sentiment analysis problem based on fuzzy majority, in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ), IEEE, 2017, pp. 1–6.

[37] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowl. Based Syst. 108 (2016) 42–49.

[38] Meaningcloud opinion mining API, 2017, https://www.meaningcloud.com/products/sentiment-analysis. Online; Accessed Jan 2017.

[39] A. Valdivia, M.V. Luzón, F. Herrera, Sentiment analysis in TripAdvisor, IEEE Intell. Syst. 32 (4) (2017) 72–77.

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

[41] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining Text Data, Springer US, 2012.

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

[43] P. OConnor, User-generated content and travel: a case study on TripAdvisor.com, Inf. Commun. Technol. Tour. 2008 (2008) 47–58.

[44] B. Lu, B.K. Tsou, Combining a large sentiment lexicon and machine learning for subjectivity classification, in: Proceedings of the 2010 International Conference on Machine Learning and Cybernetics (ICMLC), 6, 2010, pp. 3311–3316.

[45] K. Schouten, F. Frasincar, Survey on aspect-level sentiment analysis, IEEE Trans. Knowl. Data Eng. 28 (3) (2016) 813–830.

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

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

[48] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.

Ana Valdivia obtained her B.Sc. in Mathematics from Polytechnic University of Catalonia (UPC) in 2014. She then worked as a research assistant at IESE Business School of Barcelona in 2015. She received her M.Sc. in Data Science and Computer Engineering from University of Granada (UGR) in 2016. Now, she is currently pursuing the Ph.D. degree in Computer Science and Artificial Intelligence in UGR. Her research is focused on Natural Language Processing, Sentiment Analysis and Deep Learning. She is also a strong advocate for Data Science for Social Good.

Emiliya Hrabova received her both Bachelor and Master degrees in Computer Engineering from University of Sannio (Italy) in 2016 and 2017, respectively. She studied at University of Anakara in 2015 and University of Granada in 2016 with Erasmus+ Programme as well. In 2017 she attended The Cornell, Maryland, Max Planck Pre-Doctoral Research School. Her areas of interest include Data Mining, Machine Learning and Natural Language Processing.

Dr Iti Chaturvedi obtained her Bachelors in Computer Engineering from National University of Singapore on the Singapore Airlines Undergraduate Scholarship. She received her Ph.D. in Computer Engineering from Nanyang Technological University, Singapore on the SCE Ph.D. Scholarship. She also worked there as Research Fellow for three years. Her Ph.D. was on dynamic Bayesian networks with application to time series data. Her Post-doctoral work focused on Deep Learning of large datasets such as classification of documents. Currently she is working at NTU Temasek Labs on sentiment detection from microblogs such as Twitter. She is also applying deep learning to YouTube Product reviews in new languages such as Spanish. She has co-authored over 35 journal and conference publications.

M. Victoria Luzn is an associate professor in the Software Engineering Department at University of Granada. Her research interests include sentiment analysis, artificial intelligence, and computer graphics. Luzn has a PhD in industrial engineering from the University of Vigo.

Luigi Troiano (M.Eng 2000, Ph.D. 2004) is assistant professor of Artificial Intelligence, Data Science and Machine Learning at University of Sannio, Department of Engineering, Italy. His research is devoted to mathematical modelling and algorithm development with applications, Media and Finance among the others. He investigated different fields concerning artificial neural networks, fuzzy set theory, aggregation functions, evolutionary algorithms, probabilistic reasoning, data mining. His expertise is designing, experimenting and validating algorithms, along their implementation in software systems for industrial environments, including some large international companies. He is coordinator of Computational and Intelligent Systems Engineering Laboratory (CISELab) at University of Sannio.

Erik Cambria received his Ph.D. in Computing Science and Mathematics in 2012 following the completion of an EPSRC project in collaboration with MIT Media Lab, which was selected as impact case study by the University of Stirling for the UK Research Excellence Framework (REF2014). After working at HP Labs India, Microsoft Research Asia, and NUS Temasek Labs, in 2014 he joined NTU SCSE as an assistant professor. Dr Cambria is associate editor of several journals edited by Elsevier, e.g., INFFUS and KBS, Springer, e.g., AIRE and Cognitive Computation, and IEEE, e.g., CIM and Intelligent Systems. He is recipient of many awards, e.g., AI's 10 to Watch and Emerald Citations of Excellence, and is involved in several international conferences as PC member, e.g., AAAI, UAI, and ACL, workshop organizer, e.g., ICDM SENTIRE, program chair, e.g., ELM, and invited speaker, e.g., IEEE SSCI.

Francisco Herrera (SM'15) received his M.Sc. in Mathematics in 1988 and Ph.D. in Mathematics in 1991, both from the University of Granada, Spain. He is currently a Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He has been the supervisor of 41 Ph.D. students. He has published more than 360 journal papers, receiving more than 57000 citations (Scholar Google, H-index 118). He is coauthor of the books "Genetic Fuzzy Systems" (World Scientific, 2001) and "Data Preprocessing in Data Mining" (Springer, 2015), "The 2-tuple Linguistic Model. Computing with Words in Decision Making" (Springer, 2015), "Multilabel Classification. Problem analysis, metrics and techniques" (Springer, 2016), among others. He currently acts as Editor in Chief of the international journals "Information Fusion" (Elsevier) and Progress in Artificial Intelligence (Springer). He acts as editorial member of a dozen of journals. He received the following honors and awards: ECCAI Fellow 2009, IFSA Fellow 2013, 2010 Spanish National Award on Computer Science ARITMEL to the "Spanish Engineer on Computer Science", International Cajastur "Mamdani" Prize for Soft Computing (Fourth Edition, 2010), IEEE Transactions on Fuzzy System Outstanding 2008 and 2012 Paper Award (bestowed in 2011 and 2015 respectively), 2011 Lotfi A. Zadeh Prize Best paper Award (IFSA Association), 2013 AEPIA Award to a scientific career in Artificial Intelligence, 2014 XV Andaluca Research Prize Maimnides, 2017 Security Forum I+D+I Prize, and 2017 Andaluca Medal (by the regional government of Andaluca). He has been selected as a Highly Cited Researcher http://highlycited.com/ (in the fields of Computer Science and Engineering, respectively, 2014 to present, Clarivate Analytics). His current research interests include among others, soft computing (including fuzzy modeling, evolutionary algorithms and deep learning), computing with words, information fusion and decision making, and data science (including data preprocessing, prediction and big data).