ORIGINAL RESEARCH

CrossMark

What do people think about this monument? Understanding negative reviews via deep learning, clustering and descriptive rules

Ana Valdivia[1] · Eugenio Martínez-Cámara[1] · Iti Chaturvedi[2] · M. Victoria Luzón[1] · Erik Cambria[2] · Yew-Soon Ong[2] · Francisco Herrera[1]

Abstract

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

Keywords Sentiment analysis · Deep learning · Aspect clustering · Subgroup discovery

✉ Eugenio Martínez-Cámara
emcamara@decsai.ugr.es

Ana Valdivia
avaldivia@ugr.es

Iti Chaturvedi
iti@ntu.edu.sg

M. Victoria Luzón
luzon@ugr.es

Erik Cambria
cambria@ntu.edu.sg

Yew-Soon Ong
asysong@ntu.edu.sg

Francisco Herrera
herrera@decsai.ugr.es

[1] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain

[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

1 Introduction

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

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

However, we observe that ABSA methods work at entity level which implies that the information obtained is very specific for each opinion. These approaches permit

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

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

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

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

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

2 Background

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

2.1 Sentiment analysis

SA is an area which aims at identifying sentiments in opinions towards a specific entities. More formally, an opinion can be defined as a quintuple (e, a, s, h, t), in which e refers to the $entity$ of the opinion, a to the $aspects$ and components of this entity, s is the $sentiment$ of the opinion, h the author or *opinion holder* of the review, and t the date when it was expressed (Liu 2015). Hence, the main target of sentiment analysis is to discover the underlying polarity s. For example, in restaurant reviews the entity is the restaurant and the aspects are typical characteristics of the restoration field as: the service, the food, the price, etc.

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

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

2.2 Aspect-based sentiment analysis

ABSA focuses on extracting aspects and entities that have been evaluated in the reviews and gives a more detailed

information about the purpose of the opinion (Schouten and Frasincar 2016). People tend to review different aspects of the same entity rather than give an opinion of the whole object. For example, if we analyze the following statement about the Alhambra monument:

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

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

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

2.3 Deep learning for ABSA

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

In recent years, deep learning has revolutionized a large part of the computer science field. They provide the versatility of supervised learning and do not need to design and previously select a set of features. Furthermore, deep learning models are non-linear supervised classifiers which can fit the data in a more accurately way. Collobert et al. (2011) was the

first to introduce the use of Convolutional Neural Networks (CNN) in Natural Language Processing (NLP) tasks. Poria et al. (2016) presented a deep learning-based approach to ABSA, which is built upon two CNN layers combined with a set of linguistic patterns.

2.4 Descriptive rules and subgroup discovery

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

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

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

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

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

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

R :{SubgroupConditions} $\longrightarrow$ {Class},

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

R_1 : {Staff = 1} $\longrightarrow$ {Sentiment = negative},

R_2 : {WheelChairAccessible = 1, Staff = 0} $\longrightarrow$ {Sentiment = positive},

R_3 : {Ticket = 1, Staff = 1} $\longrightarrow$ {Sentiment = negative}.

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

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

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

where Covered Instances is the total number of instances that satisfies the subgroup conditions, and N is the total number of instances in the dataset.

– **Support:** Number of instances in the dataset that satisfies the Subgroup Conditions and the value of the Class. This can be computed as:

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

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

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

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

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

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

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

2.5 Related work

This work is presented as an approach for improving cultural experiences. We aim at describing patterns in cultural monument reviews through a methodology that combines SD and ABSA techniques, descriptive and deep learning models. In the literature, we find studies that also combine both areas of knowledge. Li et al. (2015) presented a system for identifying hotel features of interest and understanding customers behavior. They developed an approach based on EPR to detect important changes or trends in travelers' concerns. Hai et al. (2011) proposed an association rule mining approach for identifying implicit features in products on-line reviews based on co-occurrences. They build a set of rules with opinion words and explicit features and then given an opinion with an implicit feature, they assign it the feature of the rule that best fits. Li et al. (2010) developed a rule association method on tourist data of Hong Kong, which gave useful insights about tourist patterns in that city. Poria et al. (2014) proposed a rule-based model for extracting explicit and implicit aspects for product review. The rules were based on parsing dependences like: sentences having subject, or having auxiliary verbs, etc. Their model was fully unsupervised and it outperformed the state of the art. As we can observe, rules models have been always used in sentiment analysis for analyzing features or aspects relations. As far as we know, this work is the first that presents a methodology that combines aspects and sentiments for describing patterns on reviews.
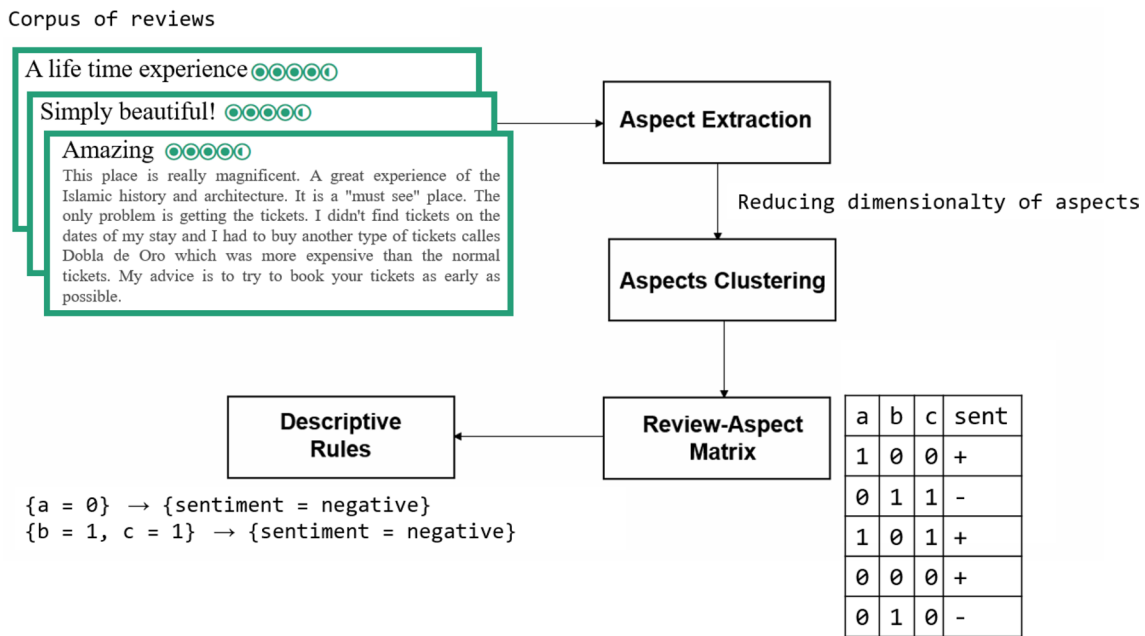
Fig. 1 Work-flow of the proposed methodology

3 Methodology for describing negative reviews based on deep learning, lustering and subgroup discovery

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

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

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

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

3.1 Deep learning: CNN for extracting aspects

Deep learning models are non-linear classifiers that are the state of the art in most of NLP tasks. Thus, we use the deep learning method presented in Poria et al. (2016) for aspect extraction, which is grounded in the use of convolutional layers. The features of an aspect term depend on its surrounding words. Thus, we used a window of five words around each targeted word in a sentence, i.e., two words. We formed the local features of that window and considered them to be features of the middle word. Then, the feature vector was fed to a CNN. The network contained one input layer, two convolution layers, two max-pool layers, and a fully connected layer with softmax output. The first convolution layer consisted of 100 feature maps with filter size 2. The second convolution layer had 50 feature maps with filter size 3. The stride in each convolution layer is 1 as we wanted to tag each word. A max-pooling layer followed each convolution layer. The pool size we use in the max-pool layers was 2. We used regularization with dropout on the penultimate layer with a constraint on L2-norms of the weight vectors, with 30 epochs. The output of each convolution layer was computed using a non-linear function; in our case we used *tanh*. The network architecture is formed by one input layer, two convolution layers, two max-pool

layers, and a fully connected layer with softmax output. The output of each convolution layer was computed using the *tanh* function. A set of heuristic linguistic patterns, which leverage on SenticNet (Cambria et al. 2018) and its extensions (Poria et al. 2012a, b), are run on the output of the deep learning model, which enhances the performance of the aspect extraction method.

Figure 2 depicts the process of extracting aspects from each sentence. Here we consider three types of labels for each word B-A (Begin Aspect), I-A (Internal Aspect) and O(Non-aspect). CNN use a sliding window of 3 or more words to look for features in the training data. For example here a convolutional kernel of 3 words is shown in blue. Each word is converted to feature representation using pre-trained word vectors. In this diagram we consider *n* features. Next, to encode the position information of the *aspect* word we include a position feature. For example, when training CNN for the aspect word *Gardens*, we set its position to 0, the word in front is set to 1 and the word before it is set to -1 and so on. The output layer of CNN predicts the aspect

category that is B-A, I-A or O for each word. To predict the final aspect category for a word we make use of the predicted labels of all the words in the sentence. This can be done by using a Conditional Random Field (CRF) where the label at each position is predicted using the previous two or three position words. The traditional CRF is unable to model long-range dependencies between words far apart in a sentence. However, CNN is able to capture such dependencies. Hence, the combined model is ideal for aspect term extraction.

3.2 Clustering: `K-means` **for clustering aspects**

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

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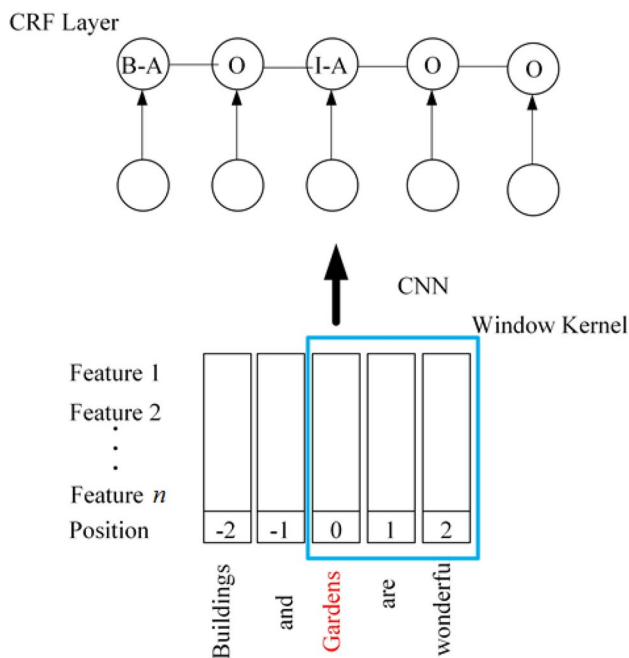

Fig. 2 Aspect Extraction using Convolutional Neural Networks

The great diversity of language implies: (1) we have to face the high dimensionality of aspects, because there are many words that express the same idea; (2) aspects with similar meanings have different representations.

To address these problems, we propose to cluster those aspects into the same group to decrease the dimensionality of features and produce a more descriptive summary by using a distributional representation of the aspects.

3.2.1 From words to vectors

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

Fig. 3 Pseudocode for the Apriori-SD algorithm (Kavšek and Lavrač 2006)

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

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

3.2.2 From vectors to clusters

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

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

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

where μ_i is the mean of points in C_i.

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

3.3 Subgroup discovery: Apriori-SD for descriptive rules

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

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

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

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

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

4 Experiments

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

4.1 Datasets

TripAdvisor is a travel website company which provides reviews from traveler experiences about hotels, restaurants,

---

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

Table 1 Summary of reviews, reviews with aspects and total number of unique aspects per monument

Monument	Reviews	Reviews with aspects	Aspects
Alhambra	7217	6186	9284
Mezquita	3526	2802	3688
Sagrada Familia	34,558	26,386	18,553

Table 2 Distributions of positive and negative polarities per monument

Monument	Positive	Negative
Alhambra	6781 (93.96 %)	436 (6.04 %)
Mezquita	3454 (97.96 %)	72 (2.04 %)
Sagrada Familia	32,664 (94.52 %)	1894 (5.48 %)

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

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

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

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

4.2 When thousand words represents a common idea

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

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

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

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

4.3 Depicting negative reviews of cultural monuments

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

We also analyze the most frequent clustered aspects in the three datasets. As we observe in Table 5, over the three monuments the most popular words are related with architectural topics. Therefore, we conclude that users tend to describe the monument while they are reviewing their visit in TripAdvisor.
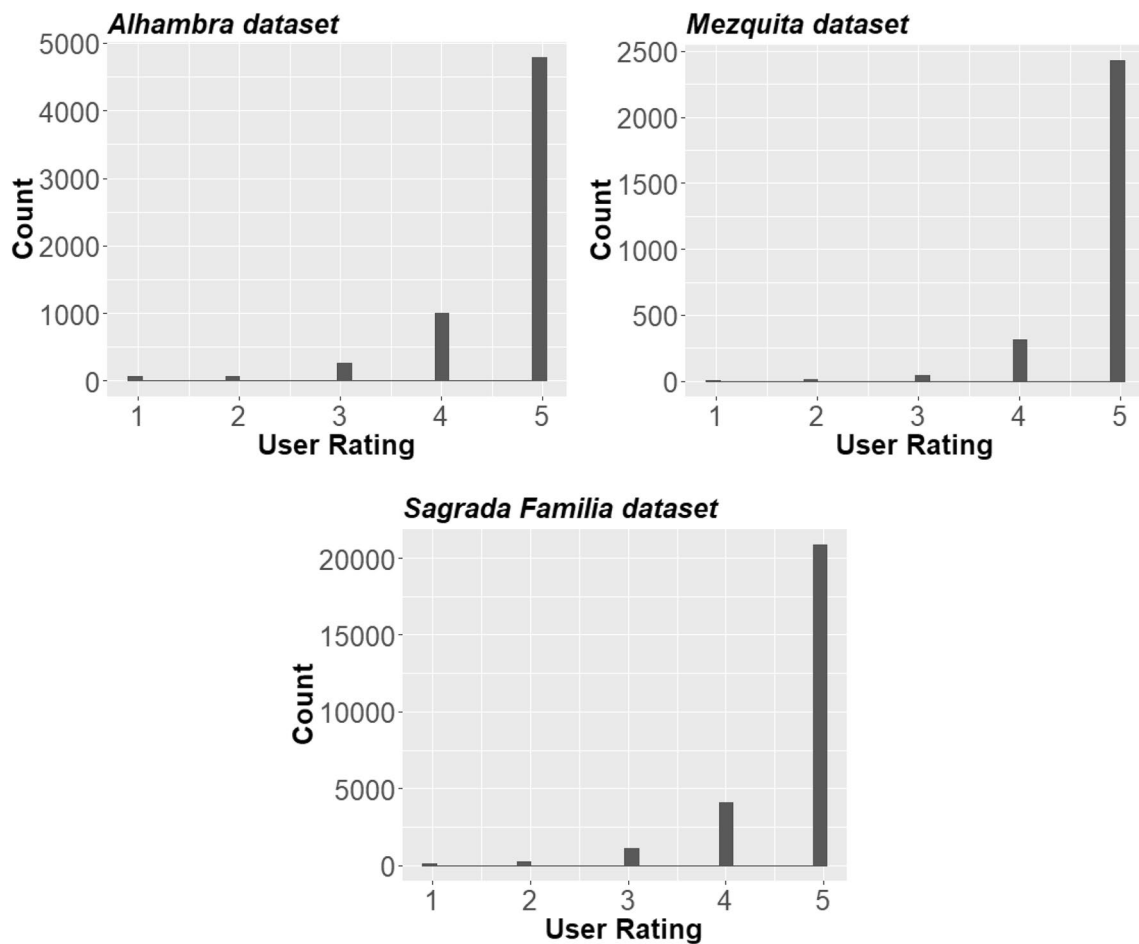
Fig. 4 Distribution of TripAdvisor rates

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

Monument	Total aspects	Aspects with embeddings	Aspects out-of-vocabulary
Alhambra	9284	5430	3854
Mezquita	3688	2291	1397
Sagrada Familia	18,553	10,247	8306

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

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

Analyzing the content of the rules, we detect some interesting patterns in the data. In the Alhambra dataset, the clustered aspects related with *staff, guard* and *cashier* are the
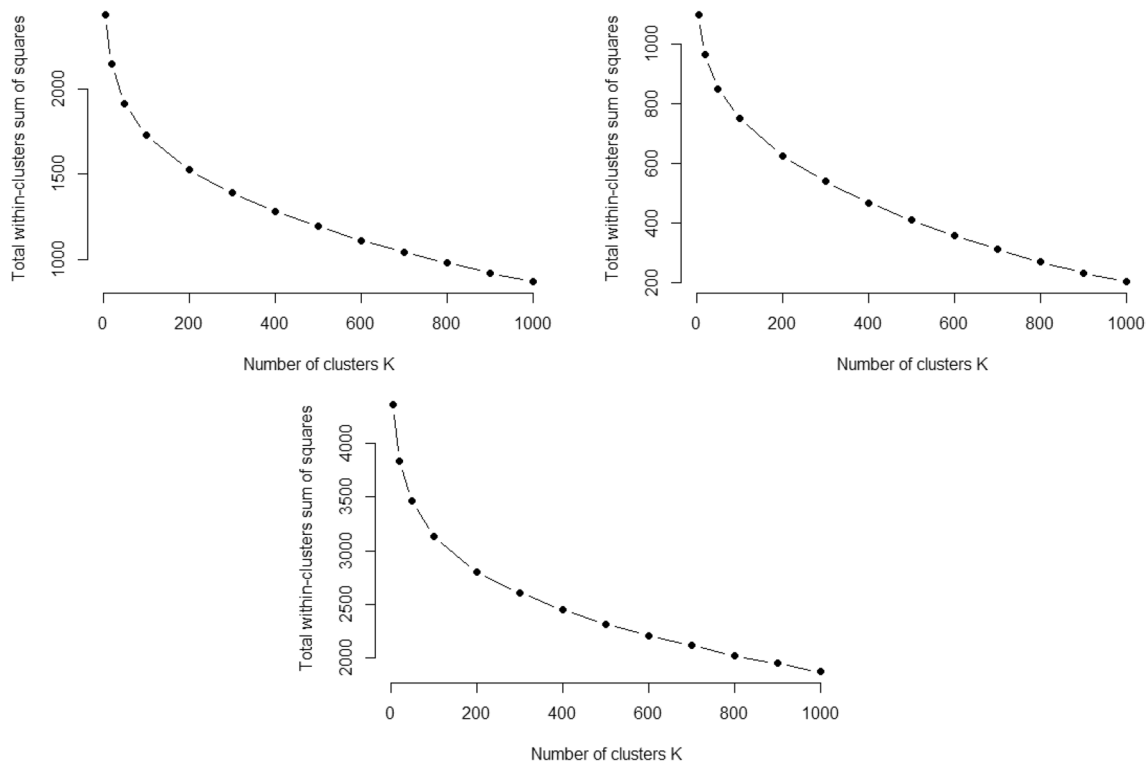
Fig. 5 Elbow plots for different *k* clusters

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

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

most significant rules with length equals to 2. The rule that is highly distinctive of the negative polarity, i.e., obtains the higher confidence is the one with length 3. In this rule, the clusters contains the words *staff* and *price*. That means that TripAdvisor users tend to complain about these two aspects together in negative reviews.

The metrics of the Mezquita and Sagrada Familia datasets are lower than the Alhambra. In these two datasets, all relevant rules obtain a length of two, which means that they have only one element in the antecedent. We also observe that in these datasets there exist other type of DRs more related with the type and characteristics of monument itself {*garden, architecture, ceiling, arches, ...*}. Consequently, we conclude that TripAdvisor users usually give an objective description of the building which they have visited.
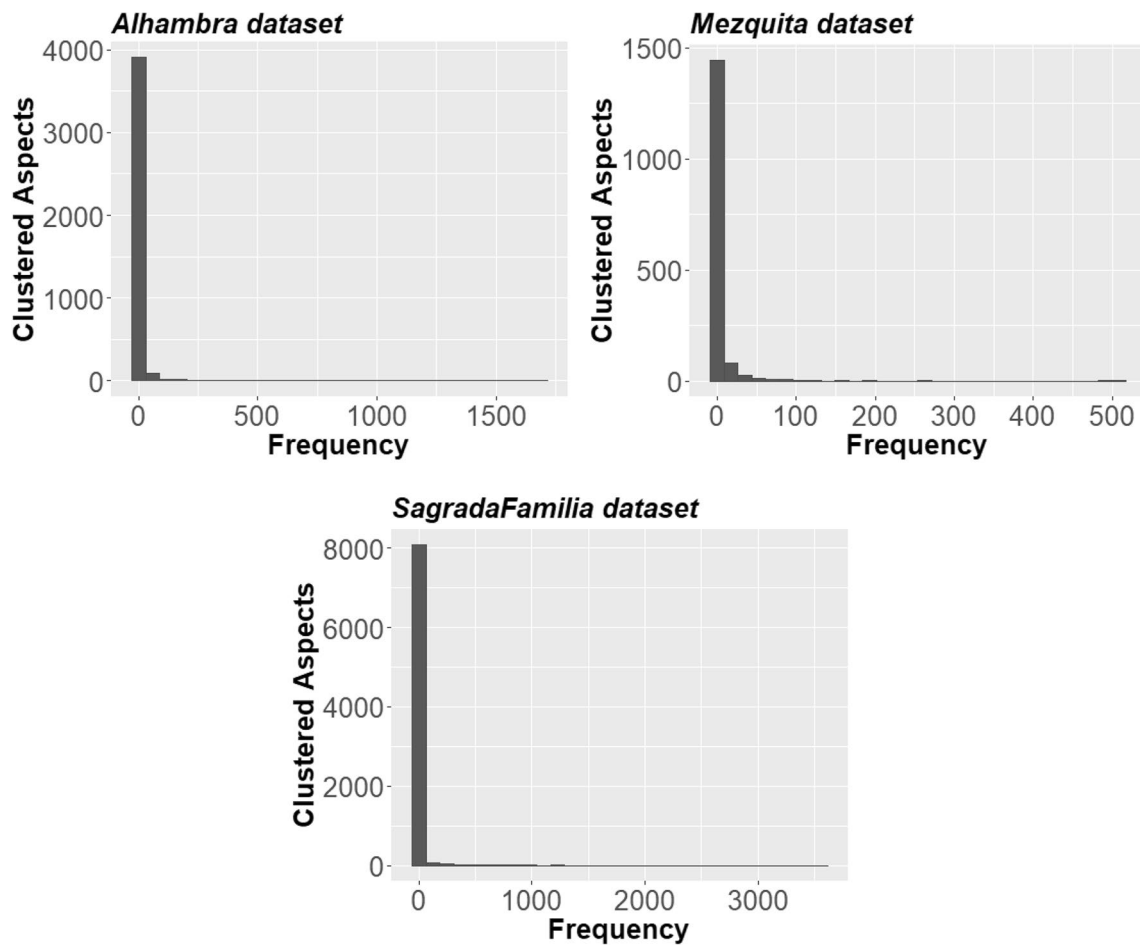
Fig. 6 Histograms of clustered aspects of the three monuments

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

Monument	Cluster label	Total	Cluster content (list of aspects)
Alhambra	BGB	1687	Alhambra, alhambra monument, alhambra ticketing, alhambra opens, alhambra museum, alhambra bookshop, alhambra tours, alhambra walking back
Alhambra	BFD	1162	Gardens, garden, generalife gardens, generalife garden, rose gardens, garden benches, garden landscaping, walk-in garden
Alhambra	BIA	735	Building, palaces, monuments, palaces fortresses, arab palaces, castles
Mezquita	BHF	509	Mosque, mosque complex, double-arched mosque, mezquita mosque, cordoban mosque
Mezquita	BEJ	489	Architecure, architecture design, architecture buff, cathedrals architecture, cultures architecture
Mezquita	FI	256	Building, views, view, perspectives, viewpoints
Sagrada Familia	ID	3555	Architecture, skeletal architecture, colors architecture, architecture engineering, designed architecture, catalan architecture
Sagrada Familia	HE	2798	Church, basilica, cathedral, church building, church glass, working church, gaudis unfinished church
Sagrada Familia	BIA	2128	Construction, continuous construction, construction noise, construction cranes, construction plaster, construction stones

Table 7 Most relevant rules of the Mezquita monument

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

Table 6 Most relevant rules of the Alhambra monument

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

Table 8 Most relevant rules of the Sagrada Familia monument

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

5 Discussion

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

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

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

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

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

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

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

This review was scored with a 4, so we set it as positive. However, we can find some negativity in the last two sentences. The user is complaining about the lack of refreshments and the adaptability of the monument to people with

disabilities. Therefore, these aspects are labelled with the overall sentiment of the review (positive) while they must be labelled as negative. This fact results in low confidence scores.

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

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

6 Conclusion

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

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

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

References

Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. Adv Knowl Discov Data Min 12(1):307–328

Atzmueller M (2015) Subgroup discovery. Wiley Interdiscip Rev Data Min Knowl Discov 5(1):35–49. https://doi.org/10.1002/widm.1144

Bay SD, Pazzani MJ (2001) Detecting group differences: mining contrast sets. Data Min Knowl Discov 5(3):213–246

Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107

Cambria E, Poria S, Hazarika D, Kwok K (2018) SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In: AAAI, pp 1795–1802

Carmona C, del Jesus M, Herrera F (2018) A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy. Knowl Based Syst 139:89–100

Carmona CJ, González P, Del Jesus M, Navío-Acosta M, Jiménez-Trevino L (2011) Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. Soft Comput 15(12):2435–2448

Carmona CJ, González P, del Jesus MJ, Herrera F (2014) Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms. Wiley Int Rev Data Min Knowl Discov 4(2):87–103

Chaturvedi I, Ragusa E, Gastaldo P, Zunino R, Cambria E (2018) Bayesian network based extreme learning machine for subjectivity detection. J Frankl Inst 355(4):1780–1797

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537

De Boom C, Van Canneyt S, Demeester T, Dhoedt B (2016) Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognit Lett 80:150–156

Dong G, Li J (1999) Efficient mining of emerging patterns: discovering trends and differences. In: KDD, ACM, pp 43–52

Fan H, Ramamohanarao K (2003) A bayesian approach to use emerging patterns for classification. In: ADC, Australian Computer Society, Inc., pp 39–48

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

Hai Z, Chang K, Kim Jj (2011) Implicit feature identification via co-occurrence association rule mining. In: CICLING, Springer, pp 393–404

Herrera F, Carmona CJ, González P, Del Jesus MJ (2011) An overview on subgroup discovery: foundations and applications. Knowl Inform Syst 29(3):495–525

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: KDD, ACM, pp 168–177

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

Jovanoski V, Lavrač N (2001) Classification rule learning with aprioric. In: EPIA, Springer, pp 44–51

Kasper W, Vela M (2011) Sentiment analysis for hotel reviews. Comput Linguist Appl Conf 231527:45–52

Kavšek B, Lavrač N (2006) Apriori-sd: adapting association rule learning to subgroup discovery. Appl Artif Intell 20(7):543–583

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

Lavrač N, Flach P, Zupan B (1999) Rule evaluation measures: a unifying view. In: ILP, Springer, pp 174–185

Lavrač N, Kavšek B, Flach P, Todorovski L (2004) Subgroup discovery with cn2-sd. J Mach Learn Res 5:153–188

Levy O, Goldberg Y (2014) Dependency-based word embeddings. ACL 2:302–308

Li G, Law R, Rong J, Vu HQ (2010) Incorporating both positive and negative association rules into the analysis of outbound tourism in Hong Kong. J Travel Tour Market 27(8):812–828

Li G, Law R, Vu HQ, Rong J, Zhao XR (2015) Identifying emerging hotel preferences using emerging pattern mining technique. Tour Manage 46:311–321

Liu B (2015) Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press, Cambridge

Lu B, Ott M, Cardie C, Tsou BK (2011) Multi-aspect sentiment analysis with topic models. In: 2011 IEEE 11th international conference on data mining workshops (ICDMW), IEEE, pp 81–88

Marrese-Taylor E, Velásquez JD, Bravo-Marquez F, Matsuo Y (2013) Identifying customer preferences about tourism products using an aspect-based opinion mining approach. Proc Comput Sci 22:182–191

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

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013 781

Nguyen HL, Jung JE (2017) Statistical approach for figurative sentiment analysis on social networking services: a case study on twitter. Multimed Tools Appl 76(6):8901–8914

Novak PK, Lavrač N, Webb GI (2009) Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J Mach Learn Res 10:377–403

Plutchik R (1984) Emotions: a general psychoevolutionary theory. Approaches Emot 1984:197–219

Poria S, Gelbukh A, Cambria E, Das D, Bandyopadhyay S (2012a) Enriching senticnet polarity scores through semi-supervised fuzzy clustering. In: ICDMW, pp 709–716

Poria S, Gelbukh A, Cambria E, Yang P, Hussain A, Durrani T (2012b) Merging senticnet and wordnet-affect emotion lists for sentiment analysis. ICSP 2:1251–1255

Poria S, Cambria E, Ku LW, Gui C, Gelbukh A (2014) A rule-based approach to aspect extraction from product reviews. SocialNLP 2014:28

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

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

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

Thorndike RL (1953) Who belongs in the family? Psychometrika 18(4):267–276

Toh Z, Wang W (2014) Dlirec: aspect term extraction and term polarity classification system. In: SemEval, pp 235–240

Valdivia A, Luzón MV, Herrera F (2017) Sentiment analysis in tripadvisor. IEEE Intell Syst 32(4):72–77

Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: European symposium on principles of data mining and knowledge discovery, Springer, pp 78–87

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