Mining on Line General Opinions About Sustainability of Hotels: A Systematic Literature Mapping

Thiago de Oliveira Lima(✉), Methanias Colaco Junior(✉), and Maria Augusta S. N. Nunes(✉)

Federal University of Sergipe (UFS), São Cristóvão, Sergipe, Brazil
thiagodeolima@gmail.com, mjrse@hotmail.com, gutanunes@gmail.com

Abstract. Context: Nowadays, people do not only navigate, but also contribute content to the Internet. Thoughts and opinions are written on rating sites, forums, social networks, blogs and other media. Such opinions constitute a valuable source of information for companies, governs and consumers, but it would be humanly impossible to analyze and locate the opinions in those assessments, due to the large volume and different origins of the data. For this, approaches and techniques of opinion mining in texts are used. **Objective:** To identify and characterize the techniques used for mining data in public opinion repositories regarding hotels, since the opinion mining area has offered necessary subsidies for decision-making related to hotel management. Besides, to identify, specifically, studies that investigated the opinions about the sustainability of hotels. **Method:** A systematic mapping was performed to characterize the research area. **Results:** It was identified that, among the main approaches, 31% of the works found use only data mining, while 55% exclusively use machine learning techniques, and 14% both. **Conclusion:** The most relevant studies in such research lines adopt machine learning algorithms such as Naive Bayes, SVM, LDA, decision tree, besides aspect-based techniques and SentiWordNet lexicon dictionaries. There are still opportunities to explore opinion mining solutions in online hotel reviews, mainly by taking into consideration aspects related to sustainable practices and sustainability levels practiced by each hotel.

Keywords: Data mining · Big data · Machine learning · NLP
Sustainability

1 Introduction

The specialized websites and the social networks have become the most popular platforms for sharing traveling information, with varied commentaries of diverse virtual origins, published every day. The automatically generated hotel reviews could help travelers when selecting hotels [1]. A convenient approach to divulging

O. Gervasi et al. (Eds.): ICCSA 2018, LNCS 10960, pp. 558–574, 2018.
https://doi.org/10.1007/978-3-319-95162-1_38

and promoting the tourism industry is through website, fact that has been improving nowadays. The researchers noted that the reviews and commentaries gathered by websites are useful information for clients and hotel managers [2].

However, the analysis of the huge set of review texts is a hard task not only for clients, but also for the parties interested in hotel management. On the other hand, the analysis of information based on few items may generate biased situations [3]. In the last years, some researchers have proposed opinion extraction systems, domain-independent mainly, to automatically extract structured representations of opinion contained in such texts [1,4–7]. As an example, consumers have been verifying the opinion of other consumers before buying a product with the intention to make a good purchase [8].

This systematic mapping aims to identify the techniques used for data mining in public written reviews on the Web. The purpose is to evaluate the state of the art of data mining in textual repositories aiming to extract the opinion of the consumers on hotels, in the context of tourists and managers in the hospitality field. A secondary goal was to verify the existence of works that mined the opinion of guests on the practice of sustainable initiatives by the hotels by analyzing the popularity of the opinions and possibly gauge the level of sustainability practiced by each hotel. Sustainability is certainly an ample term, but the present research has as secondary goal to verify if the sustainable practices implemented by the hotels such as: selective garbage collection, rain water captation, use of organic foods, use of low-consumption light bulbs, motion sensors to automatically turn light bulbs on and off in the environments, regular treatment of swimming pool water, may or may not influence the opinion polarity of the guests reviews. With such purpose, articles from important databases for computer science were mapped.

In such context, it is intended to answer the following research questions: **Q1**: What are the most used text mining techniques to detect the opinion of consumers in online reviews? **Q2**: Are the online text mining techniques used for detecting the consumers' satisfaction regarding the sustainable practices by the hotels and gauging the sustainability levels practiced by each hotel? **Q3**: What are the countries that have more researchers publishing on the theme? **Q4**: What years had more publications in the area? **Q5**: What are the main periodicals and conferences on the theme?

It was identified, when answering these questions, that Data Mining, with 31%, and Machine Learning, with 55%, were the most used approaches. Among the main approaches, we found opinion analysis based on lexicon dictionaries [4,6,7,9–12], emoticon [4], interjection [4] and acronym dictionaries [4], besides statistical methods such as Cohen's Kappa [5], maximum entropy [4,5,9] and Chi-Square [2]. In other studies, we also identified aspect-based methods [8,13–16], grammatical classes of words, supposedly named POS-Tagging [9,17] and Machine Learning algorithms, such as k-medoids [1], J48 [6,18], SVM [4,6,12,18–22], logistic regression [23], Naive Bayes [3,4,6,11,12,18,20], LDA [7,20,24–27], AdaBoost [19,20], decision tree [3,6,12,20], LSA (Latent Semantic Analysis) [28], linear regression [28] and TSC (Topic Sentiment Criteria) [29]. In relation

to the aspects and sustainability levels practiced by hotels, we identified only one study which relates the influence of sustainable issues on the opinion of the clients [24].

On the technical resources for opinion mining, the Naive Bayes algorithms [3,4,6,11,12,18,20] and SVM [4,6,12,18–22] were the ones which highlighted more in the Machine Learning sphere, while the lexicon dictionary related to the identification of texts in English named SentiWordNet had the greatest highlight in the Data Mining area [4,6,7,9,12]. In relation to countries all over the world, the USA was the country with more publications in the area.

This article is organized in the following way: in Sect. 2, the method adopted in this mapping is presented; in Sect. 3, the analysis results are described; in Sect. 4, the threats to validity are presented; finally, in Sect. 5, the conclusion is presented.

2 Method

The Systematic Mapping consists of a systematic protocol for search and selection of relevant studies aiming to extract information and map the results for a specific research issue [30,31]. Such protocol was proposed by Kitchenham et al. for systematic review in 2004 and Petersen et al. for systematic mapping in 2008. The objective of this work is limited to developing a Systematic Mapping with the intention to identify, analyze and evaluate case studies, primary works or surveys, to characterize the use of algorithms, methods and techniques for opinion mining in public reviews of hotels that take or do not take sustainability aspects into consideration. Initially, for the definition of research questions, the approaches that used opinion mining in texts related to the public reviews of hotels were detailed.

Thus, aiming to follow the systematic mapping method, how the process of search and selection of primary studies was developed is described in the following section. To that end, it was necessary to define the research questions, search and selection strategy and selection criteria. Thus, the detailed description of the process of search and selection of primary studies follows below.

2.1 Research Questions

To achieve the objective proposed by this mapping, the following research questions were elaborated:

- **Q1:** What are the most used text mining techniques to detect the opinion of consumers in online reviews?
- **Q2:** Are the online text mining techniques used for detecting the consumers' satisfaction regarding the sustainable practices by the hotels and gauging the sustainability levels practiced by each hotel?
- **Q3:** What are the countries that have more researchers publishing on the theme?

- **Q4:** What years had more publications in the area?
- **Q5:** What are the main periodicals and conferences on the theme?

For the characterization of the used computing approaches, we initially believed that the studies would be contained in the three computing areas that deal with pattern recognition: Data Mining, Artificial Intelligence and Machine Learning. As Data Mining is the application of specific algorithms for the extraction of patterns from a database [32], such area commonly presents an overlapping with the areas of Artificial Intelligence, Statistics and Database. The studies were classified as Data Mining when they explicitly referred to the term. If they did not, even if the study had used Artificial Intelligence, we consider Data Mining when the use of algorithm was made from a Database Managing System (DMS) with data aspects.

The separate classification of the Machine Learning area in an area originated from Artificial Intelligence, actually happened due to the fact that such area has periodicals and exclusive conferences, such as: Machine Learning, ISSN: 0885-6125; Machine Learning (DORDRECHT. ONLINE), ISSN: 1573-0565; ICML - International Conference on Machine Learning and ICMLA - IEEE International Conference on Machine Learning and Applications. Besides, it is an area classified on the same level as Artificial Intelligence, by the ACM classification system and in conferences indexed by IEEE.

2.2 Research Scope

For the execution of the Systematic Mapping, the Scopus database was used. For unrestricted downloads in the database the access to the Capes periodical portal [33] was used.

The choice of the Scopus database happened due to the fact that its collection incorporates articles from many other databases, such as: IEEE, ACM, Springer and Elsevier. Those databases are responsible for publishing the main periodicals in the computer science area.

The sources were select according to the availability of consultation over the Internet, which were indexed in the databases cited above, being able to be found through keyword search. In relation to the language, only works in English and Portuguese were selected. In relation to the area, only works referring to Computer Science were selected. In relation to the type of publication, only articles published in conferences, periodicals or book chapters were selected.

In the Scopus database, after using the search function by title, summary or keyword, the advanced search option to select only results belonging to the Computer Science area only in the English or Portuguese languages was used, and results that referred to recapitulations of conferences and notes were also excluded.

2.3 Publication Search Method

The sources were accessed over the Internet. The Search String that was later used in the selected database generated through the combination of keywords is:

Search String 01. ((TITLE-ABS-KEY (("review" OR "evaluation" OR "ratingOR "appraisal" OR "valuation" OR "appreciation" OR "rate" OR "estimate" OR "reckoning" OR "appraisement" OR "account" OR "putting" OR "opinion" OR "avaliação" OR "opinião" OR "comentário") AND ("hotel" OR "hospitality" OR "hostel" OR "house for season" OR "hospitabilidade" OR "pousada" OR "casa para temporada") AND ("data mining" OR "data analytics" OR "big data" OR "business intelligence" OR "data science" OR "artificial intelligence" OR "NLP" OR "natural language" OR "opinion mining" OR "sentiment classification" OR "sentiment analysis" OR "mineração de dados" OR "análise de dados" OR "ciência de dados" OR "inteligência artificia" OR "PLN" OR "linguagem natural" OR "mineração de opinião" OR "machine learning" OR "aprendizado de máquina"))) AND (LIMIT-TO (SUBJAREA, "COMP")) AND (EXCLUDE (DOCTYPE, "cr") OR EXCLUDE (DOCTYPE, "no")).

Search String 02. ((TITLE-ABS-KEY (("review" OR "evaluation" OR "rating" OR "appraisal" OR "valuation" OR "appreciation" OR "rate" OR "estimate" OR "reckoning" OR "appraisement" OR "account" OR "putting" OR "opinion" OR "avaliação" OR "opinião" OR "comentário") AND ("hotel" OR "hospitality" OR "hostel" OR "house for season" OR "hospitabilidade" OR "pousada" OR "casa para temporada") AND ("data mining" OR "data analytics" OR "big data" OR "business intelligence" OR "data science" OR "artificial intelligence" OR "NLP" OR "natural language" OR "opinion mining" OR "sentiment classification" OR "sentiment analysis" OR "mineração de dados" OR "análise de dados" OR "ciência de dados" OR "inteligência artificial" OR "PLN" OR "linguagem natural" OR "mineração de opinião" OR "machine learning" OR "aprendizado de máquina"))) AND ("sustainability" OR "sustainable" OR "tenable" OR "sustentável" OR "sustentabilidade") AND (LIMIT-TO (SUBJAREA, "COMP")) AND (EXCLUDE (DOCTYPE, "cr") OR EXCLUDE (DOCTYPE, "no")).

The search was performed during October 2017, the search string returned 250 results. After the search stage, the article selection stage, which will be detailed below, was initiated.

2.4 Criteria Selection Procedures

The performed searches used the string search from Sect. 2.3 and the collected results were calculated taking into consideration only the selected studies for evaluation. The results that obeyed the inclusion criteria were selected for summary reading and titles. The selected articles were read, analyzed and sent to the result extraction stage.

The inclusion criteria were:

1. The result must contain the theme of this study in the title, summary or keywords;
2. The result must be available for online search;
3. The result must explore an algorithm, technique, mechanism or approach of opinion mining in online hotel reviews;
4. The articles must have indicatives of how the opinion of the consumers is influenced by sustainable practices of the hotels that used them (exclusive for the research question Q2).

The exclusion criteria were:

1. Articles that do not have relation with the computer science field;
2. Secondary studies, because they deal with approaches from third parties;
3. Unavailable articles;
4. Preliminary studies.

Among the 250 works found, after the inclusion and exclusion criteria, 29 were selected for complete reading and analysis. Figure 1 shows the amount of articles by scientific repository after the application of the selection criteria and primary studies analyses.

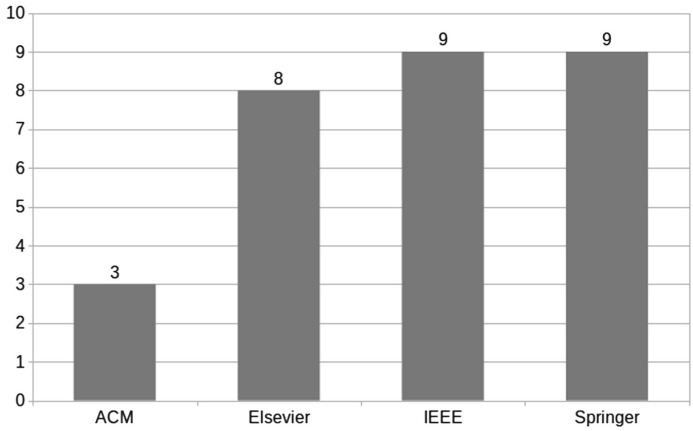

Fig. 1. Selection of articles by base

In Table 1, we can identify a general summary of the research result and application of the selection criteria, resulting in 29 studies.

Table 1. Results of the research in the databases and results of the application of the selection criteria.

Data repository	Research result	Application of selection criteria
ACM	24	4
Elsevier	19	9
IEEE	65	8
Springer	70	8
Others	72	0
Total	**250**	**29**

3 Discussion and Results

In this section, the analysis results of the primary study answering the previously defined research questions are presented. In Table 2, the 29 articles selected for this work, as well as the reference to each author are listed.

Table 2. Referred papers.

Paper title	Reference number
Sentiment analysis method for tracking touristics reviews in social media network	[4]
Word2Vec approach for sentiment classification relating to hotel reviews	[5]
Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling	[24]
Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews	[23]
Word of mouth quality classification based on contextual sentiment lexicons	[6]
The analysis and prediction of customer review rating using opinion mining	[3]
Opinion mining from online hotel reviews - A text summarization approach	[1]
Aspect identification and ratings inference for hotel reviews	[13]
Proposal of LDA-Based Sentiment Visualization of Hotel Reviews	[7]
Sentiment Polarity Classification Using Structural Features	[9]
An Investigation of Effectiveness Using Topic Information Order to Classify Tourists Reviews	[20]
Sentiment analysis of hotel reviews in greek: A comparison of unigram features	[21]
Learning sentiment based ranked-lexicons for opinion retrieval	[10]
A novel deterministic approach for aspect-based opinion mining in tourism products reviews	[14]
Are influential writers more objective? An analysis of emotionality in review comments	[11]
Opinion mining and summarization of hotel reviews	[12]
Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification	[25]
Hierarchical multi-label conditional random fields for aspect-oriented opinion mining	[15]
The ensemble of Naive Bayes classifiers for hotel searching	[2]
OpinionZoom, a modular tool to explore tourism opinions on the Web	[17]
A boosted SVM based sentiment analysis approach for online opinionated text	[19]
A distant supervision method for product aspect extraction from customer reviews	[26]
Long autonomy or long delay?' the importance of domain in opinion mining	[16]
Identifying customer preferences about tourism products using an aspect-based opinion mining approach	[8]
Analyzing user reviews in tourism with topic models	[29]
Machine learning approach to recognize subject based sentiment values of reviews	[18]
A Study on Text-Score Disagreement in Online Reviews	[22]
Aspect based Sentiment Oriented Summarization of Hotel Reviews	[27]
Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors	[28]

In the chart in Fig. 2 the characterization of the main approaches found for opinion mining is presented. As an answer to research question **Q1**, the most used approaches were the SVM machine learning techniques (27.5%) [4,6,12,18–22], *Naive Bayes* (24%) [3,4,6,11,12,20] and the LDA machine learning algorithm [7,20,24–26], followed by aspect-based data mining techniques [8,13–16,18] and the SentiWordNet lexicon dictionary [4,6,7,9,12], both with 17.24%. In 13.79% of the works analyzed, the machine learning algorithm based on decision tree was adopted, [3,6,12,20]. Machine learning statistical methods based on maximum entropy were used in 10.34% of the studies [4,5,9]. With 6.9% of the primary studies, the application of the AdaBoost machine learning algorithm

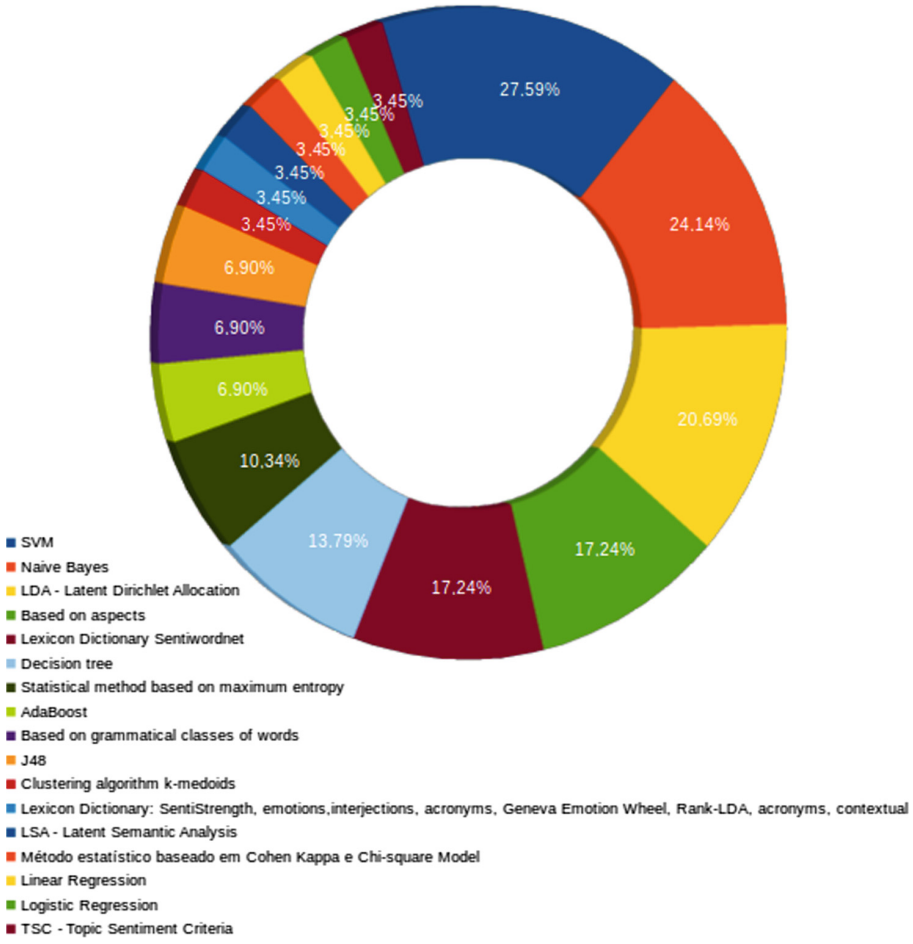

Fig. 2. Characterization of approaches, techniques and mechanisms

[19,20] and data mining based on word classes (Pos-Tagging) [9,17] and J48 [6] were identified. Finally, with 3.45%, we found data mining techniques based on the SentiStrength Lexicon dictionary [4,6,7,9,12], emoticon dictionary [4], interjection dictionary [4], acronym dictionary [4], Geneva Emotion Wheel [11], Rank-LDA [10], contextual dictionary [6] and machine learning algorithms such as logistic regression [23], k-medoids clustering [1] and Cohen's Kappa statistical methods [5], Chi-square Model [2], linear regression [28], LSA - Latent Semantic Analysis [27] and TSC - Topic Sentiment Criteria [29].

Table 3 brings a detailed mapping of the identified approaches, techniques and mechanisms. By performing a deeper analysis of the synthesized data, it was verified that the works [4,6,7,9–12] combine machine learning and data mining techniques.

Table 3. Results of the research in the databases and results of the application of the selection criteria.

Num.	Approach, technique and mechanism	Reference number
01	SentiWordnet Lexicon Dictionary	[4,6,7,9,12]
02	SentiStrength Lexicon Dictionary	[7]
03	Emoticons Dictionary	[4]
04	Interjections Dictionary	[4]
05	Acronyms Dictionary	[4]
06	GALC lexicon dictionary based on Geneva Emotion Wheel	[11]
07	Lexicon Dictionary based on Rank-LDA	[10]
08	Contextual Lexicon Dictionary	[6]
09	Based on aspects	[8,13–16]
10	Based on grammatical classes of words	[9,17]
11	Statistical method based on Cohen Kappa	[5]
12	Statistical method based on maximum entropy	[4,5,9]
13	Statistical method based on Chi-square Model	[2]
14	K-medoids clustering algorithm	[1]
15	J48	[6] [18]
16	SVM	[4,6,12,18–22]
17	Logistic regression	[23]
18	Naive Bayes	[3,4,6,11,12,20,27]
19	LDA - Latent Dirichlet Allocation	[7,20,24–26]
20	LSA - Latent Semantic Analysis	[28]
21	TSC - Topic Sentiment Criteria	[29]
22	AdaBoost	[19,20]
23	Decision tree	[3,6,12,20]

In Fig. 3, the uses of approaches, techniques and mechanisms in primary studies are presented. The Naive Bayes machine learning techniques [3,4,6,11, 12,18,20] and SVM [4,6,12,18–22] have the highest number of works (7), followed by aspect-based techniques [8,13–16], LDA [7,20,24–27] and the SentiWordNet lexicon dictionary (6) [4,6,7,9,12].

As the answer to research question **Q2**, only one work gauged questions related to sustainability [24]. According to Calheiros et al., from the 401 reviews used as corpus of the work, 13% were from the TripAdvisor website, 68% from the Suggestion book, 10% from follow-up emails, 1% from the website review, 6% from direct emails and 1% from other means. However, there were only 95 reviews indicating that the sustainable practices performed by hotels positively influenced the clients opinion.

To answer the research question **Q3**, according to Fig. 4, we verified that the USA had the highest number of publications (9), while Thailand, Japan, India, Portugal, England, Australia and Canada (2) and only with one publication Holland, Greece, Scotland, South Korea, Austria, Australia and Germany.

Research question **Q4** can be answered through the analysis of Fig. 5. It has been verified a considerable increase of interest in researches on the opinion mining area applied to opinion mining in the last two years. Five studies in 2013, six studies in 2014, only two studies in 2015, five studies in 2016 were

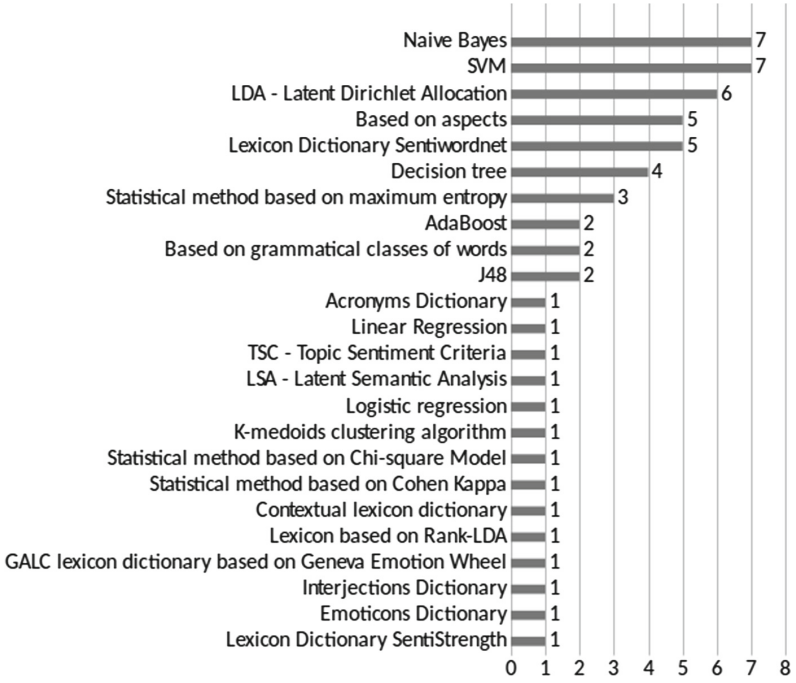

Fig. 3. Use of approaches, techniques and mechanisms

selected, while eight studies in 2017 that are adequate for our research line were considered.

Figure 6 presents a refinement of the graphic present in Fig. 5, in which we can observe the distribution of publications per year, divided by their respective means of publication. The conferences were responsible for the highest number of publications in all years.

Lastly, the answer to **Q5** is found in Table 4, where it can be observed that the primary studies were published in different periods or conferences, in which we can observe that the most popular means of publication of articles on the theme is the conferences. Such pattern does not surprise us, because the conferences are famously the most accessible means for scientific publications. The Scopus database, for example, indexes more than 100 thousand events of conference worldwide, while the number of indexed journals is almost 22 thousand [34]. What surprises us is the absolute low number of publications in journals, because such fact can denote that the works published in conferences did not stand out to deserve any extension in a journal, it means, they may not have been well-executed. Besides, they may not have gotten enough riveting results for their continuity and deepening. We can conclude that conferences The International Conference on Data Mining Workshop (ICDMW) and International Conference on Advances in Computing & Communications had, together, the largest number of publications (4).

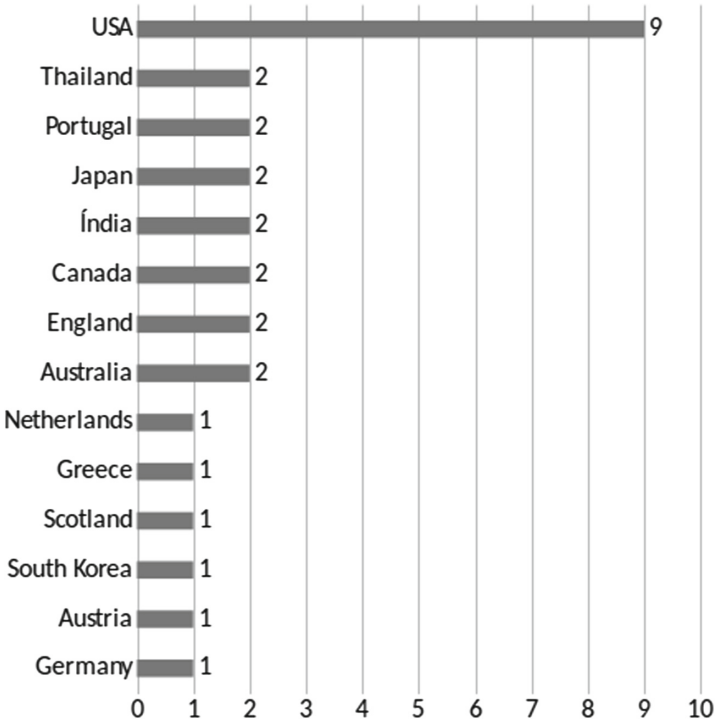

Fig. 4. Articles by country

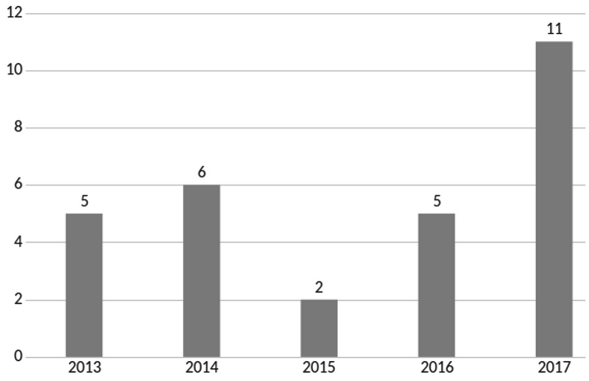

Fig. 5. Publications by year

We must highlight that the journals were published and conferences occurred in more than 12 distinct countries, characterizing homogeneity of publications and interests in different countries. In Fig. 7 we can observe the highest number of publications in conferences (17) and journals (12).

Table 4. Papers vs vehicles

Vehicles	Publications	Journal	Proceedings
Smart Innovation, Systems and Technologies	1		X
2014 International Computer Science and Engineering Conference, ICSEC 2014	2		X
Advances in Intelligent Systems and Computing	1		X
Expert Systems with Applications	2	X	
Information Processing and Management	2	X	
Journal of Hospitality Marketing and Management	1	X	
Knowledge-Based Systems	1	X	
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	1		X
Procedia Computer Science	1		X
Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015	2		X
Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013	1		X
Proceedings - 2013 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IATW 2013	1		X
Proceedings - 2014 6th International Conference on Computational Intelligence and Communication Networks, CICN 2014	1		X
Proceedings - 2015 International Conference on Computer Application Technologies, CCATS 2015	1		X
Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017	1		X
Proceedings of the 2013 Research in Adaptive and Convergent Systems, RACS 2013	1		X
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	1		X
Springer Proceedings in Mathematics	1		X
World Wide Web	1	X	
Information Technology and Tourism	1	X	
2nd International Moratuwa Engineering Research Conference, MERCon 2016	1	X	
Cognitive Computation	1	X	
7th International Conference on Advances in Computing & Communications	2		X
WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web	1		X

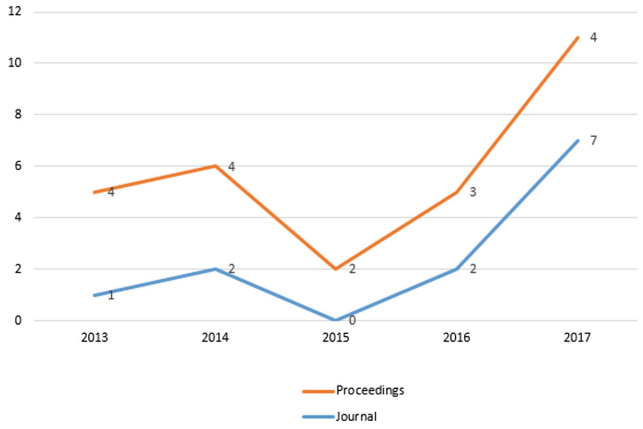

Fig. 6. Publications type vs year

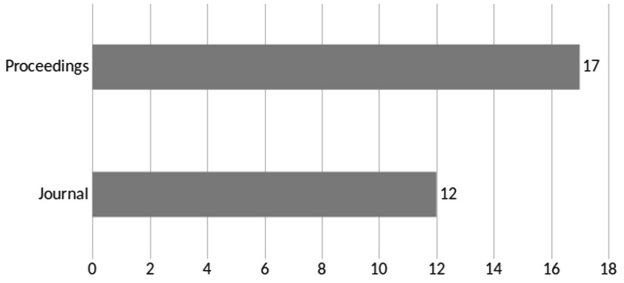

Fig. 7. Distribution of publications and journals

4 Threats to Validity

The threats to the validity of this study were:

- **Construction Validity:** The search String and the research questions used may not cover the opinion mining area in online reviews of hotels. To mitigate such threat, we tried to build a string that was as comprehensive as possible, regarding the terms that could be used in the area, using a control article and the opinion of 3 researchers.
- **Internal Validity:** (Data extraction): As three researchers were responsible for classifying and extracting the main algorithms from each publication, biases or problems in the data extraction may threaten the validity of data characterization; (Selection bias): In the beginning of the study, the articles were included or excluded from the mapping according to the judgment of the researchers themselves. It means that some studies may have been categorized incorrectly. To mitigate such threats, the selection and extraction reviews were performed by all the involved researchers, with a final voting due to disagreements.

– **External Validity:** Besides the fact that the Scopus database is the largest scientific literature database with more than 21,500 journals and 60 million registers [34], it is not possible to affirm that the results of this mapping comprehended the whole computer science area. However, this work presented evidence of the main techniques used, identifying gaps to be explored and serving as a guideline for future studies in this research line.

5 Conclusion

In this work, we performed a systematic mapping that aimed to identify and analyze approaches, techniques and mechanisms used in computer science for promoting data mining in online reviews of hotels.

The systematic mapping process was conducted by using a search protocol and study selection that specified the method used in this work. The data from 29 articles that meet the chosen research line were extracted and analyzed from the specified method.

Some strategies and approaches were used for data mining in online reviews of hotels. However, the most relevant ones were identified through Bayesian classifiers [3,4,6,11,12,20], SVM [4,6,12,19–21], LDA [7,20,24–26], aspect based [8,13–16], through the SentiWordNet lexicon dictionary [4,6,7,9,12] and based on maximum entropy [3,6,9,12,20] **(Q1)**.

Based on the results, we verified that there is a huge potential to be explored referring to the guests' opinions on the sustainable initiative practices by part of the hotels. As it was mentioned before, only the study [24] gauges the positive influences of sustainable practices **(Q2)**. In this regard, another potentiality is to identify, in the guests' commentaries, sustainability levels in their opinions, independent of the polarities of the reviews. It means, to elaborate data mining strategies to list the most sustainable hotels in relation to pre-established sustainability levels. We can also deepen in the aspect-based mining techniques to verify if it is viable to identify the sustainability levels of the hotels.

Among the primary studies selected for this mapping, we identify that Canada, India, Japan, Portugal, Thailand and the USA are responsible for about 60% of the countries with researches in our research line **(Q3)**.

Through the analysis, we can gauge that the last two years (2016 and 2017) were responsible for about 55% of the study databases that contribute towards the data collection proposed in this mapping **(Q4)**. Finally, we found 24 distinct conferences with publications in the research line, being 58% of the studies published in conferences and 42% in journals **(Q5)**.

Besides the apparent need to deepen the researches in the area under discussion, the results found in this work map the state of the art of opinion mining in online reviews of hotels, making it clear that it is an area of interest for researchers from various countries and it has great growth potential.

As a future work, we will investigate the gaps identified in this systematic mapping. We believe that this research presents relevant results for the academic field, providing support on how the mining opinion in hotels and sustainable

aspects were discussed in the Computer Science field, becoming a relevant consulting source for this research line.

This mapping can be extended by modifying the research search strings, the research questions or inclusion and exclusion criteria. A systematic review may be performed in this research line.

References

1. Hu, Y.H., Chen, Y.L., Chou, H.L.: Opinion mining from online hotel reviews - a text summarization approach. Inf. Process. Manag. **53**(2), 436–449 (2017). https://doi.org/10.1016/j.ipm.2016.12.002
2. Srisuan, J., Hanskunatai, A.: The ensemble of Naïve Bayes classifiers for hotel searching. In: 2014 International Computer Science and Engineering Conference, ICSEC 2014, no. 1, pp. 168–173 (2014)
3. Songpan, W.: The analysis and prediction of customer review rating using opinion mining. In: Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017, pp. 71–77 (2017)
4. Chaabani, Y., Toujani, R., Akaichi, J.: Sentiment analysis method for tracking touristics reviews in social media network. In: De Pietro, G., Gallo, L., Howlett, R.J., Jain, L.C. (eds.) KES-IIMSS 2017. SIST, vol. 76, pp. 299–310. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-59480-4_30
5. Polpinij, J., Srikanjanapert, N., Sopon, P.: Word2Vec approach for sentiment classification relating to hotel reviews. In: Meesad, P., Sodsee, S., Unger, H. (eds.) IC2IT 2017. AISC, vol. 566, pp. 308–316. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60663-7_29
6. Hung, C.: Word of mouth quality classification based on contextual sentiment lexicons. Inf. Process. Manag. **53**(4), 751–763 (2017)
7. Chen, Y.S., Chen, L.H., Takama, Y.: Proposal of LDA-based sentiment visualization of hotel reviews. In: Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, pp. 687–693 (2016)
8. Marrese-Taylor, E., Velásquez, J.D., Bravo-Marquez, F., Matsuo, Y.: Identifying customer preferences about tourism products using an aspect-based opinion mining approach. Procedia Comput. Sci. **22**, 182–191 (2013)
9. Ansari, D.: Sentiment polarity classification using structural features. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1270–1273 (2015). http://ieeexplore.ieee.org/document/7395814/
10. Peleja, F., Magalhães, J.: Learning sentiment based ranked-lexicons for opinion retrieval. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 435–440. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_47
11. Martin, L.: Are influential writers more objective? An analysis of emotionality in review comments categories and subject descriptors. In: WWW, pp. 799–804 (2014). https://doi.org/10.1145/2567948.2579242
12. Raut, V.B., Londhe, D.D.: Opinion mining and summarization of hotel reviews. In: Proceedings - 2014 6th International Conference on Computational Intelligence and Communication Networks, CICN 2014, pp. 556–559 (2014)
13. Xue, W., Li, T., Rishe, N.: Aspect identification and ratings inference for hotel reviews. World Wide Web **20**(1), 23–37 (2017). https://doi.org/10.1007/s11280-016-0398-9

14. Marrese-Taylor, E., Velásquez, J.D., Bravo-Marquez, F.: A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Syst. Appl. **41**(17), 7764–7775 (2014). https://doi.org/10.1016/j.eswa.2014.05.045
15. Marcheggiani, D., Täckström, O., Esuli, A., Sebastiani, F.: Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 273–285. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_23
16. Cruz, F.L., Troyano, J.A., Enríquez, F., Ortega, F.J., Vallejo, C.G.: 'Long autonomy or long delay?' The importance of domain in opinion mining. Expert Syst. Appl. **40**(8), 3174–3184 (2013)
17. Marrese-Taylor, E., Velasquez, J.D., Bravo-Marquez, F.: OpinionZoom, a modular tool to explore tourism opinions on the Web. In: Proceedings - 2013 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IATW 2013, vol. 3, pp. 261–264 (2013)
18. De Mel, N.M., Hettiarachchi, H.H., Madusanka, W.P., Malaka, G.L., Perera, A.S., Kohomban, U.: Machine learning approach to recognize subject based sentiment values of reviews. In: 2nd International Moratuwa Engineering Research Conference, MERCon 2016, pp. 6–11 (2016)
19. Sharma, A., Dey, S.: A boosted SVM based sentiment analysis approach for online opinionated text. In: Proceedings of the 2013 Research in Adaptive and Convergent Systems on - RACS 2013, pp. 28–34 (2013). http://dl.acm.org/citation.cfm?doid=2513228.2513311
20. Nakamura, S., Okada, M., Hashimoto, K.: An investigation of effectiveness using topic information order to classify tourists reviews. In: Proceedings - 2015 International Conference on Computer Application Technologies, CCATS 2015, pp. 94–97 (2016)
21. Markopoulos, G., Mikros, G., Iliadi, A., Liontos, M.: Sentiment analysis of hotel reviews in greek: a comparison of unigram features. In: Katsoni, V. (ed.) Cultural Tourism in a Digital Era. SPBE, pp. 373–383. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15859-4_31
22. Fazzolari, M., Cozza, V., Petrocchi, M., Spognardi, A.: A study on text-score disagreement in online reviews. Cogn. Comput. 1–13 (2017). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85026501769&doi=10.1007%2Fs12559-017-9496-y&partnerID=40&md5=e59e090854e197e0cf9d52d61659282a
23. Bauman, K., Liu, B., Tuzhilin, A.: Aspect based recommendations. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2017, pp. 717–725 (2017). http://dl.acm.org/citation.cfm?doid=3097983.3098170
24. Calheiros, A.C., Moro, S., Rita, P.: Sentiment classification of consumer-generated online reviews using topic modeling. J. Hosp. Mark. Manag. **8623**(March), 1–19 (2017)
25. Zheng, X., Lin, Z., Wang, X., Lin, K.J., Song, M.: Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowl.-Based Syst. **61**, 29–47 (2014). https://doi.org/10.1016/j.knosys.2014.02.003
26. Bross, J.: A distant supervision method for product aspect extraction from customer reviews. In: Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013, pp. 339–346 (2013)

27. Akhtar, N., Zubair, N., Kumar, A., Ahmad, T.: Aspect based sentiment oriented summarization of hotel reviews. Procedia Comput. Sci. **115**, 563–571 (2017). https://doi.org/10.1016/j.procs.2017.09.115
28. Xu, X., Wang, X., Li, Y., Haghighi, M.: Business intelligence in online customer textual reviews: understanding consumer perceptions and influential factors. Int. J. Inf. Manag. **37**(6), 673–683 (2017). https://doi.org/10.1016/j.ijinfomgt.2017.06.004
29. Rossetti, M., Stella, F., Zanker, M.: Analyzing user reviews in tourism with topic models. Inf. Technol. Tourism **16**(1), 5–21 (2016)
30. Kitchenham, B.: Procedures for performing systematic reviews, vol. 33, no. 2004, pp. 1–26. Keele University, Keele, UK (2004)
31. Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Systematic mapping studies in software engineering. In: EASE, vol. 8, pp. 68–77 (2008)
32. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37 (1996)
33. Periódicos CAPES (2017). https://www.periodicos.capes.gov.br/
34. Scopus: Scopus - Elsevier Database (2017). http://www.scopus.com/