WIREs Data Mining and Knowledge Discovery

Data and text mining from online reviews: An automatic literature analysis

Sérgio Moro[1] & Paulo Rita[2]

[1]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

[2]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Lisboa, Portugal

This is the peer reviewed version of the following article:

Moro, S., & Rita, P. (2022). Data and text mining from online reviews: An automatic literature analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. https://doi.org/10.1002/widm.1448

Funding Information

The work of Sérgio Moro was supported by Fundação para a Ciência e Tecnologia (FCT) within the following Projects: UIDB/04466/2020 and UIDP/04466/2020.
The work by Paulo Rita was supported by the Fundação para a Ciência e a Tecnologia (FCT) within the Project: UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC).

WIREs Wiley Interdisciplinary Reviews

Article Title: Data and text mining from online reviews: an automatic literature analysis

Article Type:

○ OPINION ○ PRIMER ○ OVERVIEW

◉ ADVANCED REVIEW ○ FOCUS ARTICLE ○ SOFTWARE FOCUS

Authors:

Sérgio Moro
ORCID=0000-0002-4861-6686
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal
E-mail: scmoro@gmail.com

Paulo Rita *
ORCID= 0000-0001-6050-9958
NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Lisboa, Portugal
E-mail: prita@novaims.unl.pt
*corresponding author

Abstract

This paper reports on a thorough analysis of the scientific literature using data and text mining to uncover knowledge from online reviews due to their importance as user-generated content. In this context, more than twelve hundred papers were extracted from publications indexed in the Scopus database within the last 15 years. Regarding the type of data, most previous studies focused on qualitative textual data to perform their analysis, with fewer looking for quantitative scores and/or characterising reviewer profiles. In terms of application domains, information management and technology, e-commerce and tourism stand out. It is also clear that other areas of potentially valuable applications should be addressed in future research, such as arts and education, as well as more interdisciplinary approaches, namely in the spectrum of the social sciences.

Keywords: online reviews; users' opinions; consumer feedback; data and text mining.

Graphical/Visual Abstract and Caption

Word cloud of online reviews literature.

1. INTRODUCTION

Online reviews have emerged from the dawn of the social media era to enable users worldwide to report their opinions about products and services (Viviani & Pasi, 2017; Zhao et al., 2020). Online reviews appear in many websites, including those that are specifically dedicated to sharing users' experiences and for which online reviews are a central element, such as TripAdvisor (Borges-Tiago et al., 2021), as well as those in which online reviews were introduced as an additional feature, but its users' adoption turned them into a feature of paramount importance (e.g., Amazon.com) (Kordrostami & Rahmani, 2020; Kundu & Chakraborti, 2020; O'Connor, 2010). Organisations have quickly understood the power of online reviews as a worldwide network of interconnected users who influence each other with their opinions. Thus, reviews have been harvested and harnessed to understand users' perspectives and behaviour (Eslami et al., 2018; Schuller et al., 2015). Organisations not only consume reviews but also reply to users, contributing to the production of large volumes of information (Meek et al., 2021; Xie et al., 2017).

Scholars have also acknowledged the potential of online reviews in a wide range of domains. These include, among others, Marketing and Psychology, with an emphasis on consumer behaviour (Chakraborty & Bhat, 2018; Wu, 2013), and Computer Science and Artificial Intelligence, with researchers taking advantage of the richness and volume of information to test algorithms and techniques (Saeed et al., 2020). Besides the standard freely written text and quantitative score provided by users, online review platforms also provide some user profile information (some mandatory, other optional) and contribution counters (e.g., number of reviews) as well as gamification features (badges awarded for specific accomplishments) (Moro et al., 2019a). Such information richness has been used in a wide range of studies with different goals. Nevertheless, studies aiming to analyse a large volume of online reviews require an automated approach based on data and text mining techniques.

Given the interest that online reviews have raised within the scientific community, the literature on related subjects has flourished. Subsequently, some studies have emerged that compile and analyse such a body of knowledge. From those, most of them are focused on marketing (e.g., Alalwan et al., 2017; Tuma & Decker, 2013) and electronic word-of-mouth (e.g., Cheung and Thadani, 2012; Verma & Yadav, 2021) perspectives. Thus, there is clear research gap in the scientific literature calling to the need for a refreshed analysis to address the following research problem: cover the vast body of knowledge, including the undertaken approaches to tackle volumes of data, provide a picture of the main research trends, and guidance on future directions that may enlighten researchers interested in this critical domain. Our literature analysis proposes an automated review of all articles covering online reviews, thus aiming to address the previously identified purpose. Hence, this research contributes not only to systematize the existing literature focused on online reviews but also to identify areas for future studies in order to raise the stakes of exploring as much as possible their highest potential for both academia and practice.

2. LITERATURE REVIEW

Online reviews have been studied from an information management perspective. For example, Yagci & Das (2018) used a design-level information quality measure to evaluate three information components, namely content, complexity and relevancy and found the number of reviews, sentences, words, noun words and feature matching noun words as key determinants. Further, Ma et al. (2017) developed a novel collaborative filtering approach to predict the overall ratings of entities to target users for personalised recommendations, which also demonstrated an advantage in dealing with sparse data. Ngo-Ye et al. (2017) focused on the analysis of the helpfulness of online reviews by incorporating higher-level knowledge constructs and using a scripts-enriched text regression model, which outperformed both the baseline and bag-of-words models not only in terms of accuracy but also demanding lowest training, testing, and feature selection times.

Moreover, Salehan & Kim (2016) studied the predictors of both readership and helpfulness of online reviews using a sentiment mining approach for big data analytics due to the challenge posed by the very nature of the 4 V´s (volume, variety, velocity and veracity). These authors found that not only reviews with higher levels of positive sentiment in their title received more readerships but also the length and longevity of a review positively influenced its readership and helpfulness. The latter has also been addressed by Singh et al. (2017), who developed models through machine learning to predict the helpfulness of reviews using textual features like ease of reading, subjectivity, polarity and entropy. In addition, online review helpfulness was studied by Karimi & Wang (2017), who found that reviewer profile image could significantly improve consumer evaluation of reviews helpfulness, which was also moderated by review length since the former served as a visual decoration creating affective responses instead of identity information.

Researchers have also used data and text mining techniques to analyse online reviews across a broad spectrum of application areas. One predominant area is hospitality and tourism. Calheiros et al. (2017) used a multitude of sources of customer reviews in order to perform text mining and sentiment analysis of an eco-hotel. Guo et al. (2017) were able to extract nineteen dimensions of visitor satisfaction from more than a quarter-million of online hotel reviews for over twenty-five thousand hotels in sixteen countries and also find differences among various demographic segments. Moro et al. (2017) used reviews from twenty-one hotels in the Las Vegas Strip to develop a predictive model for quantitative scores in TripAdvisor. Xiang et al. (2017) applied text analytics for comparing three leading online review platforms, specifically TripAdvisor, Expedia and Yelp, concerning information quality on all hotels situated in Manhattan. They found significant discrepancies in the way those hotels are representing throughout those platforms. Finally, Bilgihan et al. (2018)

analysed over two-thousand customer reviews in Yelp and identified both restaurant satisfiers and dissatisfiers impacting dining experience.

Another main area of application is marketing. Qi et al. (2016) mined online reviews to assess customer requirements aiming to learn how to design and improve products as well as develop appropriate strategies within the electronic commerce context. Zhang et al. (2018) also used online reviews as a source for product innovation with respect to Huawei smartphones and found a strong correlation between the change in the level of feature satisfaction and phone improvement. Zhang et al. (2019) used an online review-based approach to identify product features to be improved for product redesign and argued in favour of its combination with traditional survey data. Fan et al. (2017) combined real-world historical sales data with online reviews from the automotive industry and improved forecast accuracy of product sales. Chong et al. (2016) also predicted sales based on online reviews together with online promotional strategies by designing a big data architecture and using a neural network approach.

Given the relevance of research based on online reviews, several studies have recently emerged that aggregate and summarize existing literature to identify trends and research avenues. Table 1 summarizes eight literature review studies focused on online reviews. Three of them were published in 2020, showing an increasing need from scholars to understand the main trends in a mature topic. From the eight studies, only one of the most recent (i.e., Vanhala et al., 2020) adopted an automated approach to analyse a larger set through text mining. Six of the remaining seven adopted a systematic review by classifying the analysed body of knowledge. Given the increasingly more extensive body of knowledge, we adopt text mining and topic modelling, in contrast with the majority of studies. Also, the studies from Table 1 are focused on a specific subject (e.g., tourism, sales, fake reviews). Additionally, only three of them emphasise the automated approaches used by the analysed bodies of literature, which is our study's focus.

Table 1 - Literature analyses on automated approaches to analyse online reviews.

Reference	Scope	Nr. articles	Period
(De Maeyer, 2012)	Impact of online consumer reviews on sales and price strategies	*not specified*	*not specified*
(Tuma & Decker, 2013)	Online reviews as a source of marketing research data	204	1997-2012

(Balan & Mathew, 2015)	Online word of mouth using text mining	46	*not specified*
(Hlee et al., 2018)	Hospitality and tourism	55	2008-2017
(Reyes-Menendez et al., 2019)	To identify fake reviews in tourism	124	2014-2018
(Istanto et al., 2020)	Automatic SPAM detection	36	2015-2019
(Wu et al., 2020)	Fake reviews	165	2010-2019
(Vanhala et al., 2020)	Usage of large data sets in online consumer behaviour	495	2000-2021

In summary, contrary to our research, previous studies were focused on narrower topics, thus leaving a research gap to be covered with regard to performing a literature analysis about online reviews at large. Furthermore, with just one exception all of them did not take advantage of text mining and topic modelling which are used in the current research. It is important to underline that these techniques enable to discover topics occurring in a collection of documents aiming at exploring hidden semantic structures in the body of the texts. Text mining improves the search from literature bringing many advantages (Hashimi et al., 2015) such as to assist in extracting useful information from bulk of data efficiently, and to help create and build patterns from the provided data leading to the identification of increasing or decreasing trends.

3. METHODOLOGY

There are several scientific databases used on a worldwide scale. Some of the most recognized include Google Scholar (GS), Web of Science (WoS), and Scopus (Harzing & Alakangas, 2016). GS is a popular database that gathers all types of documents, which raises the challenge of finding scientific peer-reviewed literature on a given subject. Also, by indexing all sections of a document, it ends up retrieving documents that are sometimes not relevant (e.g., because a search term is mentioned in a reference title in the references list) (Harzing & van der Wal, 2008). In contrast, both WoS and Scopus are restrictive databases that only index sources that have been validated as relevant from a scientific perspective and that publish peer-reviewed manuscripts (Moro et al., 2019b). From the latter two, Scopus covers a wider range of sources. However, it has the disadvantage of also covering less relevant content, such as many conferences and short articles. Nevertheless, given the pros and cons, and also following the choice of Nobre & Tavares (2017) and Ahi & Searcy (2013), we chose Scopus to collect the articles.

To retrieve the relevant articles from Scopus, we defined a query string composed by broad scope terms related to data analysis, such as “text mining”, in combination with terms associated with online reviews. Since a search query is composed by a limited set of search terms combined with Boolean expressions, and it is difficult to encompass every term that may be used by distinct authors worldwide, we performed the search over the title, keywords and abstract through the TITLE-ABS-KEY Scopus function. By considering also the abstract, we included a larger body of text where it was more likely that any of the terms deemed as relevant in the search query might exist, enabling to capture a larger number of relevant articles. Thus, the query executed is as follows:

TITLE-ABS-KEY (("online reviews" OR "online comments" OR "online feedback" OR "online opinions") AND ("data mining" OR "text mining" OR "analytics" OR "data science" OR "sentiment analysis" OR "NLP" OR "natural language processing" OR "machine learning" OR "deep learning"))

The result was a total of 1,928 articles (as of November 2021). Considering the relevance of the Web of Science database, acknowledged by the scientific community, we executed a similar query in that database to assure no relevant publication was missing. We found none, which is justified by the broader scope of titles covered by Scopus in comparison. We chose not to use Google Scholar considering it covers too many non-peer-reviewed sources, including lecturing slides and teaching (i.e., non-scientific) books. Figure 1 shows the distribution of those articles throughout the years (i.e., period Jan/2005-Dec/2020). We omitted the 342 articles published in 2021 from Figure 1 because the full year was not considered. There is clearly an increasing interest across the 2005-2020 period.

Figure 1 - Distribution of selected articles.

Table 2 shows the distribution of articles per document type, considering the main types identified by Scopus, i.e., journal articles, conference papers, book chapters, and editorial notes. As research increases in maturity, the journal publications are also steadily increasing, while conference publications slightly decreased in 2021. Such results provide additional justification for the proposed literature analysis.

Table 2 - Distribution per document type.

Year	Journal article	Conference Paper	Book Chapter	Editorial Note	Total
2005	0	1	0	0	1
2006	1	3	0	0	4
2007	0	3	0	0	3
2008	1	4	0	0	5
2009	4	10	0	0	14
2010	1	25	0	0	26
2011	11	34	0	0	45
2012	20	39	1	0	60
2013	20	45	3	0	68

2014	32	44	2	0	78
2015	33	61	0	0	94
2016	51	71	4	0	126
2017	90	102	2	1	195
2018	97	131	0	1	229
2019	173	111	4	0	288
2020	211	135	4	0	350
2021	250	88	4	0	342
Total	995	907	24	2	1928

Table 3 exhibits the publishers most represented in terms of the number of publications. There is a prevalence of publishers related to IT, such as the IEEE and the ACM, suggesting that a large part of research uses online review domains to assess technological innovations.

Table 3 - Distribution per publisher.

Publisher	**Nr. Articles**
IEEE	344
Elsevier	253
Springer	247
Emerald	83

ACM	76
Taylor & Francis	76
SAGE	35
Association for Information Systems	29
MDPI	28
Wiley-Blackwell	22
Association for Computational Linguistics	18
Others	717
Total	**1,928**

The results in Table 4 show all sources that contribute to our set with 15 or more articles. Indeed, the sources that individually contribute the most are technology-related publications. Nevertheless, some important sources also emerge from tourism (e.g., International Journal of Hospitality Management, Tourism Management) and from electronic commerce (e.g., Electronic Commerce Research and Applications), which is justified by the fact that those are two of the main areas of application of online reviews, including large players such as TripAdvisor and Amazon.com.

Table 4 - Distribution per main source titles.

Source title	Nr. Articles
Lecture Notes in Computer Science	104
ACM International Conference Proceeding Series	48
Advances in Intelligent Systems and Computing	30
Communications in Computer and Information Science	28
Sustainability (MDPI)	28
International Journal of Hospitality Management	24
PervasiveHealth: Pervasive Computing Technologies for Healthcare	23
Decision Support Systems	22
Journal of Physiccs: Conference Series	19
IEEE Access	18
Tourism Management	17
International Journal of Contemporary Hospitality Management	16
Electronic Commerce Research and Applications	15
Others	1,536
Total	**1,928**

While some studies adopt each article's full text for analysis (e.g., Rezaeian et al., 2017), others only consider the usually indexed sections of each article (i.e., title, abstract and keywords; e.g., Moro et al., 2019a). We chose the latter since those sections tend to capture all relevant concepts underlined in the full article and constitute appropriate surrogates of full-text articles (Fontelo et al., 2013). Thus, after the previously described assessment, all titles and abstracts were gathered in a single dataset for textual analysis. Figure 2 exhibits the full procedure. We did not consider the articles´ keywords because we needed them for filtering, as explained next. The large corpus and volume of text justify an automated approach based on text mining to help summarise the selected literature (Delen & Crossland, 2008). Furthermore, the dimension of such corpus prevented us from performing a full reading of all articles, which might bring important insights by identifying common variables such as data size and type of data used from online reviews (e.g., review text, quantitative rating, helpful votes). However, text mining poses several challenges to overcome, most notably the large number of different (unique) words available in the total set of words for all the documents raises the difficulty for algorithms to find patterns between documents. While several techniques can be adopted to address such a challenge, we chose to follow the procedure suggested by Cortez et al. (2018), who also conducted an automated literature analysis and started first by gathering all keywords identified by the authors of each selected article. Then, in a similar approach to the previously mentioned authors, we selected only unique keywords and pruned the set by removing words such as "reviews" and "platform" and by building a dictionary of equivalent words (e.g., "user-generated content" and "consumer-generated content). Finally, at this stage, we computed a document-term matrix that contains the frequency of each of the selected keyword on each article.

Figure 2 - Undertaken procedure (based on Cortez et al., 2018).

To summarise the set of articles, we chose the popular Latent Dirichlet Allocation (LDA) topic modelling algorithm (Amado et al., 2018), which enables to gather the text in topics characterized by their proximity to each of the keywords (or terms) through a β distribution and by their proximity to each of the articles (or documents) through an α distribution. This algorithm takes as inputs the previously developed document-term matrix and the number of topics. As for the matrix, it is composed of a set of columns consisting in each of the selected keywords from the articles and a set of rows consisting in each of the 1,928 articles. Thus, cell (i, j) of the matrix contains the number of occurrences of keyword j within article i. The number of topics (the other input parameter to LDA) was set to 12 by computing it through the "ldatuning" package from the R statistical tool (Canito et al., 2018). Such package uses four metrics which compute the optimal number of topics (Nikita, 2016). We also adopted the R statistical tool for the remaining experiments, including the "topicmodels" package, which implements the LDA algorithm. Given that LDA returns a tri-dimensional structure (dimensions: topics, terms, articles) as output, we followed the suggestion of Moro et al. (2019a) by considering that each article was associated with the topic to which it was closer. This method enabled us to obtain a two-dimensional table to assess each topic's contents by looking at the terms that best characterize it.

4. RESULTS

In this section, we present the results of the topic model. A total of 12 topics were computed, following the procedure also undertaken by Moro et al. (2019a). We show the obtained results in Table 5. The topics are computed through topic modelling without human/expert intervention, i.e., based solely on the term frequency for the unique keywords existing in each article. Thus, Table 5 shows each topic per row, characterised by the four most relevant keywords borrowed from the total set of articles. The keywords are ordered from the most relevant (lowest β value per topic) to the fourth most relevant, with the β values also being exhibited. A β closer to zero implies a stronger relation between the topic and the keyword. We also show the distribution of articles most closely related to each topic per 4-year period (to assess trends through the analysed years) as well as the total number of articles in the last columns. The topics are labelled with numbers in the first column to facilitate referencing.

Topics are listed in descending order according to the number of articles within each of them (last column).

5. DISCUSSION

A first analysis highlights that most topics are equally characterised by the fourth most relevant terms since the β distribution values are quite similar. For example, the most relevant keyword for topic #1 is "satisfaction", with a β of 4.3, while the fourth is "management", with a β of 4.5. Such results show that the literature has a broad scope and is not clearly compartmentalised into distinct segments, i.e., topics are characterised by several terms. Also, such results denote the interdisciplinary nature of topics in general, including keywords from both IT and application domains within the four most relevant terms with similar β values. Topic #3 is one example, being characterised by "machine learning" (β=3.6) and "website" (β=3.8).

Table 5 - Unveiled topics.

#	Keywords								Articles				
	$Term_1$	β_1	$Term_2$	β_2	$Term_3$	β_3	$Term_4$	β_4	2005-2008	2009-2012	2013-2016	2017-2021	Total
1	satisfaction	4.3	responses	4.4	assessment	4.4	management	4.5	1	24	77	219	321
2	sentiment analysis	2.9	aspect	3.8	social media	4.0	negative	4.0	0	23	65	175	263
3	negative	3.4	machine learning	3.6	website	3.8	arts	3.8	1	13	25	125	164
4	negative	3.2	service	3.7	rating	3.8	e-commerce	4.0	1	3	23	132	159
5	automated	3.7	quality	3.8	satisfaction	3.9	text mining	4.0	1	13	24	119	157
6	sentiment analysis	3.1	phrase	4.0	quality	4.0	machine learning	4.0	0	9	26	121	156
7	quality	3.5	hotels	4.0	sentiment analysis	4.0	phrase	4.2	3	26	34	86	149
8	machine learning	3.4	sentiment analysis	3.6	negative	3.7	service	3.9	0	10	18	110	138
9	aspect	3.8	sentiment analysis	3.9	improvement	4.0	rating	4.0	0	7	18	90	115
10	aspect	3.6	influence	3.9	sentiment analysis	3.9	e-commerce	4.1	3	5	23	80	111
11	rating	3.5	sentiment analysis	3.7	aspect	4.1	hotels	4.1	0	8	17	81	106
12	sentiment analysis	3.1	service	3.9	process	4.1	China	4.2	3	4	16	66	89
									13	145	366	1,404	1,928

Table 5 also provides evidence that most researchers are adopting sentiment analysis, as it is the most relevant keyword for three among the total of 12 topics (#2, #6, #12). Further, sentiment analysis is also highlighted in five of the remaining nine topics (#7, #8, #9, #10, #11). Since sentiment analysis is performed over textual content, we can conclude that the richness of text is being exhaustively explored. In comparison, the quantitative rating is receiving less attention, being highlighted in topics #4, #9, and #11. Additionally, online reviews also provide other important information such as online reviewers' profiles (Lingras & Triff, 2014) and photos uploaded by users (Oliveira & Casais, 2014), which are still not mainstream enough to appear in any of the topics. This aspect is the first identified research gap that needs to be noted.

Although there is interdisciplinarity reflected in the topics through application and method keywords, there are clearly two domains of application that prevail in terms of volume of research over the remaining ones: tourism (i.e., through the keyword "hotels"), in topics #7 and #11, and e-commerce, in topics #4, and #10. This factor is a confirmation of the source titles of the selected publications, shown in Table 4. Besides those two, "arts" also emerge as a relevant domain of application. One example of the latter is the study by Menshikova et al. (2018), who analysed the comments of art critics written online. However, there are other domains known to be relevant but which are still not mainstream, including online comments of students about their schools and teachers (e.g., Santos et al., 2018) and online reviews about non-profit organisations (Feng et al., 2017). Also, the broader discipline of marketing, which encompasses applications within e-commerce and tourism, as highlighted in section 2, did not stand out in our results, suggesting that most contributions are more applied than theoretical. This important gap has also been recently highlighted by Moro et al. (2020) and should be addressed in the future to bridge between marketing theory and practice.

Interestingly, there are several trends focused on the negativeness associated with online reviews, namely reflected in four topics (#2, #3, #4, #8). While managers appreciate positive comments and tend to use them for improving brand image (Lien et al., 2015), a negative review can have a devastating effect, and managers are aware of such importance. Likewise, scholars have also devoted efforts to studying negativeness expressed in e-commerce context (topic #3) and social media in general (topic #2) (e.g., Lee et al., 2017). Also worth noting is the fact that China emerged as the only geography/language to appear among the four most relevant keywords in any topic (specifically, in topic #12).

Regarding the technical approaches, the topics unveil that machine learning and sentiment analysis are the most relevant of them. Aspect-based approaches are also important, topping topics #9 and #10, which consist of extracting specific features (aspects) from text to enable the identification of trends regarding those aspects (Azman et al., 2017).

Finally, the most represented topic (#1), with 321 associated articles, is devoted to assessing the responses to online reviews towards increased customer satisfaction, by far more than the second topic (with 263 articles). This aspect is an important domain of research, as managers strive to reply to users not only to thank users' compliments but also to show the wide online community that they care and address each user's complaint (Xie et al., 2017).

From a broad perspective, we can consider the keywords classifying the articles are either from a methodological perspective (e.g., "sentiment analysis") or from an application perspective (e.g., "hotels"). Thus, to further draw insights from the topics highlighted in Table 5, we plotted the concepts and their relations in two figures, 3 and 4. Each of the figures was drawn by including only terms from Table 5 that appeared together with others within a topic. This means that a term that occurred in isolation, such as "responses", was not drawn. The goal was to understand the relations between concepts. Thus, each line connects two concepts that are common in at least one topic (the numbers of topics are placed next to each line, reflecting the corresponding numbering system used in Table 5). Also, after each concept, the number of topics in which it occurs in Table 5 is shown.

Figure 3 - Methodological concepts and their relations.

Figure 4 - Application concepts and their relations.

Figure 3 highlights that most of the existing body of knowledge adopts sentiment analysis (8 topics from the total of 12). Considering that sentiment analysis aims at computing a sentiment score or polarity based on textual contents (Sharma & Jain, 2020), this result clearly shows that scholars are focusing their attention on the freely written textual opinions in online reviews, neglecting the remaining usual features that also compose an online review, such as the helpful votes and the date when the review was written (Moro et al., 2017). In fact, depending on the platform, an online review can conceal many quantitative features, which may also be analysed through data mining (Moro et al., 2020). The remaining methodological concepts that orbit around "sentiment analysis" include "aspects extraction", which is a subarea within "sentiment analysis", and "machine learning" that overlaps with "sentiment analysis" in areas such as document classification. As for the application domains (Figure 4), both "hotels" and "e-commerce" receive most of the attention. While e-commerce has a broader scope, the emphasis on "hotels" confirms that online reviews in the tourism and hospitality sector are widely accepted (Lee et al., 2017). According to our findings, online reviews have been assessed to understand perceived quality and satisfaction. This outcome provides evidence that application domain literature (especially within tourism and e-commerce) acknowledges the relevance of data and text mining techniques to uncover complex concepts within social sciences. Such a result shows that interdisciplinary research using data-driven approaches to online reviews is likely to continue to flourish in the next years as the world is flooded in data but hungry for knowledge.

From our findings, several takeaways can be highlighted both from a practical/professional perspective as well as from an academic one. First, the undertaken analytical process enabled to unfold and summarize a large body of knowledge of 1928 articles, constituting an efficient and objective mean of providing a view over existing literature. Second, we unveiled that e-commerce and the tourism industry are capturing most of the attention regarding data and text mining applied to online reviews. From a practical standpoint, as both e-commerce and tourism online platforms such as Amazon, E-bay, TripAdvisor, and Booking.com are evolving to include additional structured and unstructured information (i.e., gamified features to improve increase users' interactions – Moro et al., 2019b; photos and videos that complement online reviews – Song et al., 2021), studies that combine textual comments with image/video are likely to emerge as text and image mining techniques are becoming more easy to adopt due to a myriad of user-friendly tools made available in the most recent years (Trinks & Felden, 2020). Nevertheless, there are still challenges to solve in natural processing language, which is confirmed by our findings through words such as "semantic" and "emotion". From an academic perspective, the current lack of research applied to other domains of application creates future research avenues that scholars are likely to address in the near future. For example, in the educational domain, some studies are already being published that adopt online reviews that evaluate teachers and courses (Chou et al., 2021).

6. CONCLUSION

Online reviews have opened a myriad of possibilities for both researchers and practitioners to understand users' perspectives. Since social media emerged grounded on the Web 2.0 paradigm in which users are the main contributors of content to online platforms, online review systems have been developed to enable information exchange in several domains. Nevertheless, both e-commerce and tourism have been at the forefront of such development, including renowned platforms such as Amazon.com and Booking.com. This literature analysis took a broad approach by including all articles published in the last 15 years where advanced data and text mining approaches have been adopted to analyse the contents of online reviews, including both structured information (e.g., the score granted) and unstructured information (e.g., freely written comments/reviews). This study's first contribution relates to confirming e-commerce and tourism importance by highlighting both as mainstream research themes regarding online review analysis, which reflects the fact that the largest online review platforms are from both domains. Such a finding leads to uncovering research gaps in domains where online reviews are also widely used but to which scholars have still not devoted enough attention. One of those areas is education, with students currently sharing opinions and rating institutions, curricular programmes, or even teachers (Mensink & King, 2019). On the one hand, educational platforms that offer online review services are still lagging behind in terms of technological innovation when compared to e-commerce platforms. On the other hand, while competition between schools is fierce, it pales in comparison to tourism, for example, which is a main driver of economies worldwide, and thus many competitors compete for the same market shares (Lew, 2011). Another important finding relates to the type of data used and subsequent approaches to analyse it. In most cases, researchers have mostly focused on the textual content of the review itself, adopting methods such as sentiment analysis and aspect-based extraction, both from the text mining umbrella. Although the freely written text is significant because users can write their opinions, online review platforms offer a myriad of other valuable information, including a numeric rating of the product/service, which can be general or by subject (e.g., sleep quality and cleanliness, in the case of TripAdvisor), depending on the platform (Jia, 2020). Also, apart from the quantitative review scores granted, there is information characterising the users, which has been mostly neglected (e.g., the study by Moro et al., 2019b, is one of a few exceptions) and which can be used to characterise online reviewers in a similar way to what is done by researchers using primary data collection approaches such as survey questionnaires. The dissemination of online review platforms across other domains is likely to lead to new lines of research within those domains. The aforementioned insights provide ground for identifying some important contributions. The lack of a topic emphasizing other unstructured contents (such as photos and videos) besides textual comments suggests that: (1) there is still novelty to be found in analysing text from online reviews; and (2) given the interest and some recent research that analyses photos using image mining techniques (e.g., Song et al., 2021), future research is likely to combine image and text analysis to draw novel insights about the customers' perspectives. For example, a photo can help to determine if the user is being ironic (e.g., considering a sentence such as "This was pretty clean" and a photo that shows litter spread). Thus, by combining different formats of the contents published together with online reviews, we expect improved accuracy of data science applications to online reviews.

This study has several limitations that need to be mentioned. First, the search query definition is key to ensure that the relevant articles and only these are selected. We made an effort to include in the query only terms that undoubtedly retrieve relevant articles. Also, to make sure we retrieved the largest number of relevant articles, we performed the query considering also the abstract of each article, which is a more verbose section when compared to just the title and keywords, and where is more likely that the terms considered in the query might occur. Notwithstanding, as with any search engine, Scopus is limited to retrieving articles that used the terms we considered, which might result in failing to include some articles where only other lesser used expressions were adopted by the authors. Additionally, while Scopus is one of the most comprehensive databases within peer-reviewed scientific literature (Nobre & Tavares, 2017), some relevant sources are still not yet indexed by Scopus. Also, while adopting an automated approach helps in summarising the body of knowledge and obtaining a global picture about a subject, it also has some shortfalls compared to expert analysis. Namely, the complex semantic inherent of human language cannot be fully understood by topic modelling. Thus, to address such limitation, surveys can be adopted to understand professionals, scholars, and users' perspectives (Ramos et al., 2019). This methodology would enable cross-validation if the current scholars' perspective is reflected in the scientific body of knowledge already published while simultaneously verifying if there is an alignment between theory and practice (users and professionals).

Funding Information

The work of Sérgio Moro was supported by Fundação para a Ciência e Tecnologia (FCT) within the following Projects: UIDB/04466/2020 and UIDP/04466/2020.

The work by Paulo Rita was supported by the Fundação para a Ciência e a Tecnologia (FCT) within the Project: UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)

Data Availability Statement information

This study used information from articles which can be obtained from the Scopus database through the query provided in Section 3. It requires authorized access to the Scopus database.

References

Ahi, P., & Searcy, C. (2013). A comparative literature analysis of definitions for green and sustainable supply chain management. *Journal of Cleaner Production, 52*, 329-341.

Alalwan, A. A., Rana, N. P., Dwivedi, Y. K., & Algharabat, R. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics*, 34(7), 1177-1190.

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics, 24*(1), 1-7.

Azman, S. N., Ishak, I., Sharef, N. M., & Sidi, F. (2017). Towards an enhanced aspect-based contradiction detection approach for online review content. In *Journal of Physics: Conference Series* (Vol. 892, No. 1, p. 012006). IOP Publishing.

Balan, U. M., & Mathew, S. K. (2015). Online word of mouth using text mining: A review of literature and future directions. In 2015 *IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions* (WCI) (pp. 1-6). IEEE.

Bilgihan, A., Seo, S., & Choi, J. (2018). Identifying restaurant satisfiers and dissatisfiers: Suggestions from online reviews. *Journal of Hospitality Marketing & Management*, *27*(5), 601-625.

Borges-Tiago, M. T., Arruda, C., Tiago, F., & Rita, P. (2021). Differences between TripAdvisor and Booking. com in branding co-creation. *Journal of Business Research*, 123, 380-388.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, *26*(7), 675-693.

Canito, J., Ramos, P., Moro, S., & Rita, P. (2018). Unfolding the relations between companies and technologies under the Big Data umbrella. *Computers in Industry, 99*, 1-8.

Chakraborty, U., & Bhat, S. (2018). The effects of credible online reviews on brand equity dimensions and its consequence on consumer behavior. *Journal of promotion management*, 24(1), 57-82.

Cheung, C. M., & Thadani, D. R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems, 54*(1), 461-470.

Chong, A. Y. L., Li, B., Ngai, E. W., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies. *International Journal of Operations & Production Management*, 36(4), 358-383.

Chou, S. Y., Luo, J., & Ramser, C. (2021). High-quality vs low-quality teaching: A text-mining study to understand student sentiments in public online teaching reviews. *Journal of International Education in Business, 14*(1), 93-108.

Cortez, P., Moro, S., Rita, P., King, D., & Hall, J. (2018). Insights from a text mining survey on Expert Systems research from 2000 to 2016. *Expert Systems, 35*(3), e12280.

De Maeyer, P. (2012). Impact of online consumer reviews on sales and price strategies: A review and directions for future research. *Journal of Product & Brand Management, 21*(2), 132-139.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications, 34*(3), 1707-1720.

Eslami, S. P., Ghasemaghaei, M., & Hassanein, K. (2018). Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems*, 113, 32-42.

Fan, Z. P., Che, Y. J., & Chen, Z. Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research*, *74*, 90-100.

Feng, Y., Du, L., & Ling, Q. (2017). How social media strategies of nonprofit organizations affect consumer donation intention and word-of-mouth. *Social Behavior and Personality: an International Journal, 45*(11), 1775-1786.

Fontelo, P., Gavino, A., & Sarmiento, R. F. (2013). Comparing data accuracy between structured abstracts and full-text journal articles: implications in their use for informing clinical decisions. *BMJ Evidence-Based Medicine, 18*(6), 207-211.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483.

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787-804.

Harzing, A. W., & van der Wal, R. (2008). A Google Scholar H-Index for journals: A better metric to measure journal impact in economics & business. In *Proceedings of the Academy of Management Annual Meeting*. Wiley.

Hashimi, H., Hafez, A., & Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, 51, 729-733.

Hlee, S., Lee, H., & Koo, C. (2018). Hospitality and tourism online review research: A systematic analysis and heuristic-systematic model. *Sustainability, 10*(4), 1141.

Istanto, R. S., Mahmudy, W. F., & Bachtiar, F. A. (2020). Detection of online review spam: a literature review. In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology* (pp. 57-63).

Jia, S. S. (2020). Motivation and satisfaction of Chinese and US tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management, 78*, 104071.

Karimi, S., & Wang, F. (2017). Online review helpfulness: Impact of reviewer profile image. *Decision Support Systems*, *96*, 39-48.

Kordrostami, E., & Rahmani, V. (2020). Investigating conflicting online review information: evidence from Amazon. com. *Journal of Retailing and Consumer Services*, 55, (Jult), 102125.

Kundu, S., & Chakraborti, S. (2020). A comparative study of online consumer reviews of Apple iPhone across Amazon, Twitter and MouthShut platforms. *Electronic Commerce Research*, (July), 1-26.

Lee, M., Jeong, M., & Lee, J. (2017). Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website. *International Journal of Contemporary Hospitality Management, 29*(2), 762-783.

Lew, A. A. (2011). Tourism's role in the global economy. *Tourism Geographies, 13*(1), 148-151.

Lien, C. H., Wen, M. J., Huang, L. C., & Wu, K. L. (2015). Online hotel booking: The effects of brand image, price, trust and value on purchase intentions. *Asia Pacific Management Review, 20*(4), 210-218.

Lingras, P., & Triff, M. (2014). Fuzzy and crisp recursive profiling of online reviewers and businesses. *IEEE Transactions on Fuzzy Systems, 23*(4), 1242-1258.

Ma, Y., Chen, G., & Wei, Q. (2017). Finding users preferences from large-scale online reviews for personalized recommendation. *Electronic Commerce Research*, *17*(1), 3-29.

Meek, S., Wilk, V., & Lambert, C. (2021). A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews. *Journal of Business Research,* 125, 354-367.

Menshikova, A., Maglevanaya, D., Kuleva, M., Bogdanova, S., & Alekseev, A. (2018). Art Critics and Art Producers: Interaction Through the Text. In *International Conference on Digital Transformation and Global Society* (pp. 113-124). Springer, Cham.

Mensink, P. J., & King, K. (2019). Student access of online feedback is modified by the availability of assessment marks, gender and academic performance. *British Journal of Educational Technology, 51*(1), 10-22.

Moro, S., Esmerado, J., Ramos, P., & Alturas, B. (2020). Evaluating a guest satisfaction model through data mining. International Journal of Contemporary Hospitality Management, 32(4), 1523-1538.

Moro, S., Rita, P., Ramos, P., & Esmerado, J. (2019a). Analysing recent augmented and virtual reality developments in tourism. *Journal of Hospitality and Tourism Technology, 10*(4), 571-586.

Moro, S., Pires, G., Rita, P., & Cortez, P. (2019a). A text mining and topic modelling perspective of ethnic marketing research. *Journal of Business Research, 103*, 275-285.

Moro, S., Ramos, P., Esmerado, J., & Jalali, S. M. J. (2019b). Can we trace back hotel online reviews' characteristics using gamification features?. *International Journal of Information Management, 44*, 88-95.

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism Management Perspectives, 23*, 41-52.

Ngo-Ye, T. L., Sinha, A. P., & Sen, A. (2017). Predicting the helpfulness of online reviews using a scripts-enriched text regression model. *Expert Systems with Applications*, *71*, 98-110.

Nikita, M. (2016). Package ‘ldatuning’. Retrieved from https://cran.ms.unimelb.edu.au/web/packages/ldatuning/ldatuning.pdf.

Nobre, G. C., & Tavares, E. (2017). Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. *Scientometrics, 111*(1), 463-492.

O’Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management*, 19(7), 754-772.

Oliveira, B., & Casais, B. (2019). The importance of user-generated photos in restaurant selection. *Journal of Hospitality and Tourism Technology, 10*(1), 2-14.

Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Information & Management*, *53*(8), 951-963.

Ramos, R. F., Rita, P., & Moro, S. (2019). From institutional websites to social media and mobile applications: A usability perspective. *European Research on Management and Business Economics, 25*(3), 138-143.

Reyes-Menendez, A., Saura, J. R., & Filipe, F. (2019). The importance of behavioral data to identify online fake reviews for tourism businesses: A systematic review. *PeerJ Computer Science, 5*, e219.

Rezaeian, M., Montazeri, H., & Loonen, R. C. G. M. (2017). Science foresight using life-cycle analysis, text mining and clustering: A case study on natural ventilation. *Technological Forecasting and Social Change, 118*, 270-280.

Saeed, N. M., Helal, N. A., Badr, N. L., & Gharib, T. F. (2020). An enhanced feature-based sentiment analysis approach. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(2), e1347.

Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, *81*, 30-40.

Santos, C. L., Rita, P., & Guerreiro, J. (2018). Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. *International Journal of Educational Management, 32*(3), 431-447.

Schuller, B., Mousa, A. E. D., & Vryniotis, V. (2015). Sentiment analysis and opinion mining: on optimal parameters and performances. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5*(5), 255-263.

Sharma, S., & Jain, A. (2020). Role of sentiment analysis in social media security and analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(5), e1366.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, *70*, 346-355.

Song, Y., Wang, R., Fernandez, J., & Li, D. (2021). Investigating sense of place of the Las Vegas Strip using online reviews and machine learning approaches. *Landscape and Urban Planning, 205*, 103956.

Trinks, S., & Felden, C. (2020). A differentiation between Image Mining and Computer Vision in the application area of Big Data. In 2020 *IEEE International Conference on Big Data* (Big Data) (pp. 5839-5842). IEEE.

Tuma, M. N., & Decker, R. (2013). Online reviews as a source of marketing research data: a literature analysis. *International Journal of Electronic Marketing and Retailing, 5*(4), 287-316.

Vanhala, M., Lu, C., Peltonen, J., Sundqvist, S., Nummenmaa, J., & Järvelin, K. (2020). The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining–driven analysis of previous research. *Journal of Business Research, 106*, 46-59.

Verma, S., & Yadav, N. (2021). Past, present, and future of electronic word of mouth (EWOM). *Journal of Interactive Marketing*, 53, 111-128.

Viviani, M., & Pasi, G. (2017). Credibility in social media: opinions, news, and health information—a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(5), e1209.

Wu, P. F. (2013). In search of negativity bias: An empirical study of perceived helpfulness of online reviews. *Psychology & Marketing, 30*(11), 971-984.

Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, 132, 113280.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51-65.

Xie, K. L., So, K. K. F., & Wang, W. (2017). Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management, 62*, 101-110.

Yagci, I. A., & Das, S. (2018). Measuring design-level information quality in online reviews. *Electronic Commerce Research and Applications*, *30*, 102-110.

Zhang, L., Chu, X., & Xue, D. (2019). Identification of the to-be-improved product features based on online reviews for product redesign. *International Journal of Production Research*, *57*(8), 2464-2479.

Zhang, H., Rao, H., & Feng, J. (2018). Product innovation based on online review data mining: a case study of Huawei phones. *Electronic Commerce Research*, *18*(1), 3-22.

Zhao, Y., Wen, L., Feng, X., Li, R., & Lin, X. (2020). How managerial responses to online reviews affect customer satisfaction: An empirical study based on additional reviews. *Journal of Retailing and Consumer Services*, 57, (November), 102205.