doi.org/10.26398/asaproc.0064

Data science to support decision makers

Sandro Stancampiano
Istat, Rome, Italy

1. Introduction[1]

This paper aims to present a research that extracts information from reviews published on Google Maps. People use this application in order to discover new places and experiences. When people explore locations on Google Maps, they see content provided by other users – such as reviews and photographs – to get a better understanding of those places. On Google Maps, you can read and write reviews. Comments and opinions are voluntary. Google does not pay reviewers to add reviews to Google Maps. We assume that people write only if they genuinely want to share their point of view about a location. Furthermore, Google enforces a strict policy that aims to remove deceptive content, which can lead to misinformation, fake engagement, and misrepresentation. These removal measures help maintain reviews that are relevant, helpful, and trustworthy.

Through writing reviews, people function as human sensors, providing information based on their observations. With a vast number of reviews being published daily, attempting to read them all becomes an impossible task. This is where text mining comes into the picture.

Text mining enables us to convert unstructured data into actionable knowledge, which both administrators and citizens can utilize to make informed decisions. Textual information published on the Internet can serve as a significant foundation for projects aimed at alleviating statistical burden.

When analysing reviews published on Google Maps, we take into consideration not only the actual content (which includes topic and sentiment) but also other factors such as the user's rating, the timestamp of the review, its popularity (determined by the number of likes), and, of course, the specific location. For instance, when examining reviews about the Colosseum in Rome, we compare all the aforementioned data points. Our approach involves topic modeling to identify the main subjects being discussed. Subsequently, we search for pertinent text documents and compile a comprehensive report that can be utilized by public administrators.

2. Methodology

Topic modeling presents an unsupervised algorithmic approach for exploring collections of short texts. Among these algorithms, latent Dirichlet allocation (LDA) stands out, driven by two core principles: 1) each document comprises a blend of various topics, and 2) each topic consists of a mixture of words (Silge and Robinson, 2022).

In his paper, Blei outlines the objective of topic modeling: the automated discovery of topics within a document collection. He defines a topic as a distribution over a fixed vocabulary (Blei, 2011). This statistical methodology acknowledges the inherent presence of multiple topics in documents.

For this study, our objective is to determine the topics of discussion among tourists in their comments and to assess the utility of their observations. We have gathered a collection of reviews about the Colosseum spanning from 2019 to the first semester of 2023. Our focus centers on analysing 735 reviews from the year 2022.

The analysis consists of four phases:

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of Istat (Italian National Institute of Statistics).

1. Initially, a corpus of reviews is formed using the *quanteda*[2] package, and pre-processing operations are executed.
2. Following this, the LDA algorithm is employed to create a topic model. LDA objects from the R package are utilized for this purpose.
3. Subsequently, an exploration of word-topic probabilities and document-topic probabilities is undertaken.
4. Lastly, reviews with the highest probability of belonging to each identified topic are selected, based on the outcomes of the preceding step.

3. Results

In our analysis, we used LDA, which estimates each topic as a combination of various words. It also models each document as a blend of these topics. Hence, a document is not limited to a single topic, but rather comprises a mixture of multiple topics (Blei et al., 2003).

We have identified four distinct topics in the discussions and linked them with their respective ratings and likes. The topics are distinct and cover various aspects of the Colosseum. They range from management and organization of the visit to Colosseums, to expressions of feelings about it, and historical notes related to it. This variety suggests a comprehensive analysis covering practical, emotional, and historical perspectives.

Table 1 provides a detailed breakdown of our analysis, with the column "Number of documents (N. Doc) with Gamma > 0.9" specifically indicating the number of documents strongly related to each identified topic. The gamma (γ) parameter estimates what proportion of words in a document are attributed to a specific topic. For instance, if a document (d1) has a gamma value of 0.1 for topic 1 and 0.9 for topic 2, it is more associated with topic 2 than topic 1. Therefore, we infer that documents with a gamma value exceeding 0.9 for a particular topic are predominantly about that topic. The columns titled "average likes" and "standard deviation likes" quantify the level of engagement or popularity of each topic among the audience. The columns labelled "average rating" and "standard deviation of rating" represent the mean rating and the variability of ratings of a given topic, respectively. These metrics offer insight into the overall visitor satisfaction and the consensus level among the reviews for each topic.

The "expressing feelings" topic has the highest number of strongly related documents (100), suggesting it might be a popular aspect among discussant. The topic records the lowest average likes (0.5), despite its high number of related documents and high average rating. This suggests that while visitors frequently discuss their feelings and rate their overall visit experience highly, this aspect may not be as engaging or resonate as strongly with the broader audience. The average number of likes (0.5) is notably lower compared to other topics, which could indicate a lower level of engagement of these reviews among readers.

In contrast, "historical notes" has significantly fewer documents (26), which could imply it is a less represented topic. The limited number of reviews could also explain the lack of variability in the ratings. This topic shows exceptionally high appreciation with uniformly perfect ratings. However, the moderate and varied likes suggest that these reviews, similar to those in the "expressing feelings" topic, might not strongly engage other users. Among the documents showing a per-document-per-topic probability (referred to as 'gamma') of over 0.9 for a particular topic, the average rating undergoes an increase from 4.5 to 5. The average rating shows high appreciation for all topics.

A slight decrease in the average rating was observed between "historical notes" (topic 4) and "management and organization of the Colosseum" (topic 1). The average rating declined from 5 in topic 4 to 4.5 in topic 1, on a 1 to 5 scale. This decrease of half a point could suggest the presence of problematic issues in the documents associated with topic 1. Moreover, a standard deviation of around 0.9, while not excessively high, indicates a certain variety in opinions. Not all

[2] See https://quanteda.io for a complete description of quanteda R package.

reviews are uniformly positive. This variability could suggest the presence of specific aspect within the topic that do not fully satisfy some visitors.

The data for "management and organization of the Colosseum" indicates significant visitor engagement, as reflected by 93 documents with a high topic relevance ('gamma' value over 0.9). This aspect of the Colosseum experience received an average rating of 4.5 and an average of 1.1 likes, suggesting it is fairly well regarded among visitors.

While some topics are popular in terms of the number of reviews (such as topic 3), they do not always correspond to high engagement as measured in 'likes.' This could suggest that, although some subjects are commonly discussed, they may not necessarily evoke strong emotional or social reactions, such as 'likes.'

The greater variability in ratings and 'likes' in topics 1 and 2 might reflect broader differences in visitors' experiences and expectations regarding both the organization and management of the Colosseum and the visit itself.

Table 1: Topic analysis *vs.* rating and popularity

Topic	Description	N.Doc γ > 0.9	Avg ratings	Avg likes	Std Dev ratings	Std Dev likes
1	Management and organization of the Colosseum	93	4.5	*1.1*	*0.9*	*5.2*
2	Management and organization of the visit	84	4.9	*1.2*	*0.5*	*5.8*
3	Expressing feelings	100	4.9	*0.5*	*0.4*	*1.3*
4	Historical notes	26	5	*0.9*	*0.0*	*2.4*

Table 2: Topic 1 - Management and organization of the Colosseum

Review	Rating	Likes
Ovviamente ne vale la pena, prenotare online è necessario. State attenti a false guide e truffatori tutto intorno il Colosseo	5	0
Purtroppo non sono riuscita ad entrare, bisognava prenotarsi almeno 1/2 giorni prima... ma del resto come quasi tutti i punti culturali il weekend, anche se ad ingresso gratuito... suppongo siano restrizioni preventive per il covid quindi capisco	4	5
4 stelle perché il monumento ne merita 10 ma l'organizzazione è tristemente pessima. Varco d'entrata piccolo e ti respingono anche se arrivi come espressamente indicato entro 30 minuti dall'orario indicato. Praticamente convinse presentarti all'orario e basta. Ho trovato in molti punti del percorso più calca che nella metropolitana. Ho fatto altra coda per comprare l'audio guida (15 minuti di coda) ed è carina	4	1
Il posto non si deve neanche commentare, peccato per la gestione prenotazioni, un pò difficoltosa online, il sito fa pena! Potrebbero scrivere che non fanno più entrare nessuno da quando manca un'ora alla chiusura.	3	0
Gestito malissimo. Biglietto solo online quindi non acquistabile sul posto. File interminabili per l'accesso considerato che rispetto all'enorme flusso di turisti dovrebbero esserci almeno tre varchi. Insopportabile poi l'impertinenza di pseudo guide turistiche che offrono ad ogni passo presunte visite guidate e in alcuni casi anche biglietti introvabili online	1	3

The data suggest that while certain aspects of the Colosseum, like expressed sentiments and historical notes, are highly rated, others, such as its organization and management, elicit a wide range of opinions. This indicates potentially critical areas where improvements could significantly

enhance the overall visitor experience. Moreover, engagement measured in 'likes' does not always align with the frequency or positivity of the reviews, indicating diverse dynamics in how visitors interact with online reviews. However, the ratings also imply there might be room for improvement in this area to further enhance overall visitor satisfaction.

In Tables 2, 3 and 4, we show the most significant comments representing each subset within the four identified topics, as per our assessment. The first two groups share a similarity in their focus but approach the subject from distinct perspectives. The former group holds potential value for cultural heritage administrators (Table 2), whereas the latter group offers insights beneficial for other tourists seeking to enhance their visit experience (Table 3).

Table 3: Topic 2 - Management and organization of the visit

Review	Rating	Likes
Esperienza unica. La guida Simona simpaticissima e molto molto brava!! Ho capito tutto ciò che è stato spiegato, fantastico	5	0
È assolutamente da visitare…quanta storia.... fantastico	5	0
Visitato di notte (da fuori). Meraviglioso!	4	1
Il Colosseo è Roma Visitarlo almeno una volta nella vita è d'obbligo.	5	2
Piccolo consiglio: se si è di fretta non consiglio di visitarlo internamente, non offre la stesss magnificenza di fuori.	4	1

Table 4: Topic 3 - Expressing feelings

Review	Rating	Likes
Quando lo vedi rimani a bocca aperta ... Questa sarà la quinta volta ma la sensazione è sempre la stessa ... Imponente maestoso qualcosa di meravigliosamente bello .... Roma	5	1
Imponente, magico, incredibile. Uno se non IL punto di riferimento della città di Roma, dell'Italia e del Mondo. […]. Ancora più affascinante quando illuminato, di notte. Da vedere e rivedere, da ammirare. Tappa OBBLIGATORIA visitando la Città Eterna.	5	10
Roma e i suoi monumenti è bella tutto l' anno, il Colosseo sempre magico	5	0
Come recensire il Colosseo? Se andate a Roma non potete non visitarlo, un luogo straordinario e unico al Mondo, ricco di storia, molto più impressionante dal vivo di quanto non appaia nelle foto!	5	0
Oltre alla classica passeggiata interna giornaliera, ammirate questa meraviglia la sera l'atmosfera è stupenda.	5	0

Within the first subset, tourists express that there are significant issues with online booking, site organization, and dealing with unauthorized activities (false guides and scammers). On the other hand, the second group of reviews (Table 4) highlights tourists' recommendations for guided tours, noting the clear and comprehensive explanations provided during such tours.

The reviews categorized within the third group ("expressing feelings") delve into aspects such as strolls, sentiments and emotions. In these reviews, thoughts and personal reflections play a pivotal role. On the other hand, the fourth group adopts a didactic approach: Upon initial inspection, it appears that these tourists have studied the subject prior to (or possibly after) their visit. It is conceivable that they compose their comments while seating at their desks, referencing sources like Wikipedia or a Colosseum guide to provide historical and architectural insights.

Table 5: Topic 4 - Historical notes

Review	Rating	Likes
Uno dei luoghi più belli ed importanti da visitare a Roma. Situato nel centro della città di Roma, è il più grande anfiteatro del mondo. In grado di contenere un numero di spettatori stimato tra 50 000 e 87 000 unità, è il più importante anfiteatro romano, nonché il più imponente monumento dell'antica Roma che sia giunto fino a noi. […] La struttura esprime con chiarezza le concezioni architettoniche e costruttive romane della prima Età imperiale, basate rispettivamente sulla linea curva e avvolgente offerta dalla pianta ellittica e sulla complessità dei sistemi costruttivi. Archi e volte sono concatenati tra loro in un serrato rapporto strutturale.	5	0
È il simbolo di Roma l'eccellenza dei monumenti più visitati	5	0
È una delle costruzioni più spettacolari ed imponenti che ci siano. Situato al centro della città di Roma il Colosseo ne è il simbolo indiscusso. È il più grande Anfiteatro di romano al mondo. E' stato inserito fra le 7 meraviglie del mondo moderno nel 2007 e questo aiuta a ricordare la grandezza e i successi che l'impero romano era riuscito a raggiungere	5	4
È il monumento simbolo della Romanità, è stato l'ombelico del mondo romano, il centro della Storia, ce ne sono passate di storie tra queste mura.	5	0

4. **Deciphering tourist feedback: key conclusions**

We have embarked on an exploration of topic models as a solution for effectively managing an extensive repository of reviews focused on the Colosseum. The application of the LDA technique holds the potential to uncover genuine insights from these reviews, shedding light on various points of interest. This same methodology can be adapted to investigate, condense, and visually represent collections of reviews pertaining to other sites of significance.

The significance of topic modeling in this context cannot be overstated. It serves as a filter that sieves through the avalanche of reviews, enabling decision-makers to uncover prevalent sentiments, recurring topics, and emerging concerns. This distilled perspective eliminates the arduous task of manually sifting through an overwhelming volume of reviews, providing a macroscopic yet nuanced understanding of visitor impressions and concerns. Linking to our previous discussion, the document on the investigation initiated by the Italian Competition Authority (AGCM) into the ticket sales for the Colosseum Archaeological Park sheds light on issues related to topic 1[3].

The investigation addresses the difficulty consumers face in purchasing tickets from the official retailer, CoopCulture. Consumers struggle to buy tickets from the official seller due to bulk purchases by secondary sellers using automated systems. In topic 1, reviews feature complaints about ticket prices and difficulties in purchasing them online. This is echoed by recent media reports on the issue of secondary ticketing[4], which has also been addressed legislatively (Legislative Decree of March 7, 2023, No. 26, implementing EU Directive 2019/2161). These tickets are then resold at higher prices on platforms like Musement and GetYourGuide, often bundled with extra services. This practice prevents consumers from accessing standard-priced tickets, aligning with the complaints of high prices and online purchasing difficulties discussed in topic 1.

Moreover, topic modeling's capacity to adapt and evolve over time ensures its relevance in a dynamic landscape. This adaptability ensures that decision-makers remain informed about

[3] https://www.agcm.it/media/comunicati-stampa/2023/7/PS12603

[4] https://tg24.sky.it/economia/2023/07/19/bagarini-online-colosseo-antitrust
https://roma.repubblica.it/cronaca/2023/07/18/news/colosseo_biglietti_secondary_ticketing_antitrust_indagine_finanza_concorrenza-408160957/

shifting perceptions, allowing them to adjust strategies and prioritize interventions in alignment with visitor expectations.

In summary, topic modeling stands as an indispensable tool in the pursuit of comprehending the diverse tapestry of reviews about the Colosseum on platforms like Google Maps. It not only makes the seemingly insurmountable task of review analysis feasible but also enhances the decision-making process by providing a succinct and holistic overview of visitor sentiments. By harnessing the power of topic modeling, we transcend the limitations of information overload, ensuring that the voice of each tourist contributes meaningfully to the preservation, enhancement, and understanding of this iconic cultural heritage site.

Given the paramount importance of preserving Italy's cultural heritage, we recognize the necessity to harness every available resource for managing our historical legacy. Text mining emerges as a valuable component in this process, offering increased efficiency and the potential for seamless collaboration with complementary tools (Stancampiano, 2023). As part of our ongoing efforts, we are extending our study to encompass additional cultural assets, with the aim of providing administrators with the means to optimize their decision-making procedures.

References

Blei, D.M., Ng A.Y., Jordan M.I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, **3**, pp 993-1022.

Blei, D. (2011*). Introduction to Probabilistic Topic Models*, Communications of the ACM. 55.

Silge J., Robinson, D. (2017). *Text Mining with R: A Tidy Approach (1st. Ed.).* O'Reilly Media, Inc.

Stancampiano, S. (2023), Text mining e turismo culturale: l'analisi testuale delle recensioni, in *Il turismo culturale in Italia: analisi territoriale integrata dei dati*, Istat, Letture Statistiche: Territori, Roma, pp.116-130.