Big Data Research 37 (2024) 100474

Contents lists available at ScienceDirect

Big Data Research

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

Semi-supervised topic representation through sentiment analysis and semantic networks

Marco Ortu *, Maurizio Romano, Andrea Carta

Dept. of Business and Economics Sciences, University of Cagliari, Viale Fra Ignazio 17, Cagliari, Italy

ARTICLE INFO

Keywords:
Semi-supervised clustering
Topic modeling
Natural language processing
Threshold-based naïve Bayes classifier

ABSTRACT

This paper proposes a novel approach to topic detection aimed at improving the semi-supervised clustering of customer reviews in the context of customers' services. The proposed methodology, named SeMi-supervised clustering for Assessment of Reviews using Topic and Sentiment (SMARTS) for Topic-Community Representation with Semantic Networks, combines semantic and sentiment analysis of words to derive topics related to positive and negative reviews of specific services. To achieve this, a semantic network of words is constructed based on word embedding semantic similarity to identify relationships between words used in the reviews. The resulting network is then used to derive the topics present in users' reviews, which are grouped by positive and negative sentiment based on words related to specific services. Clusters of words, obtained from the network's communities, are used to extract topics related to particular services and to improve the interpretation of users' assessments of those services. The proposed methodology is applied to tourism review data from Booking.com, and the results demonstrate the efficacy of the approach in enhancing the interpretability of the topics obtained by semi-supervised clustering. The methodology has the potential to provide valuable insights into the sentiment of customers toward tourism services, which could be utilized by service providers and decision-makers to enhance the quality of their services.

1. Introduction

Network analysis-based topic detection has recently emerged as an alternative approach to the widely-used Latent Dirichlet Allocation (LDA) method for topic mining in Natural Language Processing (NLP) field [16,13]. While LDA has been considered the most advanced tool in this field, further extensions have aimed to improve topic coherence using various information from document collections to detect topics. In this study, we investigate language models, such as word embeddings [18], to construct a topic model based on a semantic network of words, and to exploit these methodologies, we focus our work on topic modeling for semi-supervised clustering of online tourism review data.

In fact, the analysis of online reviews has become a crucial tool and a source of information for the decision process, to interpret and understand customers' assessment of the quality of services and products, which influences the reputation of businesses. Focusing on word network-based topic detection, our approach aims to handle sparse and imbalanced text representations, without relying on special assumptions about the pre-defined number of topics, which is one of the main flaws of unsupervised topic modeling. Nowadays, it is a common activity for businesses to analyze customer feedback to gain insights into the sentiment and opinions of their customers [21]. Topic modeling is one of the most popular methods to analyze documents (in our case tourist customer reviews) and to cluster them to understand which groups of reviews share similar content, in order to ease the interpretation of the overall sentiment of customers toward a product or a service. The interpretation of customer feedback expressed in online reviews is a challenging task due to the high dimensionality of the textual data and the intrinsic ambiguity and subjectivity of the language used in the reviews. Topic modeling methodology, such as LDA, has emerged as a promising approach to tackle such challenges. Topic modeling is an unsupervised clustering technique that automatically identifies latent topics present in a large collection of documents. By identifying topics, topic modeling can reduce the dimensionality of the data and provide a more interpretable representation of the reviews.

Although the effectiveness of topic modeling in clustering online users' reviews has been proved in several domains, such as retail and tourism feedback, there are still some challenges in applying it to real-

* Corresponding author.
E-mail addresses: marco.ortu@unica.it (M. Ortu), romano.maurizio@unica.it (M. Romano), andrea.carta88@unica.it (A. Carta).

https://doi.org/10.1016/j.bdr.2024.100474
Received 6 September 2023; Received in revised form 22 March 2024; Accepted 8 June 2024
Available online 13 June 2024

world datasets. For instance, the limited availability of labeled data sets makes it difficult to train supervised models with sufficient quality for industrial applications. For this reason, semi-supervised methods have been proposed to address this challenge, which leverages both labeled and unlabeled data to improve the accuracy of clustering [8]. In this paper, we propose a novel methodology for topic detection designed to improve semi-supervised clustering of users' reviews with an application for tourism review data. The proposed methodology, called SeMi-supervised clustering for Assessment of Reviews using Topic and Sentiment (SMARTS) for Topic-Community Representation with Semantic Networks, leverages an ensemble of semantic network and sentiment analysis for semi-supervised clustering of reviews to obtain interpretable topics.

Our methodology exploits the construction of a semantic network of words, based on word embedding, to identify the semantic similarity between different words used in the reviews. The semantic networks are constructed using two subsets of reviews, grouped by their sentiment (positive and negative), and topics are then identified using community detections algorithms. The results of the sentiment analysis are used to improve the interpretation of quality assessment expressed by customers in online reviews of specific services.

The proposed methodology is applied to a dataset of tourism reviews, extracted from Booking.com, and our findings show the effectiveness of our approach detecting interpretable topics. The proposed methodology could be used to provide insights into the sentiment of customers towards products and services and could support decision-making processes.

The paper is organized as follows: Section 2 describes the current literature, Section 3 illustrates in detail the SMARTS methodology, Section 4 shows the results of the case-study, while Section 5 draws the conclusion and paves the way for future works.

2. Background

Paving the way for the proposed methodology, this section is composed of three parts that discuss the main features of three key building blocks of the proposal. Section 2.1 and Section 2.2 refer to Topic Modeling and Sentiment Analysis, respectively, briefly describing the most important concepts and some related work. Finally, Section 2.3 illustrates how Network Analysis and Modularity Clustering are linked, focusing on Overlapping Community Detection algorithms.

2.1. Topic modeling

Topic Modeling is an unsupervised learning method for identifying underlying themes and patterns in a set of documents and facilitating their representation based on the frequency of words that comprise them. One of the most popular methods of topic modeling, which has the advantage of working with large groups of text documents, is LDA [2]. This method is a hierarchical Bayesian model in which every item in a corpus is represented as a finite combination of latent topics. Each topic is created by combining an infinite number of latent probabilities of words that directly represent documents [2]. One of the key advantages of this approach is the ability to maintain contextual usage of distinct terms, as the original terms are preserved in the analyses, resulting in much more interpretable outcomes.

Topic modeling is a powerful approach that has reshaped the NLP, its ability to detect latent topics in document corpora has made it an indispensable tool for text analysis in a variety of fields, and it is expected to increase its role significantly as the amount of digital data produced every day grows. Topic modeling has an extensive spectrum of applications, including social networking research, information retrieval, and market analysis. It can, for example, assist in identifying themes or sentiments displayed in tweets or comments, improve search accuracy by recognizing relevant themes or topics in documents, and deliver insights into consumer preferences and trends [17,15]. Recent researchers have focused on exploiting deep neural networks for extracting topics [35,9].

2.2. Sentiment analysis

Sentiment analysis is a method that uses NLP to examine and extract subjective texts containing user opinions, preferences, and sentiments. This sort of analysis can be implemented at different degrees of granularity, such as an entire document or the individual words that define it as an entity.

Sentiment Analysis is commonly implemented using machine learning techniques (typically in a supervised learning setting) by training a model with an assigned natural language text and then classifying a text as a negative, positive, or neutral sentiment. Sentiment Analysis serves many applications (e.g., client satisfaction, social media monitoring) with numerous business, executive, political, and academic consequences [5,6,30,29]. Furthermore, Sentiment Analysis is rising in importance as a result of the increased availability of massive amounts of textual data generated by social media. Researchers are still developing new methodologies while improving existing ones (see [23,21]). Nonetheless, because words can have distinct meanings in distinct situations, this area has numerous challenges. Challenges such as detection of other forms of language (such as sarcasm or irony), sentiment subjectivity, and topic comprehension based on the reader's background knowledge (cultural). How it will be better explained in Sec. 3, despite it would be possible to use any type of classifier (e.g., Random Forest, Support Vector Machine, standard Naïve Bayes), the default proposal focus on the use of the classifier defined by [23], called Threshold-based Naïve Bayes (Sec. 3.1).

2.3. Network analysis and community detection

Network analysis is a multidisciplinary field of research that investigates the structure and behavior of complex systems that can be represented as networks, which are a collection of nodes or entities that are connected by a set of relationships or edges [33]. The central premise of network analysis is that analyzing the relationships between entities generates better explanations for certain phenomena with respect to considering individual entities in isolation. It involves evaluating patterns of connections, interactions, and information flow within and between networks [11]. Finally, by representing phenomena as networks, we can study the mathematical properties of their structure, and typically we quantify these properties using network metrics, such as the centrality measures of degree, closeness, and betweenness [7,1]. It is also possible to apply clustering algorithms to a network, and, an interesting application is Modularity Clustering.

Network Modularity clustering [20] is an unsupervised machine learning technique used to identify extremely linked communities of nodes across a complex network. Therefore, this approach finds a set of clusters that maximizes modularity, a clustering metric, which has a scale of 0 to 1, with 0 suggesting a random structure and 1 corresponding to a strong community structure; however, analytically, these values usually lie in a smaller interval [19]. In order to maximize the modularity measure, modularity clustering algorithms iteratively split the network into smaller subgroups according to the connectivity patterns of the nodes until a stopping condition is met. There are several modularity clustering algorithms (see [14] for an overview), and one of the most popular is the Louvain method [3], which has the advantage of high efficiency and scalability. Network Modularity clustering has a wide range of applications in fields such as social network analysis, biology, finance, and web mining. However, many real world networks have complex structures in which nodes can have multiple roles or memberships [4]. For this reason, in recent years overlapping community detection algorithms have been introduced.

Traditional community detection algorithms, such as Network Modularity clustering, usually assign each node to a single community,

presuming that nodes can only belong to one group [22], whilst, overlapping community detection, seeks to identify groups or communities within a network where nodes may belong to multiple communities at the same time [27]. There are many approaches to overlapping community detection, such as graph neural networks [36], link-based methods [28], and algorithms centered around node importance and random walks [32]. These techniques aim to detect communities while requiring minimal computation, making them appropriate for large-scale complex network analysis [26]. Thus, multiple comparative studies [10,34] emphasize the importance of comprehending the structural properties of communities in order to design more efficient community detection methods [31].

3. Methodology

This section deeply formalizes all the methodological key points and is organized as follows: Section 3.1 recalls the Threshold-based Naïve Bayes classifier by describing its main features and strengths and how is applied for sentiment analysis. Section 3.2 introduces the SMARTS methodology algorithm.

3.1. Threshold-based naïve Bayes classifier

Recalling that it would be possible to use any type of classifier (like those mentioned in Sec. 2.2), the default proposal focus on the use of the classifier defined by [23]. Highlighting that Threshold-based Naïve Bayes (Tb-NB) can be applied when dealing with a labeled context, we hereby briefly describe the previously mentioned classifier.

Considering a collection of n_D documents (i.e. reviews), within a labeled context each document d_j is a priory known as a positive (d_j^+) or negative (d_j^-). Notationally, the set of documents $\mathcal{D} = \{d_1, \dots, d_j, \dots, d_{n_D}\}$ is split into a training set of size n_d, and a test set of size $n_D - n_d$. Following a preprocessing step that consists in removing stopwords, punctuations, and all non-alphabetic and non-relevant character, all the n_w words included in the n_D documents are collected in a Bag-of-Words (BoW) $\mathcal{W} = \{w_1, \dots, w_i, \dots w_{n_w}\}$.

Considering a probability function $\pi(\cdot)$, Tb-NB builds on the Bayes' rule and computes a scoring function $\Lambda(\cdot)$ for all the n_d included in the training set in order to predict if a document d_j ($j = 1, \dots, n_d$) that contains a certain word $w_k \in \mathcal{W}$ has a negative or positive sentiment:

$$\begin{aligned}\Lambda\left(d_j|w_k\right) &= \log\left[\frac{\pi(d_j^+|w_k)}{\pi(d_j^-|w_k)}\right] = \\ &= \log\left[\frac{\pi(w_k|d_j^+)}{\pi(w_k|d_j^-)} \cdot \frac{\pi(\bar{w}_k|d_j^+)}{\pi(\bar{w}_k|d_j^-)} \cdot \frac{\pi(d_j^+)}{\pi(d_j^-)}\right] = \\ &= \underbrace{\left[\log \pi(w_k|d_j^+) - \log \pi(w_k|d_j^-)\right]}_{\mathcal{L}(w_k)} \\ &\quad + \underbrace{\left[\log \pi(\bar{w}_k|d_j^+) - \log \pi(\bar{w}_k|d_j^-)\right]}_{\mathcal{L}(\bar{w}_k)} \\ &\quad + \left[\log \pi(d_j^+) - \log \pi(d_j^-)\right] \approx \\ &\approx \mathcal{L}(w_k) + \mathcal{L}(\bar{w}_k) \end{aligned} \tag{1}$$

As shown in [23], $\Lambda(d_j|w_k)$ derives from the sum of two components: a function $\mathcal{L}(w_k)$ that measures how likely a specific word w_k is present in a document, and a function $\mathcal{L}(\bar{w}_k)$ that measures how likely w_k is not present in the same one. These two functions derive from the log-likelihood ratio of the event ($w_k \in d_j$) and ($w_k \notin d_j$), respectively.

Equation (1) allows us to understand if a document d_j has a negative (positive) sentiment by computing the scoring function $\Lambda(d_j)$ for all the M words included in its content:

Table 1
Algorithm symbols.

Symbol	Meaning
S	Sentiment function
Sim	Similarity function
$Rank$	Ranking function
$\mathcal{F}_c$	Network decomposing function for communities detection
$\mathcal{D}$	Documents
$\mathcal{W}_d$	Word vector
$\mathcal{W}_e$	Embeddings vector
$e_{i,i+1}$	Semantic similarity between w_i and w_{i+1}
$\mathcal{N}$	Network
K	List of numbers of clusters
$Cvrg$	List of coverages
C_k	Coverage of the k_{th} community
ϕ_s	Detected community representing a topic
Φ_s	Set of all nodes in ϕ_s

$$\begin{aligned}\Lambda\left(d_j\right) = \Lambda\left(d_j|w_1, \dots, w_m, \dots, w_M\right) = \sum_{m=1}^{M} \Lambda\left(d_j|w_m\right) = \\ = \sum_{m=1}^{M} \mathcal{L}(w_m) + \mathcal{L}(\bar{w}_m)\end{aligned} \tag{2}$$

with $(w_1, \dots, w_m, \dots w_M) \in d_j \in \mathcal{W}$. Thus, Tb-NB proceeds by computing $\mathcal{L}(w_k)$ and $\mathcal{L}(\bar{w}_K)$ (Eq. (1)) for all the words $w_k \in \mathcal{W}$ and next it aggregates the computed quantities with respect to the set of words included in each document d_j to obtain $\Lambda(d_j)$ according to Eq. (2).

Once the set of scores $\Lambda\left(d_j\right)$ is computed, a decision rule $\mathcal{T}$ has to be defined in order to classify the comment d_j^* included in the test set, as positive (+1) or negative (-1). Thus, $\mathcal{T}$ is defined based on the estimated value of the threshold parameter τ corresponding to a specific value of $\Lambda(\cdot)$:

$$\mathcal{T}_{d_j^*}: \begin{cases} \Lambda\left(d_j^*\right) > \hat{\tau} & \rightarrow d_j^* = +1 \\ \Lambda\left(d_j^*\right) \leq \hat{\tau} & \rightarrow d_j^* = -1 \end{cases} \qquad (j^* = n_d + 1, \dots, n_D) \tag{3}$$

The threshold τ is the unique parameter of the Tb-NB classifier, which is estimated from the training data (for instance the value minimizing the Type I error).

3.2. Topic-community representation with SMARTS

SMARTS methodology consists of four main phases represented in Fig. 1: i) Natural Language text pre-processing; ii) Vectorization of textual data using word embedding and Sentiment analysis; iii) Semantic similarity network construction; iv) Topic extraction and words ranking.

Algorithm 1 shows the SMARTS methodology in detail (see also Table 1). In the preprocessing phase, the document set $\mathcal{D}$ underwent stopword removal, punctuation elimination, and the exclusion of non-alphabetic and irrelevant characters, resulting in the word vector $\mathcal{W}_d$. Subsequently, $\mathcal{W}_d$ was processed through Word Embeddings [18] using the SpaCy library [12], which incorporates a pre-trained model for the Italian language, to derive the embeddings vector $\mathcal{W}_e$. These word embeddings effectively capture semantic relationships between words. This capability is leveraged to construct a semantic network, where nodes represent individual words and edges denote the semantic similarities between these words, as inferred from their embeddings. The sentiment category is assigned to the $\mathcal{W}_e$ word embeddings using the S function in this phase.

The Semantic Network is created using words as nodes and the semantic similarity as weight denoted by $e_{i,i+1}$. In each review, we construct a node for every word, forming the basis of the network. To ensure sparsity within this network, edges are strategically assigned only between adjacent words. This is achieved through the semantic similarity function, $Sim(w_i, w_{i+1})$, which calculates a weight for

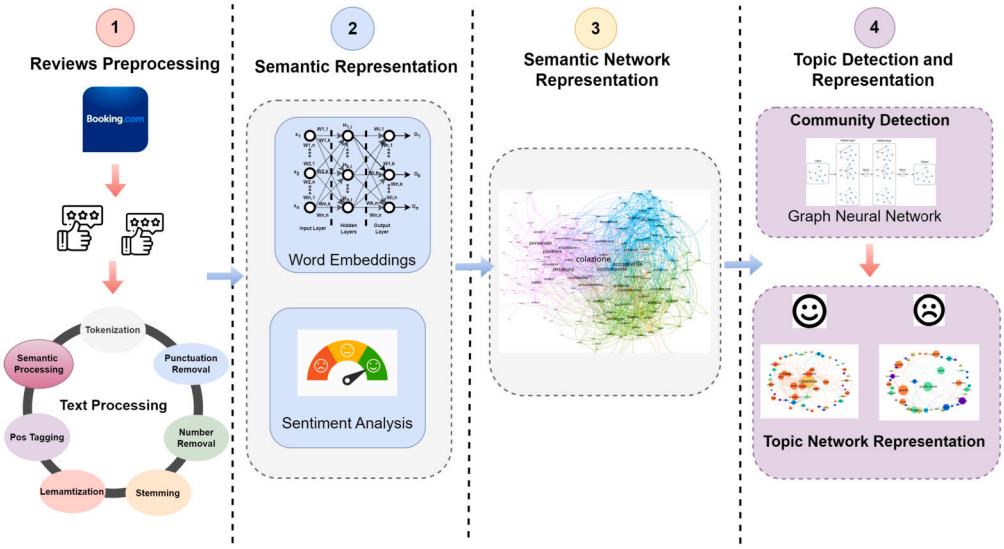

Fig. 1. SMART architecture.

Algorithm 1 SMARTS methodology algorithm.

Require $\mathcal{D}$: Documents (i.e. reviews) set of size n_D;
Require $\mathcal{S} : d \rightarrow s$; A function that assign a sentiment s to a document in D;
Require $Sim : (w_i, w_j) \rightarrow \Re$; A function to compute a semantic similarity of words (w_i, w_j);
Require $Rank : \mathcal{N} : \rightarrow \Re^z$; A function that ranks all nodes of a network $\mathcal{N}$ of size z;
Require $\mathcal{F}_c : (\mathcal{N}, k) \rightarrow (C_1, \ldots, C_k)$: A function that decomposes a network $\mathcal{N}$ into k communities C_i;
Step 1:
  Input: $\mathcal{D} = \{d_1, \ldots, d_{n_D}\}$
  Output: $\mathcal{W}_d = \{w_1, \ldots, w_{n_D}\}$
Step 2:
  Input: $\mathcal{W}_d = \{w_1, \ldots, w_{n_D}\}$
  Output: $\mathcal{W}_e = \{w_{e,1}, \ldots, w_{e,n_D}\}$
  Output: $\mathcal{S}_e = \mathcal{S}(\mathcal{W}_e)$
Step 3:
    $\mathcal{N} \leftarrow \emptyset$
    for $s \in S_e$ **do**
      for $w_{e,i} \in \mathcal{W}_e$ **do**
        $e_{i,i+1} \leftarrow Sim(w_i, w_{i+1})$
        $\mathcal{N} \leftarrow \mathcal{N} \cup (w_{e,i}, w_{e,i+1}, e_{i,i+1})$
      end for
    end for
Step 4:
  Input: w_s
  $K \leftarrow (2, \ldots, K_{max})$
  Step 4.1:
    $Cvrg \leftarrow \emptyset$
    for $k \in K$ **do**
      $\mathcal{N} \leftarrow C_1 \cup \cdots \cup C_k \leftarrow \mathcal{F}_c(\mathcal{N}, k)$
      $Cvrg \leftarrow Cvrg \cup \text{Coverage}(C_1, ..., C_k)$
    end for
    Output: $k_{best} \leftarrow argmax(Cvrg)$
  Output: $\mathcal{N}_{s,k} \subseteq \mathcal{N}$
  Output: $\phi_{s,k} : \mathcal{N}_{s,k} \rightarrow Rank(\Phi_{s,k})$

a word w_i and its immediate successor w_{i+1}. Consequently, this results in the formation of an edge $(w_i, w_{i+1}, e_{i,i+1})$ in the network, linking node w_i to node w_{i+1}, with the edge weight denoted by $e_{i,i+1}$. This ensures that the corresponding network will be more coherent as words will be connected with other words that appear in similar contexts.

In the last phase, two separate networks are created using positive and negative reviews. For a given word w_s representing a specific service (such as "Wi-Fi" or "swimming pool"), a subnetwork of $\mathcal{N}$ is selected considering all adjacent nodes to the specific word such as that $\mathcal{N}_s \subseteq \mathcal{N}$. Next, the subnetwork is clustered using the overlapping community detection method proposed by [27]. Here each detected community, denoted by ϕ_s, represents a topic (words are represented by nodes). Each node of the topic subnetwork is then ranked, using the word's degree centrality in the detected community as the $Rank(\Phi_s)$ function, where Φ_s is the set of all nodes in ϕ_s.

To quantify the goodness of the detected communities, we used the following unsupervised metrics [27]. Coverage, measures the percent-

age of the edges explained by at least one community (i.e. if (u, v) is an edge, both nodes share at least one community).

$$\text{Coverage}(C_1, ..., C_k) = \frac{1}{|E|} \sum_{u,v \in E} 1[z_u^T z_v > 0] \tag{4}$$

Density, measures the average density of the detected communities (weighted by community size): $\rho(C) = \frac{\#\text{ existing edges in } C}{\#\text{ of possible edges in } C}$.

$$\text{AvgDensity}(C_1, ..., C_k) = \frac{1}{\sum_i |C_i|} \sum_i \rho(C_i) \cdot |C_i| \tag{5}$$

Conductance, average conductance of the detected communities (weighted by community size).

$$\text{outside}(C) = \sum_{u \in C, v \notin C} A_{uv} \tag{6}$$

$$\text{inside}(C) = \sum_{u \in C, v \in C, v \neq u} A_{uv} \tag{7}$$

$$\text{AvgConductance}(C_1, ..., C_k) = \frac{1}{\sum_i |C_i|} \sum_i \text{Conductance}(C_i) \cdot |C_i| \tag{8}$$

Clustering coefficient: average clustering coefficient of the detected communities (weighted by community size).

$$\text{ClustCoef}(C) = \frac{\#\text{ existing triangles in C}}{\#\text{ of possible triangles in C}} \tag{9}$$

$$\text{AvgClustCoef}(C_1, ..., C_k) = \frac{1}{\sum_i |C_i|} \sum_i \text{ClustCoef}(C_i) \cdot |C_i| \tag{10}$$

4. Motivating example: booking.com Italian reviews

The proposed methodology has been applied to the Booking.com reviews data of Italian tourism facilities. The results showed how our approach detected the interpretable topics obtained by the semi-supervised clustering. Data from Booking.com has been collected with web-scraping made by an ad-hoc Python extractor and concerns 619 Sardinian hotels, 106,800 reviews (4/5 Italian – 1/5 English) from January 3rd, 2015 to May 27th, 2018, and their polarity (62,291 positive, 44,509 negative). Booking.com has been chosen for two main reasons: only real guests are allowed to create a review, and each one is made of one positive section and one negative section. We hereby considered the positive (negative) section as a single positive (negative) review, knowing a priori the polarity of each review. That permits to work within a supervised framework.

We extracted the topics related to specific services using the clusters of words obtained from the semantic network. We used the sentiment scores, obtained by training the Tb-NB (Sec. 3.1) with this data, to interpret users' assessment of specific services by constructing a specific semantic network for positive and negative reviews.

4.1. Topic detection results

Table 2 and 3 show what we found when we looked at customer reviews from Booking.com. We used three different ways to find topics in the reviews: Louvain's Community Detection, Overlapping Community Detection, and Latent Dirichlet Allocation (LDA). Each row in the table is a topic that we found. We looked at specific services like 'wifi' or 'breakfast' and whether the sentiment was 'positive' or 'negative'. We used the top five keywords for each method to describe the topics. Table 2 and 3 let us compare the topics that the different methods found. We selected Latent Dirichlet Allocation (LDA) as the comparison baseline, given its widespread recognition and use in topic modeling. To ensure optimal performance and a fair comparison, we fine-tuned LDA's parameters. This fine-tuning enable LDA to effectively identify the most relevant topics. The degree of similarity between topics generated by LDA and our proposed method was assessed by examining the overlap in keywords identified by each technique. We can do this by looking at the keywords and seeing how much they overlap between the methods. The more overlap, the more the methods agree on what the topics are. We also looked at how many different words the methods used. This is called lexical diversity, and it's calculated by dividing the number of different words by the total number of words. If the lexical diversity is high, it means the method uses a wide range of keywords to describe the topics. Fig. 2 shows the results, showing the overlapping community detection as a good overlapping of terms identified with LDA.

Fig. 3 shows the lexical diversity of the obtained topics for the three methods considered in our model. We can see that, in general, the network-based methods produce higher lexical diversity, in particular the Louvains' community detection. After inspecting the generated topics, we proceed with our in-deep analysis considering the overlapping community detection since it is able to detect more relevant words for the services analyzed.

To figure that out, we represented the topic obtained with the overlapping community detection using a graph representation. Table 4 shows the clustering metrics (Equations from (4) to (10)) topics for the positive reviews using the subnetwork for the service "swimming-pool". Fig. 4 shows the relevant subnetwork and the detected communities. The topics shown in Fig. 5 are related to the rooms, the hotel services, the swimming-pool, and a cluster of words in topic 3 are strongly related to positive feedback provided by satisfied users.

Table 4 provides a comprehensive overview of key metrics for topics identified by Overlapping Community Detection, one of the methods used to analyze customer reviews from Booking.com. These topics are associated with various services like 'wifi', 'swimming pool', 'breakfast', and 'transportation' and sentiments, either 'positive' or 'negative'. Each metric in the table provides a perspective on the structure and quality of the identified topics. The first two columns describe the sub-network obtained for the specific service and sentiment. The number of nodes tells us how many unique words each sub-network consists of. Edges show the number of connections between these words, representing how interconnected a sub-network is. Coverage gives us an idea of the density of a topic by showing the proportion of all possible connections that are actually present. Table 4 also presents the Conductance, which measures how self-contained a community is; lower values are desirable here as they imply fewer connections to other communities. Density, similar to Coverage, provides insight into how interconnected a topic is by showing the ratio of actual connections to all possible ones. Lastly, the Average Clustering Coefficient reveals the degree of clustering in a graph, i.e., how likely it is that words related to the same topic connect to each other. Each service and sentiment combination is analyzed individually, allowing us to compare the quality and structure of the topics identified. For example, for the service 'wifi' with positive sentiment, the topic consists of 217 unique words, interconnected through 4803 edges, with a coverage of 0.6648, a conductance of 0.4045, a density of 0.6209, and an average clustering coefficient of 0.0104.

4.2. Topic representation with semantic networks

We proceed with the inspection of topics using a visual representation of topics using the community detected. Each plot uses color to represent the topics and the nodes and edges size to represent the importance of the node in the topics' sub-network and the semantic similarity between nodes, respectively.

Figs. 4 and 5 present the network representation of topics pertaining to the service 'breakfast' by positive and negative sentiments. For the positive sentiment, seven communities are detected. The keywords suggest that positive reviews often refer to the overall experience at the hotel such as 'staff', 'structure', 'excellent', 'room', and 'hotel'. Specific mentions of 'breakfast' also appear, indicating direct positive feedback about the service, such as 'abbondante' (abundant, plentiful) with orange color. The green topic community suggests positive feedback with good food with keywords such as 'ricotta' (ricotta cheese), 'macedonia' (fruit salad) and 'buonissima' (very good). Metrics show a high

Table 2
Comparison of top-5 keywords per topic considering **wifi** and **pool** services and the positive and negative reviews. We compared the topic detection using the semantic network and Louvain's and Overlapping community detention methods. The last column is obtained using the LDA topic modeling method as a comparison with a traditional topic modeling method.

Service	Sentiment	Louvain's Community		Overlapping Community		LDA Topic	
		Topic	**Keywords**	**Topic**	**Keywords**	**Topic**	**Keywords**
wifi	positive	0	comfortable, breakfast, structure, furnishings, efficient	0	bathroom, bedroom, breakfast, fridge, rooms	0	broad, euro, free, power, present
		1	welcoming, spotless, convenient, clean, comfort	1	staff, breakfast, excellent, structure, swimming pool	1	room, wifi, breakfast, structure, bathroom
		2	spacious, bathroom, room, comfortable, very convenient	2	breakfast, excellent, restaurant, location, room	2	wifi, ottimo, camera, colazione, buono
	negative	0	breakfast, room, furniture, cleaning, bathroom	0	painful, generation, was worth, miserable, receives	0	wifi, room, breakfast, work, room
				1	hotel, room, structure, parking, breakfast		
				2	breakfast, room, structure, bathroom, hotel	1	wifi, room, signal, structure, breakfast
		1	conditioned, deficient, insufficient, improvable, problem	3	room, breakfast, bathroom, room, why		
				4	signal, internet, connection, reception, service		
		2	personal, improve, internet, present, non-existent	5	breakfast, restaurant, expensive, bad, crowded, dry	2	wifi, room, room, bathroom, work
				6	bathroom, breakfast, room, structure, room		
pool	positive	0	welcoming, comfortable, relaxing, swimming pool, spacious	0	breakfast, room, staff, structure, excellent, room	0	pool, excellent, breakfast, room, price
		1	breakfast, structure, staff, excellent, location	1	structure, excellent, sea, breakfast, hotel, location	1	swimming pool, breakfast, location, beautiful, excellent
				2	pool, dopp, check, improve it, operating, complain		
		2	comfortable, confortable, cozy, spacious, rooms	3	breakfast, excellent, structure, staff, hotel, location	2	swimming pool, structure, beach, small, hotel
				4	pool, happiness		
		3	truly, highly recommended, pleasant, beautiful, spotlessly clean	5	breakfast, excellent, parking, hotel, sea, beach	3	swimming pool, excellent, beautiful, staff, structure
				6	breakfast, staff, staff, structure, excellent		
	negative	0	really, friendly, problem, little, people	0	swimming pool, room, structure, breakfast, bathroom, room	0	pool, room, water, bathroom, area
		1	breakfast, structure, staff, improve, cleaning	1	swimming pool, breakfast, hotel, service, above all, the beach	1	swimming pool, room, breakfast, structure, service
		2	internal, water, external, insufficient, height	2	pool, render, slip, balloon, lap, beam, hang around		
				3	breakfast, room, structure, room, hotel, bathroom, rooms	2	pool, room, structure, breakfast, service
		3	rooms, bathrooms, furnishings, small, dirty	4	swimming pool, room, hotel, service, breakfast, structure		
				5	pool, room, bathroom, room, bed, cleaning, old	3	swimming pool, structure, room, price, restaurant
		4	room, pool, bathroom, bedroom, shower	6	pool, morning, evening, late, afternoon, morning, water		

coverage (0.7525) and density (0.1735), suggesting a well-defined and dense network of words. The conductance (0.4993) and clustering coefficient (0.0007) suggest that communities are fairly self-contained with a low degree of clustering. In contrast, for negative sentiment, nine communities are identified. Keywords like 'colazione' (breakfast), 'mattina' (morning), 'qualit' (quality), and 'scelta' (choice) appear, suggesting that negative feedback might be related to the quality, with keywords such as 'insufficiente' (insufficient), choice, with keywords such as 'categoria' (category), or timing of breakfast. Other keywords relate to other aspects of the hotel stay such as 'camera' (room), 'struttura' (structure), and 'bagno' (bathroom). The metrics show slightly higher coverage (0.7794) and lower density (0.1144) compared to the positive sentiment, suggesting a more spread-out network of words. The conductance (0.5053) and clustering coefficient (0.0001) indicate that the communities are also fairly self-contained with a very low degree of clustering.

Figs. 6 and 7 present the network representation of topics pertaining to the service 'wifi' by positive and negative sentiments. For the positive sentiment, three communities are detected. The keywords suggest that positive reviews often refer to various aspects of the hotel experience such as 'bagno' (bathroom), 'camera' (room), 'colazione' (breakfast), and 'personale' (staff). The metrics show a high coverage (0.6648) and density (0.6209), suggesting a well-defined and dense network of words. The conductance (0.4045) and clustering coefficient (0.0104) suggest that communities are fairly self-contained with a modest degree of clustering. In contrast, for negative sentiment, seven communities are identified. Some keywords like 'segnale' (signal), 'internet', 'connessione' (connection) and 'inutilizzabile' (unusable), suggest that negative feedback may be related to the quality or availability of the wifi service. Other keywords relate to different aspects of the hotel stay such as 'hotel', 'camera' (room), 'struttura' (structure), and 'colazione' (breakfast). The metrics show higher coverage (0.7681) and lower density (0.3188) compared to the positive sentiment, suggesting a more spread out network of words. The conductance (0.4927) and clustering coefficient (0.0016) indicate that the communities are also fairly self-contained with a low degree of clustering.

Table 3
Comparison of top-5 keywords per topic considering **breakfast** and **transportation** services and the positive and negative reviews. We compared the topic detection using the semantic network and Louvain's and Overlapping community detention methods. The last column is obtained using the LDA topic modeling method as a comparison with a traditional topic modeling method.

Service	Sentiment	Louvain's Community		Overlapping Community		LDA Topic	
breakfast	positive	0	comfortable, relaxing, spacious, great, restaurant	0	staff, personnel, structure, excellent, hotel, excellent	0	breakfast, quality, price, excellent, value
		1	plentiful, sweet, very good, very good, fresh	1	breakfast, kind, lady, welcomed, very kind	1	hotel, beach, structure, room, center
		2	welcoming, truly, highly recommended, pleasant, perfect	2	rooms, staff, excellent, staff, pool, location		
		3	comfort, kindness, relaxation, relaxation, tranquility	3	breakfast, room, structure, hotel, excellent	2	breakfast, sea, pool, view, beautiful
		4	comfortable, welcoming, spacious, well done, rooms	4	breakfast, excellent, staff, staff, room, good		
		5	breakfast, staff, structure, definitely, location	5	breakfast, center, structure, beaches, sea	3	breakfast, great, location, room, staff
		6	tidy, cared for, renovated, kept, organized, supplied	6	room, structure, staff, excellent, hotel		
	negative	0	really, comfortable, really, above all, annoying	0	breakfast, morning, evening, night, morning, time, late	0	breakfast, room, room, bathroom, structure
				1	room, hotel, structure, room, because, night		
		1	room, swimming pool, restaurant, bathroom, bedroom	2	breakfast, bacon, scrambled, squandered, ignoble, plastered	1	room, breakfast, bathroom, bed, room
		2	breakfast, fruit, sweets, eat, bread	3	breakfast, quality, products, fruit, choice, croissants	2	breakfast, poor, coffee, service, time
				4	breakfast, toilet, shower, room, room, bed, air, smell		
		3	staff, structure, absolutely, definitely, above all	5	breakfast, hotel, room, structure, restaurant, parking		
				6	room, room, bathroom, structure, rooms, hotel	3	breakfast, room, quality, star, structure
		4	improve,see,especially,small,review	7	room, structure, hotel, room, stars, bathroom, restaurant		
transportation	positive	0	clean, comfortable, excellent, beach	0	beaches, cars, alghero, find	0	transport, location, staff, excellent, center
						1	transport, beach, excellent, hotel, structure
						2	transport, room, excellent, beach, structure
						3	transport, room, swimming pool, beach, excellent
		1	helpful, personal, services, service, reach	1	interior, services, excellent, staff, available	4	location, transport, excellent, hotel, breakfast
						5	room, transportation, good, staff, kind
						6	transport, sea, structure, downtown, breakfast
						7	hotel, excellent, breakfast, restaurant, present
	negative	0	interior, service, personal, available, center	0	airport, center, km, shuttle, motorbike	0	transport, baggage, hotel, reveal, timetable
				1	staff, interior, service, above all, elevator, available, cost	1	room, transport, restaurant, height, star
				2	presence, above all, internal, date, need, education	2	card, ask, restaurant, price, beach, choose
				3	uncomfortable, stroller, possession, menu	3	transport, room, luggage, interior, elevator
						4	transport, pay, center, service, structure
		1	soggiorno, spiaggia, camera	4	available, personal, service, services, cost, pay	5	close, check, restaurant, room, arrival
						6	transport, beach, service, hotel, guest
				5	center, timetables, shuttle, beach, km, airport	7	unit, housing, transport, star, working
						8	transport, car, center, stay, partner, hotel

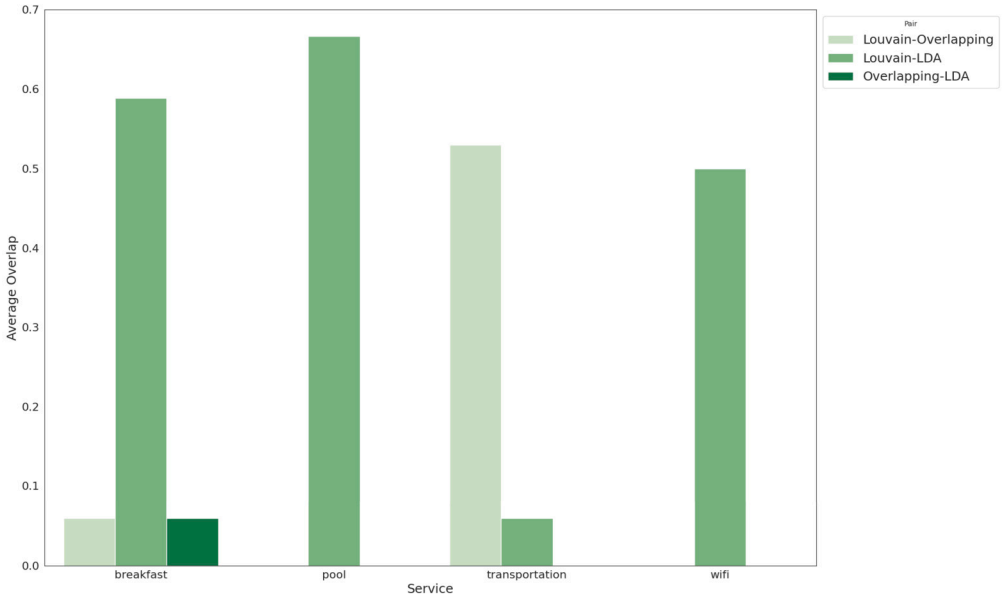

Fig. 2. Pairwise comparison of topics' words represented in Table 2 and 3. The average overlap is calculated as the average number of common words per topic for each type of service.

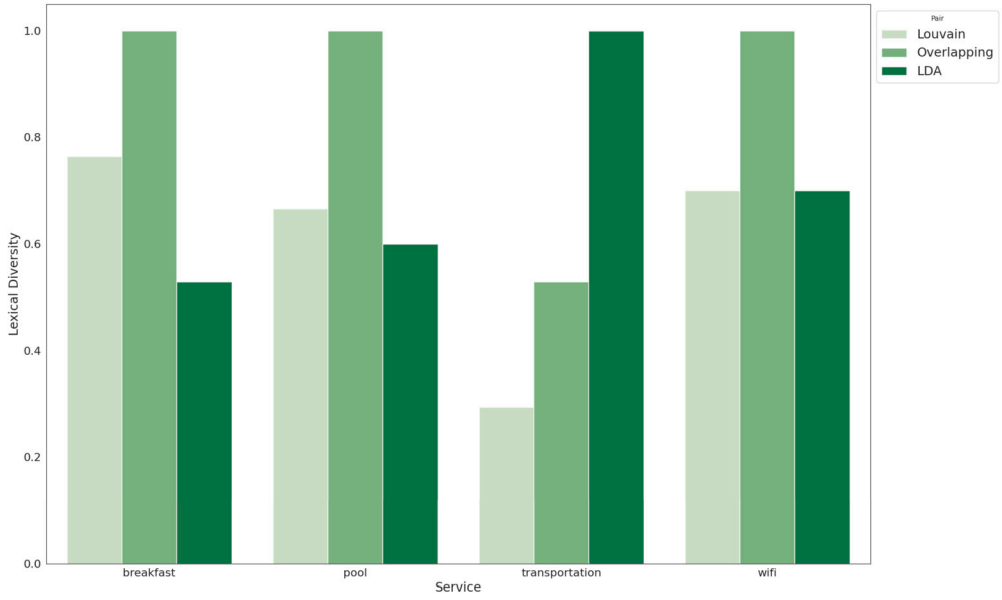

Fig. 3. Lexical diversity of topics represented in Table 2 and 3. The lexical diversity is the ration between the number of unique words and the total number of words considering all topics related to a type of service.

Table 4
Overlapping community detection metrics.

Service	Sentiment	Nodes	Edges	Coverage	Conductance	Density	Avg. Clustering Coeff.
wifi	positive	217	4803	0.6648	0.4045	0.6209	0.0104
wifi	negative	395	6847	0.7681	0.4927	0.3188	0.0016
swimming pool	positive	1128	37745	0.7525	0.4993	0.1735	0.0007
swimming pool	negative	786	15252	0.7794	0.5053	0.1144	0.0001
breakfast	positive	2235	61140	0.6903	0.4933	0.3171	0.0021
breakfast	negative	2020	40482	0.7827	0.4885	0.2352	0.0008
transportation	positive	40	306	0.4542	0.4338	0.7241	0.0123
transportation	negative	44	186	0.4731	0.6233	0.4953	0.0071

Figs. 8 and 9 present the network representation of topics pertaining to the service 'pool' by positive and negative sentiments. For the positive sentiment, nine communities are detected. The keywords suggest that positive reviews often refer to various aspects of the hotel experience such as 'piscina' (pool), 'colazione' (breakfast), 'camera' (room), 'personale' (staff), and 'struttura' (structure). The metrics show a high coverage (0.6903) and density (0.3171), suggesting a well-defined and dense network of words. The conductance (0.4933) and clustering co-

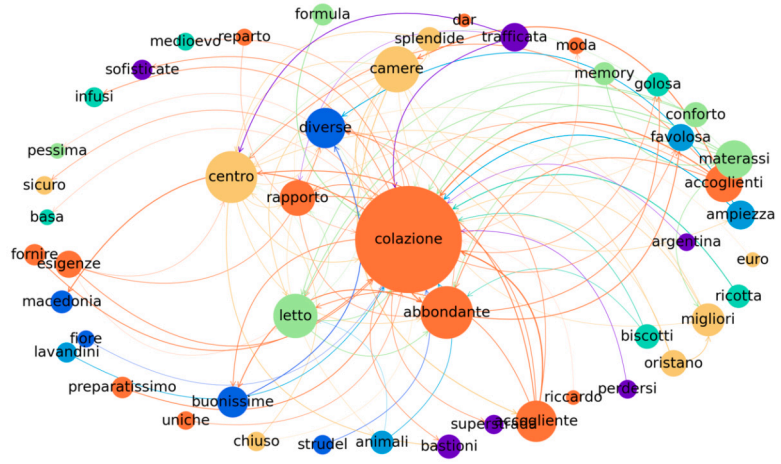

Fig. 4. Topic representation for keyword **breakfast** in positive reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

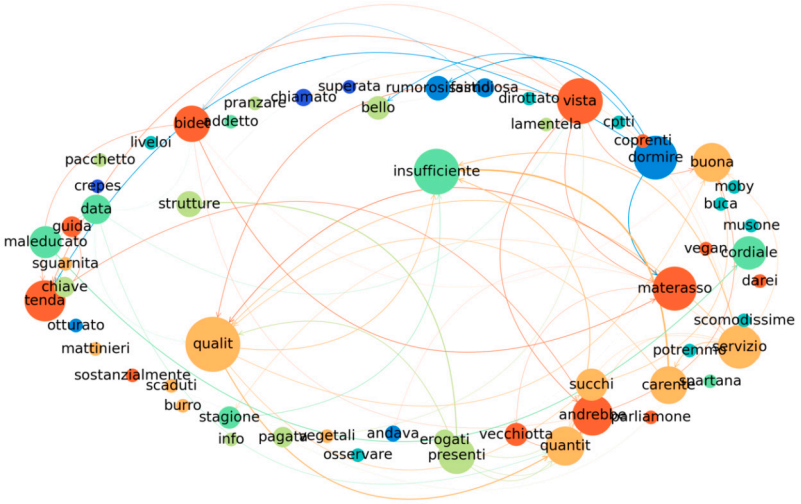

Fig. 5. Topic representation for keyword **breakfast** in negative reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

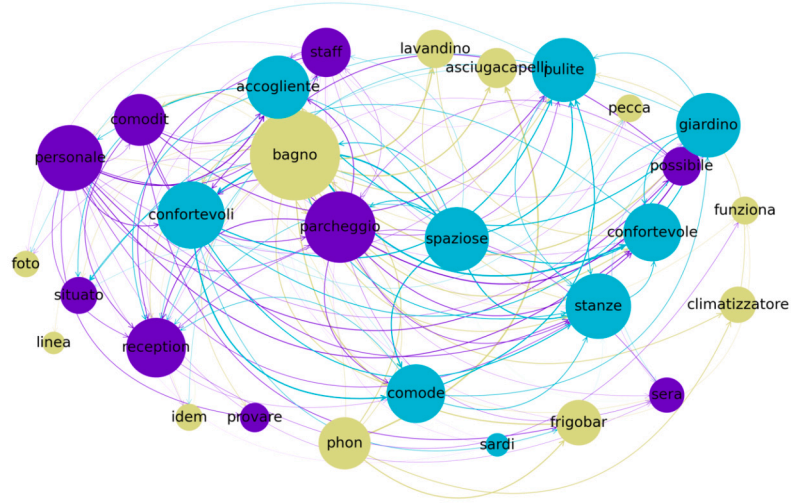

Fig. 6. Topic representation for keyword **wifi** in positive reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

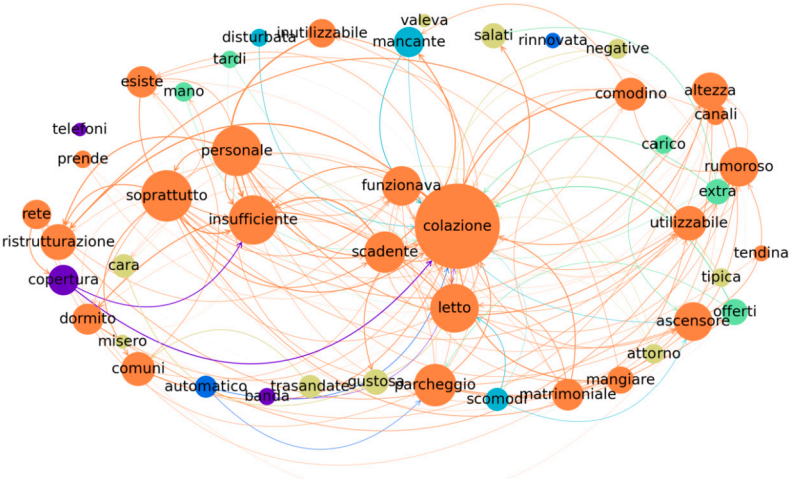

Fig. 7. Topic representation for keyword **wifi** in negative reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

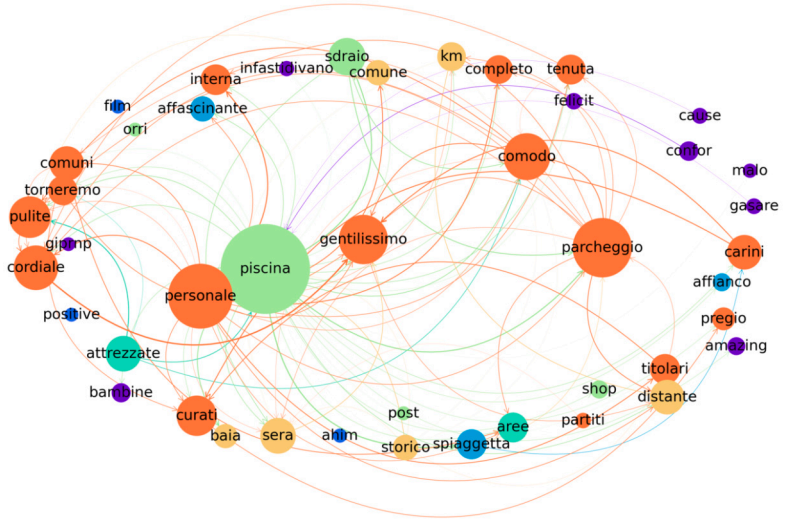

Fig. 8. Topic representation for keyword **piscina** in positive reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic they are affiliated with.

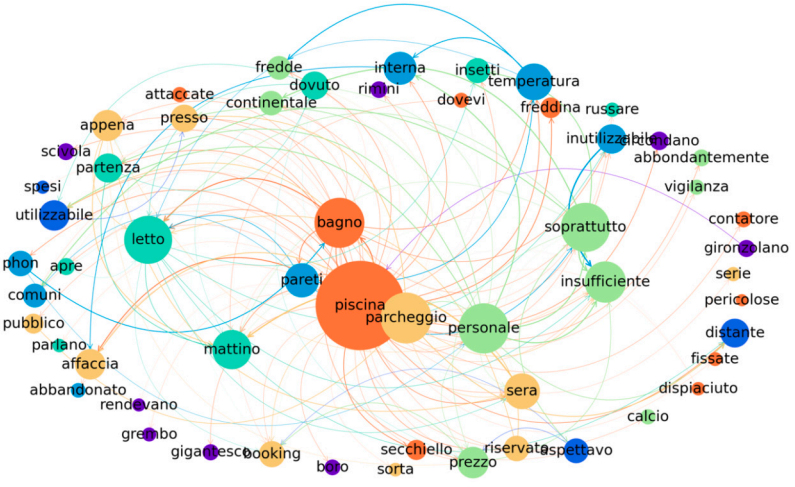

Fig. 9. Topic representation for keyword **pool** in negative reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

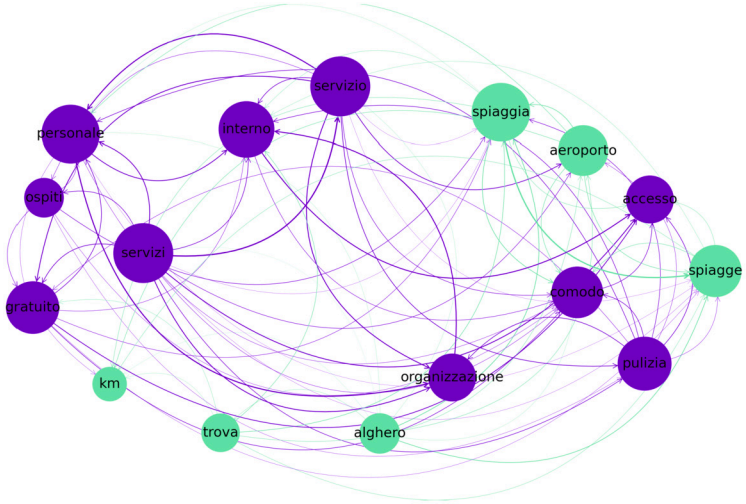

Fig. 10. Topic representation for keyword **transportation** in positive reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

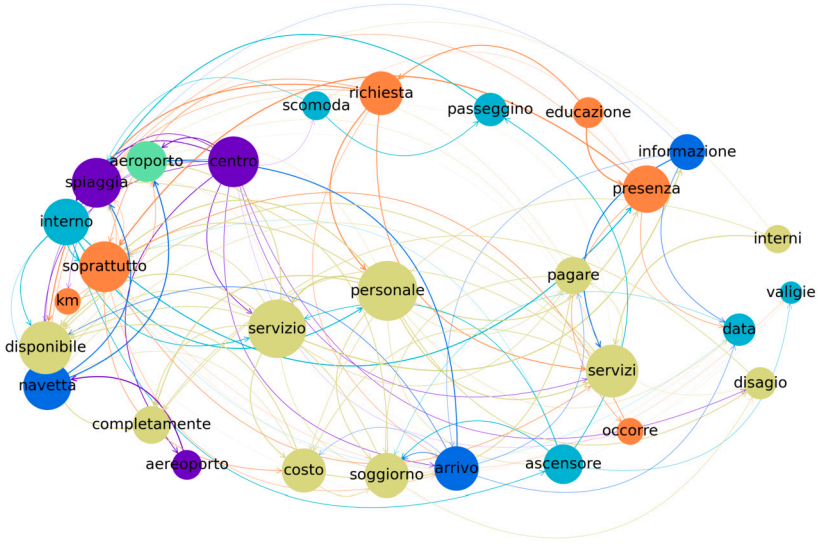

Fig. 11. Topic representation for keyword **transportation** in negative reviews. Each node corresponds to a key term within the topic, with the size and label font size of the node being indicative of its relative significance within its respective cluster. Edges delineate the conceptual linkage between terms, predicated on semantic proximity. The chromatic distinction of nodes demarcates the thematic cluster they are affiliated with.

efficient (0.0022) suggest that communities are fairly self-contained with a modest degree of clustering. In contrast, for negative sentiment, eight communities are identified. Some keywords like 'piscina' (pool), 'colazione' (breakfast), and 'camera' (room) suggest that negative feedback may be related to various aspects of the hotel experience. Other keywords like 'mattina' (morning), 'sera' (evening), 'tardi' (late), and 'pomeriggio' (afternoon) indicate possible issues with the timing or availability of the pool service. The metrics show higher coverage (0.7827) and lower density (0.2352) compared to the positive sentiment, suggesting a more spread out network of words. The conductance (0.4885) and clustering coefficient (0.0008) indicates that the communities are also fairly self-contained with a low degree of clustering.

Figs. 10 and 11 present the network representation of topics pertaining to the service 'transoportation' by positive and negative sentiments. For the positive sentiment, three communities are detected. The keywords suggest that positive reviews often refer to various aspects such as 'spiagge' (beaches), 'auto' (car), 'alghero' (Alghero, a city in Sardinia, Italy), and 'trova' (find). The metrics show a relatively low coverage (0.4542) and high density (0.7242), suggesting a well-defined and dense network of words. The conductance (0.4338) and clustering coefficient (0.0123) suggest that communities are fairly self-contained with a modest degree of clustering. For the negative sentiment, eight communities are identified. Some keywords like 'aeroporto' (airport), 'centro' (center), 'km' (kilometers), 'navetta' (shuttle), and 'moto' (motorbike) suggest that negative feedback may be related to various aspects of transportation service. Other keywords like 'valigie' (suitcases) and 'valige' (luggage) indicate possible issues with luggage handling or storage. The metrics show slightly higher coverage (0.4731) but lower density (0.4953) compared to the positive sentiment, suggesting a more spread out network of words. The conductance (0.6233) and clustering coefficient (0.007) indicate that the communities are also fairly self-contained with a low degree of clustering.

Overall, the analysis shows that both positive and negative sentiments about services are not solely focused on the service itself but extend to the overall hotel experience. The differences in keywords and metrics between positive and negative sentiments provide valuable insights into the aspects of the service that customers appreciate or find lacking. SMARTS offers several benefits for tourism decision-

makers and stakeholders. Here, we highlight seven potential areas of benefits of the proposed methodology.

Enhanced Interpretability of Customer Feedback, SMARTS employs an elaborated approach to topic detection that combines semantic and sentiment analysis of customer reviews. This enables decision-makers to better understand the specific aspects of their services that customers are praising or criticizing. By identifying topics related to positive and negative sentiment, decision-makers can pinpoint areas of strength and areas that need improvement.

Focused Service Improvement, with the identified topics, decision-makers can focus their efforts on improving specific services that are receiving negative feedback. For instance, if the "wifi" service is consistently receiving negative sentiment due to issues with signal quality or connectivity, stakeholders can allocate resources to address these concerns and enhance the overall customer experience.

Tailored Marketing and Communication, understanding the sentiments and topics in customer reviews allows stakeholders to tailor their marketing and communication strategies. Positive aspects highlighted by customers can be emphasized in promotional materials, while areas of concern can be proactively addressed in marketing campaigns to show customers that their feedback is being taken seriously.

Strategic Decision Making, the insights obtained from SMARTS can guide strategic decision making within the tourism industry. By identifying which services are most positively or negatively received, businesses can allocate resources more effectively, make informed operational changes, and develop strategies to differentiate themselves from competitors.

Continuous Monitoring and Feedback Loop, SMARTS can be used as part of a continuous feedback loop to monitor the impact of service improvements. Decision-makers can track changes in sentiment and topics over time to assess the effectiveness of their interventions and adapt their strategies accordingly.

Customer-Centric Approach, the methodology is focused on customer reviews and sentiments to empower decision-makers to take a customer-centric approach. By aligning services with customer preferences and addressing their concerns, businesses can foster greater customer loyalty and satisfaction.

Competitive Advantage, implementing SMARTS can provide a competitive advantage by allowing businesses to better understand customer sentiments and preferences compared to competitors. This can lead to more tailored and relevant offerings, ultimately attracting and retaining customers.

5. Conclusions

In this study, we proposed a novel approach to topic detection called SeMi-supervised clustering for Assessment of Reviews using Topic and Sentiment (SMARTS) applied to Topic-Community Representation with Semantic Networks, which exploits an ensemble of semantic network and sentiment analysis for semi-supervised clustering of customer reviews. Between all the possible Sentiment Analysis algorithms, we considered a new version of the NB classifier, called Tb-NB, that utilizes a data-driven decision rule to assign a new case to the most likely between two alternative classes, which is based on a threshold whose value is estimated from the training data [23]. In [23], it is reported that Tb-NB effectively discriminates positive reviews from negative ones and, at the same time, allows us to quantify the (positive or negative) impact of a specific word within a review. Moreover, the information deriving from Tb-NB can be used to support decision makers as the Tb-NB output can be used further in post-hoc analyses to evaluate different facets of customer satisfaction [24]. Last but not least, it has been shown the Tb-NB is preferable to other methods used in Sentiment Analysis in terms of classification accuracy, resistance to noise, and computational efficiency.

Our methodology leverages the construction of a semantic network of words based on word embeddings to identify the semantic similarity between different words used in the reviews, which is then used to identify topics present in the reviews grouped by positive and negative sentiment and related to particular services or products. Our findings show that our approach is effective in detecting interpretable topics in a dataset of tourism reviews extracted from Booking.com. Our novel methodology could provide valuable insights into the sentiment of customers towards products and services and could support decision-making processes. In fact, the SMARTS methodology presents a specialized example for tourism decision-makers and stakeholders, offering a more nuanced and accurate understanding of customer feedback. By integrating semantic and sentiment analysis through word embeddings and graph-based techniques, SMARTS identifies topics associated with positive and negative sentiments in customer reviews. This approach allows decision-makers to pinpoint specific service aspects for improvement, tailor marketing strategies, and make informed strategic decisions, fostering customer loyalty and competitive advantage. In comparison to traditional methods, SMARTS has a domain-specific focus, advanced analysis techniques, and seamless integration of sentiment and topic insights, providing tourism industry professionals with a powerful tool for enhancing service quality and customer satisfaction. Its semantic analysis, sentiment-topic integration, and graph-based representation provide a unique approach that enhances decision-making, service improvement, and customer satisfaction within the tourism industry.

Future works will tackle the problem of the generation of automatic topic labeling and automatic topic number selection using the information of the semantic network of words as a driver, as well as the implementation of the, recently proposed, Iterative Threshold-based Naïve Bayes classifier [25].

CRediT authorship contribution statement

Marco Ortu: Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Maurizio Romano:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Andrea Carta:** Writing – review & editing, Writing – original draft, Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

The work was partially funded by the project Partenariato Esteso "GRINS - Growing Resilient, INclusive and Sustainable", tematica "9. Economic and financial sustainability of systems and territories", CUP: PE00000018.

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – Next Generation EU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the MUR.

References

[1] A. Bavelas, Communication patterns in task-oriented groups, The Journal of the Acoustical Society of America 22 (6) (1950) 725–730.

[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (Jan 2003) 993–1022.
[3] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10) (2008) P10008.
[4] Z. Ding, X. Zhang, D. Sun, B. Luo, Overlapping community detection based on network decomposition, Scientific Reports 6 (1) (2016) 24115.
[5] Z. Drus, H. Khalid, Sentiment analysis in social media and its application: systematic literature review, Procedia Computer Science 161 (2019) 707–714.
[6] M. Ebrahimi, A.H. Yazdavar, A. Sheth, Challenges of sentiment analysis for dynamic events, IEEE Intelligent Systems 32 (5) (2017) 70–75.
[7] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry (1977) 35–41.
[8] L. Frigau, M. Romano, M. Ortu, G. Contu, Semi-supervised sentiment clustering on natural language texts, Statistical Methods & Applications (2023) 1–19.
[9] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint, arXiv:2203.05794, 2022.
[10] S.K. Gupta, D.P. Singh, J. Choudhary, A review of clique-based overlapping community detection algorithms, Knowledge and Information Systems 64 (8) (Aug 2022) 2023–2058.
[11] D. Hevey, Network analysis: a brief overview and tutorial, Health Psychology and Behavioral Medicine 6 (1) (2018) 301–328.
[12] M. Honnibal, I. Montani, Spacy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear 7 (1) (2017) 411–420.
[13] M. Huang, Y. Rao, Y. Liu, H. Xie, F.L. Wang, Siamese network-based supervised topic modeling, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4652–4662.
[14] M.A. Javed, M.S. Younis, S. Latif, J. Qadir, A. Baig, Community detection in networks: a multidisciplinary review, Journal of Network and Computer Applications 108 (2018) 87–111.
[15] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (lda) and topic modeling: models, applications, a survey, Multimedia Tools and Applications 78 (2019) 15169–15211.
[16] S. Jung, A. Segev, Analyzing the generalizability of the network-based topic emergence identification method, Semantic Web 13 (3) (2022) 423–439.
[17] P. Kherwa, P. Bansal, Topic modeling: a comprehensive review, EAI Endorsed Transactions on Scalable Information Systems 7 (24) (2019).
[18] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint, arXiv:1301.3781, 2013.
[19] T. Narayanan, M. Gersten, S. Subramaniam, A. Grama, Modularity detection in protein-protein interaction networks, BMC Research Notes 4 (1) (2011) 1–6.
[20] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69 (2) (2004) 026113.
[21] M. Ortu, L. Frigau, G. Contu, Topic based quality indexes assessment through sentiment, Computational Statistics (2022) 1–23.
[22] A. Ponomarenko, L. Pitsoulis, M. Shamshetdinov, Overlapping community detection in networks based on link partitioning and partitioning around medoids, PLoS ONE 16 (8) (08 2021) 1–43.
[23] M. Romano, G. Contu, F. Mola, C. Conversano, Threshold-based naïve Bayes classifier, Advances in Data Analysis and Classification (Mar 2023).
[24] M. Romano, G. Zammarchi, C. Conversano, Threshold-based naïve Bayes classifier: customer satisfaction evaluation, in: Short Papers IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment, 2022, pp. 90–94.
[25] M. Romano, G. Zammarchi, C. Conversano, Iterative threshold-based naïve Bayes classifier. Statistical Methods & Applications, 2023.
[26] S.M. Saif, M.E. Samie, A. Hamzeh, A subgraphs-density based overlapping community detection algorithm for large-scale complex networks, Computing 105 (1) (Jan 2023) 151–185.
[27] O. Shchur, S. Günnemann, Overlapping community detection with graph neural networks, in: Deep Learning on Graphs Workshop, KDD, 2019.
[28] C. Shi, Y. Cai, D. Fu, Y. Dong, B. Wu, A link clustering based overlapping community detection algorithm, Data & Knowledge Engineering 87 (2013) 394–404.
[29] F. Steuber, S. Schneider, M. Schoenfeld, Embedding semantic anchors to guide topic models on short text corpora, Big Data Research 27 (2022) 100293.
[30] F. Tavazoee, C. Conversano, F. Mola, Recurrent random forest for the assessment of popularity in social media, Knowledge and Information Systems 62 (2020) 1847–1879.
[31] V.d.F. Vieira, C.R. Xavier, A.G. Evsukoff, A comparative study of overlapping community detection methods from the perspective of the structural properties, Applied Network Science 5 (1) (Aug 2020) 51.
[32] P. Wang, Y. Huang, F. Tang, H. Liu, Y. Lu, Overlapping community detection based on node importance and adjacency information, Security and Communication Networks 2021 (Dec 2021) 8690662.
[33] S. Wasserman, K. Faust, Social network analysis: Methods and applications, 1994.
[34] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks, ACM Computing Surveys 45 (4) (aug 2013) 1–35.
[35] K. Xu, X. Lu, Y.-f. Li, T. Wu, G. Qi, N. Ye, D. Wang, Z. Zhou, Neural topic modeling with deep mutual information estimation, Big Data Research 30 (2022) 100344.
[36] S. Yuan, H. Zeng, Z. Zuo, C. Wang, Overlapping community detection on complex networks with graph convolutional networks, Computer Communications 199 (2023) 62–71.