SMARTS: SeMi-supervised clustering for Assessment of Reviews using Topic and Sentiment

Marco Ortu and Maurizio Romano and Andrea Carta

Abstract This paper proposes a novel approach to topic detection aimed at improving the semi-supervised clustering of customer reviews in the context of tourism services. The proposed methodology, named SeMi-supervised clustering for Assessment of Reviews using Topic detectio and Sentiment, 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 a specific service. Clusters of words obtained from the semantic network 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 reviews 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.

Key words: Semi-supervised Clustering, Sentiment Analysis, Topic Modeling, Natural Language Processing

---

Marco Ortu
Università di Cagliari, Dept. of Business and Economic Sciences e-mail: marco.ortu@unica.it

Maurizio Romano
Università di Cagliari, Dept. of Business and Economic Sciences e-mail: romano.maurizio@unica.it

Andrea Carta
Università di Cagliari, Dept. of Business and Economic Sciences e-mail: andrea.carta88@unica.it

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 (Jung & Segev 2022, Huang et al. 2018). 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 (Mikolov et al. 2013), to construct a topic model based on a semantic network of words.

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 quality toward 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 analyzing customer feedback to gain insights into the sentiment and opinions of their customers (Ortu et al. 2022). Topic modeling is one of the most popular methods to analyze customer feedback 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 feedbacks, there are still some challenges in applying it to real-world datasets. For instance, the limited availability of labeled data sets makes it difficult to train supervised models with a sufficient quality for industrial applications. For this reason, semi-supervised methods have been proposed to address this challenge, which leverage both labeled and unlabeled data to improve the accuracy of clustering. In this paper, we propose a novel methodology for topic detection designed for improving semi-supervised clustering of users' reviews with an application for tourism reviews data. The proposed methodology, called SeMi-supervised clustering for Assessment of Reviews using Topic detection and Sentiment (SMARTS), leverage 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 interpretation of quality assessment expressed by customers in online reviews toward 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 details 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

2.1 Topic modeling

Topic Modeling is a method for generating low-dimensional, multi-faceted summaries of documents or other discrete data, representing the process of analysing the words in a text to discover hidden themes and pattern. It is an unsupervised learning method because it does not require the labelling of documents (Blei 2012). Different approaches to topic modelling can be used, among them, Latent Dirichlet Allocation (Blei et al. 2003) is one of the most popular because it allows to work with large collections of text documents, and it is widely used in several contests such as conference program organization (Frigau et al. 2022). Formally, LDA is a generative probabilistic model for collections of text corpora, and it comprises a three-level hierarchical Bayesian model, in which each item of a specific collection is modelled as a finite combination over an underlying set of topics (Blei et al. 2003). Network-based techniques have been widely utilized in recent years for textual data analysis, even though the analysis of term-term matrices has a long-established history in other scientific domains (Popping 2003). One of the primary advantages of this approach is the ability to preserve the contextual usage of different terms. The original terms remain intact in the analyses, making the results easier to interpret. By selecting a threshold on the co-occurrence value, it becomes possible to highlight highly connected structures that represent the core of the document collection. However, graphs, being non-metric representations, do not account for term associations as distances. Regarding community detection, a noteworthy aspect is that prior knowledge of the number of topics present in the document collection and the corresponding parameter setting is not required. This enables a completely automatic extraction process (Misuraca et al. 2020).

Topic modeling is a powerful methodology that has transformed the field of Natural Language process. Its ability to automatically identify latent topics in a

corpus of documents has made it an essential tool for text analysis in various fields. As the amount of digital data generated every day increases, topic modeling is expected to play an increasingly important role in analysing and understanding the content of large text collections. Topic modeling has many applications in fields such as social media analysis, information retrieval, and market research. For instance, it can help identify topics or sentiments expressed in tweets or comments, improve search accuracy by identifying relevant themes or topics in documents, and provide insights into consumer preferences and trends. However, ensuring the quality and coherence of the identified topics is a significant challenge in topic modeling. Researchers can address this challenge by using appropriate coherence measures and carefully selecting the number of topics extracted to ensure the identified topics are meaningful and interpretable (Misuraca et al. 2020).

2.2 Modularity clustering

Modularity clustering (Newman & Girvan 2004) is a widely-used unsupervised machine learning method for detecting highly interconnected communities or groups of nodes within an intricate network. This technique leverages the modularity measure, which assesses the density of connections among nodes within the same group relative to other groups. The modularity measure ranges between 0 and 1, where 0 indicates a random structure, and 1 indicates a strong community structure (Newman 2003). Empirically, it has been observed that modularity values usually fall within the subinterval $[0.3, 0.7]$ (Newman & Girvan 2004).

To maximize the modularity measure, modularity clustering algorithms iteratively divide the network into smaller subgroups. The process starts with a single group containing all nodes in the network and splits it into two subgroups based on the nodes' connectivity patterns. The algorithm continues to divide the subgroups into smaller subgroups until a stopping criterion is met.

Several algorithms are available for modularity clustering, including the Louvain method, spectral clustering, and greedy algorithm (for an overview, see Javed et al. (2018)). Among them, the Louvain method (Blondel et al. 2008), is particularly popular due to its efficiency and scalability. It uses a greedy algorithm that maximizes the gain for each step to optimize the modularity score at each level of the hierarchy, thus enabling users to consider the communities in different resolutions (Frigau et al. 2021).

Modularity clustering has numerous applications in social network analysis, biology, finance, and web mining. For instance, Groza et al. (2021) developed an automated computational drug repurposing process based on drug–gene interaction data using modularity class network clustering. Moreover, Park & Kim (2021) used modularity clustering to partition the networks that were built based on the extracted attributes of customer reviews, and then the resulting clusters were evaluated by topic frequency to characterize customer segments.

One major challenge in modularity clustering is the resolution limit problem, which refers to the difficulty of detecting small communities nested within larger communities (Fortunato & Barthelemy 2007). This problem occurs because modularity tends to favor larger communities over smaller ones.

2.3 Sentiment Analysis

Sentiment analysis is a subfield of Natural Language Processing that focuses on analyzing a given text to find the emotional tone. This type of analysis can be applied within different levels of granularity, for instance, considering an entire document or the single words that compose it as an entity. Commons implementations of Sentiment Analysis are based on Machine learning techniques (usually within a supervised learning context), training models with a given natural language text, and classifying it as a positive, neutral, or negative sentiment.There are many applications of Sentiment Analysis (i.e. customer satisfaction, social media monitoring) with many businesses, managerial, political, and academic implications (Tavazoee et al. 2020, Jain et al. 2021). Moreover, Sentiment Analysis has grown in popularity thanks to the increasing availability of the vast amounts of textual data produced by social media. In fact, researchers are still both producing novel methodologies whilst improving well-known ones (see, for instance, Ortu et al. (2022), Romano et al. (2023)). Nevertheless, considering that words can have different meanings within different contexts, there are still several challenges in the field. Detection of other language expressions (like irony or sarcasm), sentiment subjectivity, and more interesting topics that focus on interpretation based on background knowledge (cultural) of the reader.

3 Methodology

SMARTS methodology consists of four main phases: 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 details. We pre-processed the reviews' set $\mathcal{D}$ by removing stopwords, punctuations, and all non-alphabetic and non-relevant characters to obtain the word vector $\mathcal{W}_d$. Next, $\mathcal{W}_d$ is transformed using the Word Embeddings (Mikolov et al. 2013) using the SpaCy library (Honnibal & Montani 2017), which provides a pre-trained model for Italian language, to obtain the embeddings vector $\mathcal{W}_e$. The embeddings of words vector is able to capture the semantic relationships among words, which is used in the next phase to build the semantic network. The sentiment category is assigned to the $\mathcal{W}_e$ word embeddings using the $\mathcal{S}_e$ function in this phase. The Semantic Network is created using words as

Algorithm 1 SMARTS methodology algorithm.

Require $\mathcal{D}$: Documents (i.e. reviews) set of size N;
Require $\mathcal{S} : d \rightarrow s$; A function that assign a sentiment s to a review 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;
Step 1:
  Input: $\mathcal{D} = \{d_1, \ldots, d_N\}$
  Output: $\mathcal{W}_d = \{w_1, \ldots, w_n\}$
Step 2:
  Input: $\mathcal{W}_d = \{w_1, \ldots, w_n\}$
  Output: $\mathcal{W}_e = \{w_{e,1}, \ldots, w_{e,n}\}$
  Output: $\mathcal{S}_e = \mathcal{S}(\mathcal{W}_e)$
Step 3:
  for $s \in S$ **do**
    for $w_{e,i} \in \mathcal{W}_e$ **do**
      $e_i \leftarrow Sim(w_i, w_{i+1})$
      $\mathcal{N} \leftarrow \mathcal{N} \cup (w_{e,i}, w_{e,i+1}, e_i)$
    end for
  end for
Step 4:
  Input: w_s
  Output: $\mathcal{N}_s \subseteq \mathcal{N}$
  Output: $\phi_s : \mathcal{N}_s \rightarrow Rank(\Phi_s)$

nodes and the semantic similarity as weight denoted by e_i. Considering each review, a node of the network is created for each of its words. An edge of the network is created using the semantic similarity function, $Sim(w_i, w_{i+1})$, computed as a weight for a word w_i and the following word w_{i+1} to obtain the tuple (w_i, w_{i+1}, e_i) which represents an edge of the network from node w_i to node w_{i+1} weighted by e_i. The semantic similarity is computed using an estimate of the cosine similarity using average word vectors, which is a number from 0 to 1. In the last phase, two separated 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 Louvain's Modularity (Blondel et al. 2008), 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 page rank (Brin & Page 1998) algorithm as the $Rank(\Phi_s)$ function, where Φ_s is the set of all nodes in ϕ_s.

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 inter-

pretable 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 concern 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 guest 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 to interpret users' assessment toward specific services by constructing a specific semantic network for positive and negative reviews.

Label Topic	Words	Topic Label
Topic 0	accogliente confortevole, rilassante, piscina, spaziosa	swimming-pool
Topic 1	colazione, struttura, personale, eccellente, posizione	hotel
Topic 2	confortevoli, confort, accoglienti, spaziose, camere	rooms
Topic 3	veramente, consigliatissimo, piacevole, bellissimo, pulitissimo	positive feedback

Table 1: Positive topics for the word service "swimming-pool".

Table 1 shows the obtained topics for the positive reviews using the subnetwork for the service "swimming-pool". Figure 1 shows the relevant subnetwork and the detected communities. The topics shown in Figure 1 are related to the rooms, the hotel services, the swimming-pool and a cluster of words in topic 3 are strongly related to positive feedbacks provided by satisfied users.

Label Topic	Words	Topic Label
Topic 0	veramente, accogliente, problema, piccolo, gente	swimming-pool service
Topic 1	colazione, struttura, personale, migliorare, pulizia	hotel services improvements
Topic 2	interno, acqua, esterno, insufficiente, altezza	swimming-pool cleaning
Topic 3	stanze, bagni, arredi, piccoli, sporchi	room cleaning
Topic 4	stanza, piscina, bagno, camera, doccia	distance from the swimming-pool

Table 2: Negative topics for the word service "swimming-pool".

Instead, in Table 2 are shows the obtained topics for the negative reviews using the subnetwork for the service "swimming-pool". Figure 2 shows the relevant subnetwork and the detected communities. The topics shown in Figure 2 are related to the rooms cleanings, the hotel services issues, the swimming-pool distance from the rooms and the swimming-pool service. These cluster of words in topics 2, 3 and 4

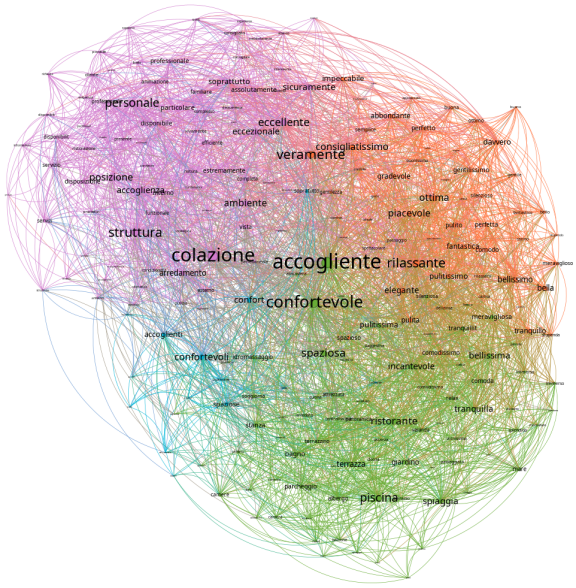

Fig. 1: Swimming pool service network for positive reviews.

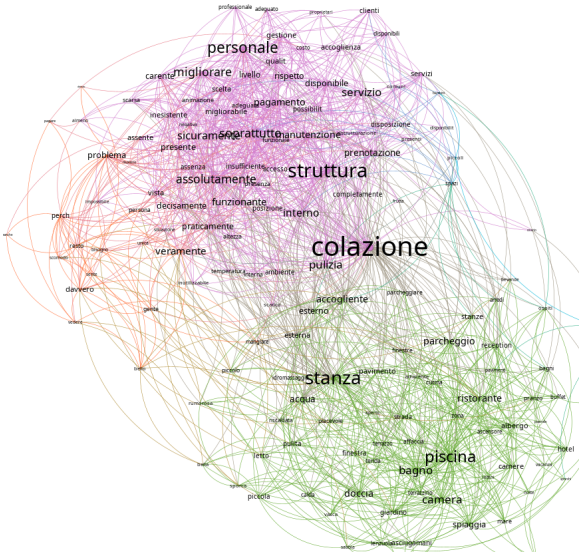

Fig. 2: Swimming pool service network for negative reviews.

are strongly related to negative feedbacks provided by unsatisfied users, and can be used to improve the overall quality of a service.

Label Topic	Words	Topic Label
Topic 0	confortevole, colazione, struttura, arredamento, efficiente	Hotel and services
Topic 1	accogliente, pulitissimo, comodo, pulito, comfort	Positiveness of services
Topic 2	spaziosa, bagno, stanza, confortevoli, comodissimo	Room

Table 3: Positive topics for the word service ”Wi-Fi”.

Table 3 shows the obtained topics for the positive reviews using the subnetwork for the service “Wi-Fi”. Figure 3 shows the relevant subnetwork and the detected communities. The topics shown in Figure 3 are related to the rooms, the hotel services, the Wi-Fi and a cluster of words in topic 3 are strongly related to positive feedbacks provided by satisfied users. These topics highlight the quality of services represented by positive reviews that contains the word ”Wi-FI”.

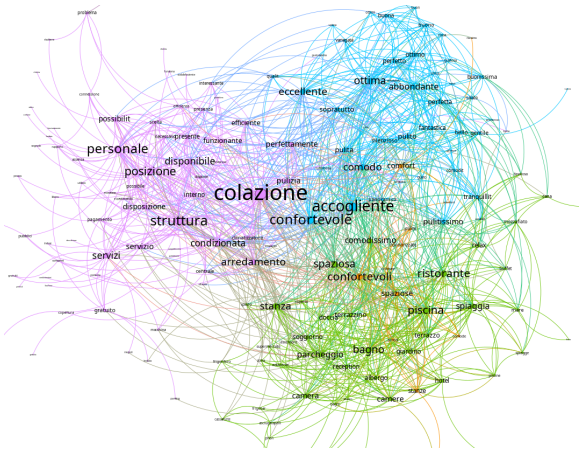

Fig. 3: Wi-Fi service network for positive reviews.

Table 4 shows the obtained topics for the negative reviews using the subnetwork for the service “Wi-Fi”. Figure 4 shows the relevant subnetwork and the detected communities. The topics shown in Figure 4 are related to the rooms, the hotel services in particular. Topic 2 and 3, highlight mainly the problems related to the services, specifically for the ”Wi-Fi”, such as the quality and extension of the Wi-Fi network, the areas where the Wi-Fi network is not present, and in general issues with the service that could be improved.

SMARTS has the potential to provide valuable insights into the sentiment of customers toward services’ assessment, which could be used by service providers and decision-makers to enhance their quality of services.

Label Topic	Words	Topic Label
Topic 0	colazione, stanza, arredamento, pulizia, bagno	Hotel and services
Topic 1	condizionata, carente, insufficiente, migliorabile, problema	Negativeness of services
Topic 2	personale, migliorare, internet, presente, inesistente	Wifi service quality

Table 4: Negative topics for the word service ”Wi-Fi”.

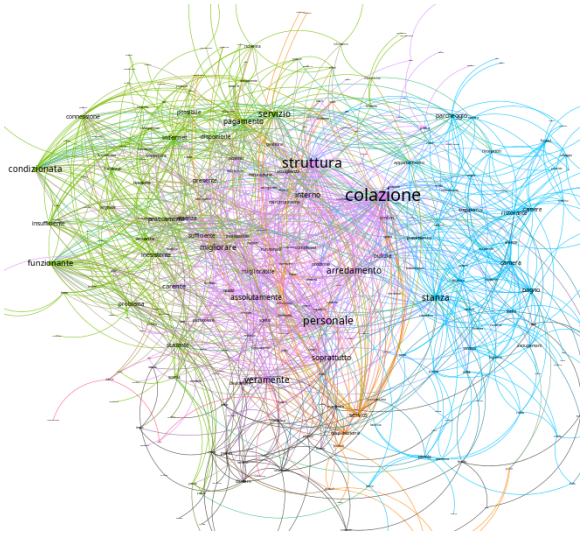

Fig. 4: Wi-Fi service network for negative reviews.

5 Conclusions

In this study, we proposed a novel approach to topic detection called SeMi-supervised clustering for Assessment of Reviews using Topic detection and Sentiment (SMARTS), which exploits an ensemble of semantic network and sentiment analysis for semi-supervised clustering of customer reviews. 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. Future works will tackle the problem of the generation of automatic topic labeling and automatic topic number selection using the information of semantic network of words as a driver.

References

Blei, D. M. (2012), 'Probabilistic topic models', *Communications of the ACM* **55**(4), 77–84.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008), 'Fast unfolding of communities in large networks', *Journal of statistical mechanics: theory and experiment* **2008**(10), P10008.

Brin, S. & Page, L. (1998), 'The anatomy of a large-scale hypertextual web search engine', *Computer networks and ISDN systems* **30**(1-7), 107–117.

Fortunato, S. & Barthelemy, M. (2007), 'Resolution limit in community detection', *Proceedings of the national academy of sciences* **104**(1), 36–41.

Frigau, L., Contu, G., Mola, F. & Conversano, C. (2021), 'Network-based semisupervised clustering', *Applied Stochastic Models in Business and Industry* **37**(2), 182–202.

Frigau, L., Wu, Q. & Banks, D. (2022), 'Optimizing the jsm program', *Journal of the American Statistical Association* **117**(538), 617–626.

Groza, V., Udrescu, M., Bozdog, A. & Udrescu, L. (2021), 'Drug repurposing using modularity clustering in drug-drug similarity networks based on drug–gene interactions', *Pharmaceutics* **13**(12), 2117.

Honnibal, M. & Montani, I. (2017), spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Huang, M., Rao, Y., Liu, Y., Xie, H. & Wang, F. L. (2018), Siamese network-based supervised topic modeling, *in* 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing', pp. 4652–4662.

Jain, P. K., Pamula, R. & Srivastava, G. (2021), 'A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews', *Computer Science Review* **41**, 100413.

Javed, M. A., Younis, M. S., Latif, S., Qadir, J. & Baig, A. (2018), 'Community detection in networks: A multidisciplinary review', *Journal of Network and Computer Applications* **108**, 87–111.

Jung, S. & Segev, A. (2022), 'Analyzing the generalizability of the network-based topic emergence identification method', *Semantic Web* **13**(3), 423–439.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Misuraca, M., Forciniti, A., Scepi, G. & Spano, M. (2020), 'Sentiment analysis for education with r: packages, methods and practical applications', *arXiv preprint arXiv:2005.12840* .

Newman, M. E. & Girvan, M. (2004), 'Finding and evaluating community structure in networks', *Physical review E* **69**(2), 026113.

Newman, M. E. J. (2003), 'The structure and function of complex networks', *SIAM Review* **45**(2), 167–256.

Ortu, M., Frigau, L. & Contu, G. (2022), 'Topic based quality indexes assessment through sentiment', *Computational Statistics* pp. 1–23.

Park, S. & Kim, H. M. (2021), Data-driven customer segmentation based on online review analysis and customer network construction, *in* 'International Design Engineering Technical Conferences and Computers and Information in Engineering Conference', Vol. 85383, American Society of Mechanical Engineers, p. V03AT03A015.

Popping, R. (2003), 'Knowledge graphs and network text analysis', *Social science information* **42**(1), 91–106.

Romano, M., Contu, G., Mola, F. & Conversano, C. (2023), 'Threshold-based naïve bayes classifier', *Advances in Data Analysis and Classification* .

Tavazoee, F., Conversano, C. & Mola, F. (2020), 'Recurrent random forest for the assessment of popularity in social media', *Knowledge and Information Systems* **62**, 1847–1879.
URL: *https://doi.org/10.1007/s10115-019-01410-w*