QUALITY PAPER

Digital voice-of-customer processing by topic modelling algorithms: insights to validate empirical results

Received 15 July 2021
Revised 11 October 2021
Accepted 4 November 2021

Federico Barravecchia, Luca Mastrogiacomo and Fiorenzo Franceschini
Department of Management and Production Engineering (DIGEP), Politecnico di Torino, Turin, Italy

Abstract

Purpose – Digital voice-of-customer (digital VoC) analysis is gaining much attention in the field of quality management. Digital VoC can be a great source of knowledge about customer needs, habits and expectations. To this end, the most popular approach is based on the application of text mining algorithms named topic modelling. These algorithms can identify latent topics discussed within digital VoC and categorise each source (e.g. each review) based on its content. This paper aims to propose a structured procedure for validating the results produced by topic modelling algorithms.

Design/methodology/approach – The proposed procedure compares, on random samples, the results produced by topic modelling algorithms with those generated by human evaluators. The use of specific metrics allows to make a comparison between the two approaches and to provide a preliminary empirical validation.

Findings – The proposed procedure can address users of topic modelling algorithms in validating the obtained results. An application case study related to some car-sharing services supports the description.

Originality/value – Despite the vast success of topic modelling-based approaches, metrics and procedures to validate the obtained results are still lacking. This paper provides a first practical and structured validation procedure specifically employed for quality-related applications.

Keywords Quality 4.0, Digital voice-of-customer, Supervised validation, Topic modelling, Customer reviews, User-generated contents

Paper type Research paper

1. Introduction

The term Web 2.0 generically indicates the second phase of development and diffusion of the internet, characterised by a substantial increase in the interaction between site and user: greater participation of users, who often also become authors (blogs, chats, forums, wikis); more efficient sharing of information, which can be more easily retrieved and exchanged with peer to peer tools or with multimedia content dissemination systems; affirmation of social networks (Tirunillai and Tellis, 2014).

The authors gratefully acknowledge Dr Silvia Franco for her support in the preliminary activities of this study. This work has been partially supported by "Ministero dell'Istruzione, dell'Università e della Ricerca" Award "TESUN-83486178370409 finanziamento dipartimenti di eccellenza CAP. 1694 TIT. 232 ART. 6".

International Journal of Quality & Reliability Management
Vol. 39 No. 6, 2022
pp. 1453-1470
Emerald Publishing Limited
0265-671X
DOI 10.1108/IJQRM-07-2021-0217

Digital voice-of-customer (digital VoC), such as blogs, social media posts and online reviews, has achieved resounding success as a source of information because it is free, easily accessible and trustworthy (Özdağoğlu *et al.*, 2018). Digital VoC proved to represent a promising alternative to customer interviews as a source of helpful information to identify consumer needs (Barravecchia *et al.*, 2020a, b; Mastrogiacomo *et al.*, 2021).

Data mining and machine learning techniques make it possible to analyse large corpora of digital VoC to extrapolate the most relevant information, avoiding the impossible task of human reading and interpretation. Digital VoC often consists of unstructured textual information. For this reason, one of the most commonly implemented data mining techniques in these contexts is topic modelling. Topic modelling approaches are based on machine learning algorithms that can detect latent topics running through a collection of unstructured textual documents (Jelodar *et al.*, 2019). Given an extensive set of documents, topic modelling algorithms deal with the problems of (1) identifying a set of topics that describe a text corpus (i.e. a collection of text documents from a variety of sources), (2) associating a set of keywords to each topic and (3) defining a specific mixture of these topics for each document (Roberts *et al.*, 2019).

Most studies applying topic modelling algorithms on digital VoC only focused on the techniques used to extrapolate the topics, neglecting the critical issues related to the validation of the obtained results. This validation remains an open question, and a structured and shared approach is still lacking.

To close this gap, this paper investigates the following research question: *How to validate the results of topic modelling algorithms employed for digital VoC analysis in quality applications?*

Key elements of the novelty of this study are: (1) the exploration of the problem of validating the results of analyses based on a large corpus of digital VoC and (2) the proposal of a structured procedure for validating the results of topic modelling algorithms used in quality management applications.

The rest of the paper is organised into three sections. Section 2 presents the literature background. Section 3 introduces the proposed approach and an application case study. Finally, the concluding section summarises the original contributions of the paper focusing on theoretical and practical implications.

2. Background and literature review

2.1 Text mining and digital voice-of-customer

While industry faced its fourth industrial revolution, research on quality management also faced its own transformation to the new Quality 4.0 paradigm (Kannan and Garad, 2020; Zonnenshain and Kenett, 2020). The digitalisation of businesses generates unique opportunities for managing the quality of products and services (Sony *et al.*, 2020). In this context, awareness of the value of data generated directly by users is growing.

Many papers in the literature have shown how digital VoC, in particular online reviews, can be used to understand consumer preferences and latent quality dimensions (Mastrogiacomo *et al.*, 2021). Digital VoC is a valuable source of customer needs, and machine learning methods are likely to be more effective and efficient than conventional techniques (Özdağoğlu *et al.*, 2018). Traditionally, the detection of determinants that influence the perception of quality has been supported by quantitative methods, primarily based on data obtained through questionnaires and interviews (DeVellis, 2016). Albeit firmly established, these methods are quite expensive in terms of people involved and time. The quality of findings stemming from the implementation of these methodologies depends on the respondent's willingness to participate and on the complexity of the questionnaire (Groves, 2006). In addition, the use of questionnaires suffers from several limitations, including: (1) the limited sample size of respondents, (2) expert bias in defining the initial pool of items and (3) potential errors included in responses (DeVellis, 2016).

An alternative way to identify latent quality determinants of a product or service may be through the analysis of digital VoC and, more specifically, online reviews, which can offer a

low-cost, unbiased, and reliable source of information for understanding customer opinions, expectations, and requirements (Mastrogiacomo *et al.*, 2021). This detection is based on the in-depth analysis of such data, leveraging text mining tools able to infer information from text documents written in a natural language (Aggarwal and Zhai, 2012). The fundamental logic of these approaches is that if a product or service feature is discussed (within the digital VoC), it is critical to the definition of the quality of the object under investigation. Most previous studies trying to leverage text mining tools to analyse digital VoC focused on keyword frequency and sentiment analysis (Bi *et al.*, 2019; Liu, 2012). Few researchers applied topic modelling to identify quality determinants (Özdağoğlu *et al.*, 2018; Tirunillai and Tellis, 2014).

2.2 Topic modelling algorithms

Text mining refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. Text mining processes usually consist of five phases: (1) document collection, (2) pre-processing of the texts, (3) preparation and selection of the data, (4) knowledge extraction and (5) evaluation and interpretation of the results (Mastrogiacomo *et al.*, 2021). Within the vast family of text mining algorithms, topic modelling algorithms assume an essential role in analysing digital VoC for quality applications (Barravecchia *et al.*, 2020a, b; Mastrogiacomo *et al.*, 2021).

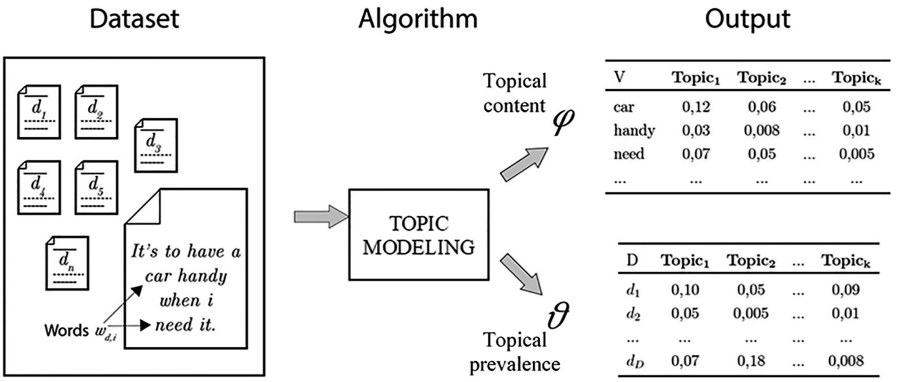
Topic modelling is a machine learning technique that allows extracting latent topics from a collection of unstructured textual documents (Blei *et al.*, 2003; Carnerud, 2017).

Figure 1 represents the general functioning of a topic modelling algorithm. Given an extensive collection of documents from a variety of sources (in the figure indicated as $d_1, \ldots, d_n$), topic modelling algorithms deal with the problems of:

(1) Identifying a set of topics that describe a text corpus (i.e. the collection of text documents);

(2) Associating a set of keywords to each topic (φ: topical content); and

(3) Defining a specific mixture of these topics for each document (ϑ: topical prevalence) (Blei *et al.*, 2003).



Source(s): $d_1, \ldots, d_D$ = textual documents (digital VoC). The output consists of φ (Topical content matrix) and θ (Topical prevalence matrix)

Figure 1. Graphical representation of the functioning of topic modelling algorithms

In recent years, a variety of topic modelling algorithms have been developed. The most diffused are the latent Dirichlet allocation (LDA) (Blei *et al.*, 2003), the hierarchical Dirichlet process (HDP) (Teh *et al.*, 2004), the structural topic model (STM) (Roberts *et al.*, 2019) and many other non-parametric models based on the Dirichlet process (Jelodar *et al.*, 2019).

2.3 Topic model validation

Validation is a crucial step in the development of a model because it establishes the trustworthiness of the obtained result. The importance of this activity is often underestimated, although the results of topic modelling can guide further decisions, including strategic ones.

Despite the importance of validating the results of topic modelling algorithms, there remains a paucity of a practical methodology for this purpose. A variety of metrics and criteria have been proposed (Wallach *et al.*, 2009), but standardised procedures to evaluate the outputs of a topic modelling algorithm are still lacking (Chang *et al.*, 2009; Kobayashi *et al.*, 2018).

Several automatic metrics have been suggested to assess the performance of topic modelling algorithms, of which predictive metrics are likely to be the most popular. Predictive metrics are calculated by developing the topic model using a set of documents, the so-called training set, and testing the model's reliability by applying it on a set of unseen documents, the so-called test set. Usually, 90% of the available documents are part of the training set, and the remaining 10% is part of the test set. The most commonly used predictive metric is the so-called "held-out likelihood". Held-out likelihood measures how likely some new unseen text documents are provided by the model that was learned earlier (Wang *et al.*, 2012). The range of the held-out likelihood is $(-\infty, 0]$. The higher this value, the more statistically strong the developed topic model is.

Other automatic metrics to evaluate the performance of topic modelling algorithms are:

(1) *Semantic coherence*, i.e. the average semantic relatedness between topic words (Newman *et al.*, 2010). Semantically coherent topics are intended as composed of words that should co-occur within the same document. Semantic coherence measures whether a topic is internally consistent (Mimno *et al.*, 2011; Roberts *et al.*, 2014). Semantic coherence ranges in the interval $(-\infty, 0]$. The higher the semantic coherence value, the more semantically consistent the identified latent topics are.

(2) *Exclusivity*, i.e. a metric that measures the extent to which the top words for each topic do not appear as top words in other topics. In detail, if words with high probability under topic i have low probabilities under other topics, then we say that topic i is exclusive. Exclusivity is defined in the range $[0, +\infty)$. The higher the exclusivity, the more distinct are the identified latent topics. A coherent and exclusive topic is more likely to be semantically valid (Bischof and Airoldi, 2012; Roberts *et al.*, 2014).

The main strength of these assessments is that they can be calculated automatically without the need for human input. This allows using these metrics to set the input parameters of the topic models and automatically measure the output quality to select their optimal values. These metrics have also found wide application in comparing the performance of different topic modelling algorithms (Wallach *et al.*, 2009). Automatic metrics allow to test the performance of different topic models in real time. On the other hand, the main weakness of this methodology is the fact that these criteria do not consider the semantic meaning of the topics, and consequently, they are not fully applicable for an assessment of the developed topic model.

Some preliminary works have begun to raise the issue of the quality of the results of these approaches, also proposing alternative solutions based on a supervised evaluation of the

outcomes of topic modelling algorithms (Chang *et al.*, 2009). Supervised criteria inherently require assessments by human evaluators and introduce evaluations based on comprehensibility, consistency of topics and document classifications. However, the major shortcoming of supervised methods is that they require a considerable amount of time and resources to be employed.

Chang *et al.* (2009) demonstrated that automatic criteria do not adequately capture whether topics are coherent or not, and that topic models that perform better on held-out likelihood may infer less semantically meaningful topics. Chang *et al.* (2009) also introduced two supervised metrics: (1) word intrusion, i.e. how well the inferred topics match human concepts; (2) topic intrusion, i.e. how well a topic model assigns topics to documents.

Costa *et al.* (2007) propose a series of metrics to assess the performance of classifier algorithms. Their work suggests tracing the problem back to a binary classification problem and identifying when the object (in our case the document) has been correctly or incorrectly classified (true positive, true negative, false positive, false negative). On the basis of these supervised evaluations, it is possible to calculate performance metrics (accuracy rate, error rate, precision, specificity, etc.).

Table 1 compares some characteristics of the automatic and supervised validation criteria. In detail, the main differences are: (1) the automation of the validation process, possible with automatic criteria and not feasible for supervised approaches; and (2) the quality of the final outcome, only supervised criteria take into account the semantics of the analysed texts and the intelligibility of the identified topic.

In short, the literature offers two main potential approaches. The first is based on automatic validation criteria (i.e. unsupervised). Such approach can be adopted when verification times need to be reduced (e.g. optimisation of the model's parameters). The second relies on supervised evaluation. These approaches are preferable when evidence of the quality of the topic model is required. Some supervised-based criteria have already been proposed in the literature. However, a formal procedure for the supervised validation of the outcomes of topic modelling algorithms is still missing. The objective of this paper is, therefore, to try to fill this gap. The following section introduces an empirical methodology for validating the results of topic modelling algorithms based on the assessment of human evaluators.

3. The proposed methodology

This section aims to provide a practical methodology to support users of topic modelling algorithms to validate obtained results.

The validation procedure is organised in four steps (Figure 2): (1) Sample extraction and human topic assignment, (2) automatic topic assignment, (3) comparison of results and (4) metrics calculation. Table 2 summarises the main inputs and outputs of the four steps.

The following subsections detail the aforementioned steps. The description of a simple case study accompanies the description of the method.

	Automatic criteria	Supervised criteria
Required time	Low	High
Fully automated evaluation process	Yes	No
Semantic evaluation	No	Yes
Need for human input	No	Yes
Topic intelligibility evaluation	No	Yes
Suitable for setting model parameters	Yes	No
Sample size	Complete database	Sample of documents

Table 1. Comparison between the automatic performance metrics and the supervised validation criteria

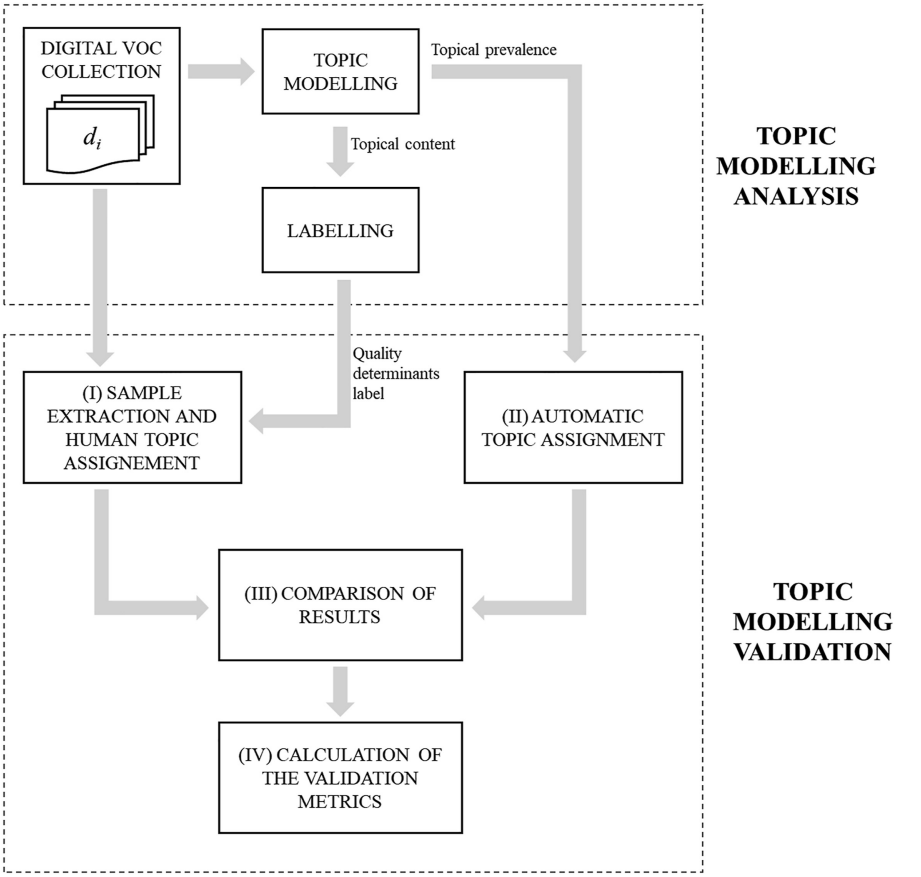


Figure 2.
Main steps of the proposed validation procedure

Step	Input	Output
(I) SAMPLE EXTRACTION AND HUMAN TOPIC ASSIGNMENT	Digital VoC collection + quality determinants labels	Human topic assignment (each item of the sample is associated with one or more quality determinants)
(II) AUTOMATIC TOPIC ASSIGNMENT	Topical prevalence distributions	Automatic topic assignment (each item of the sample is associated with one or more quality determinants)
(III) RESULT COMPARISON	Human topic assignment + automatic topic assignment	Confusion matrix
(IV) VALIDATION METRICS CALCULATION	Confusion matrix	Validation metrics

Table 2.
Input and output of the main steps of the proposed validation procedure

3.1 Application case

The case study concerns the analysis of digital VoC regarding car-sharing services (Mastrogiacomo *et al.*, 2021). Analysed data are reviews retrieved in December 2019 from

different review aggregators: Yelp, Google, Trustpilot, Facebook and Play Store. Reviews were published from January 2010 to December 2019. Only English-language reviews were selected, with a total of almost 17,000 reviews from 22 car-sharing providers (Car2go, DriveNow, Maven, Zipcar, Goget), operating in three different countries (USA, Canada, and UK). STM has been applied for the identification of topics (Roberts *et al.*, 2019). Using the held-out likelihood criteria, the optimal number of identified topics was 20. The graph in Figure 3 shows the values of the held-out likelihood as a function of T (from 5 to 100). From the graph, we can observe that starting from a value of T equal to 20, there is an almost stationary held-out likelihood.

Table 3 reports the topics described by the keywords and the corresponding assigned labels.

3.2 Sample extraction and human topic assignment

The first step in the procedure regards the extraction of a random sample n of documents, subsequently categorised by a human evaluator. The value of n should be high enough to allow a reliable validation of the results, but at the same time, it should not be too high to avoid an excessive workload for the human evaluators. In practice, a value of $n = 100$ can be considered sufficient to obtain good results and is also convenient for human topic assignments.

The proposed supervised approach requires that the extracted sample of digital VoC to be read and classified by human evaluators. This assessment can be carried out by one or more subjects. Its robustness increases with the number and the agreement of subject involved. Each evaluator is required to carefully read the extracted sample of digital VoC and classify the content of each document according to the labels of the identified quality determinants (Table 3). Using the label list of quality determinants allows concurrently validating both the generated topical prevalence distributions (for each document) and the list of quality determinants labels. To comply with the underlying assumption of topic modelling algorithms – according to which each document may contain a mix of different topics – the evaluator is required to identify one or more quality determinants discussed within the document. If the content of the analysed digital VoC is unclear or non-classifiable, the human evaluator may also report the absence of a quality determinant.

Table 4 shows some examples of topic assignments performed for the proposed case study.

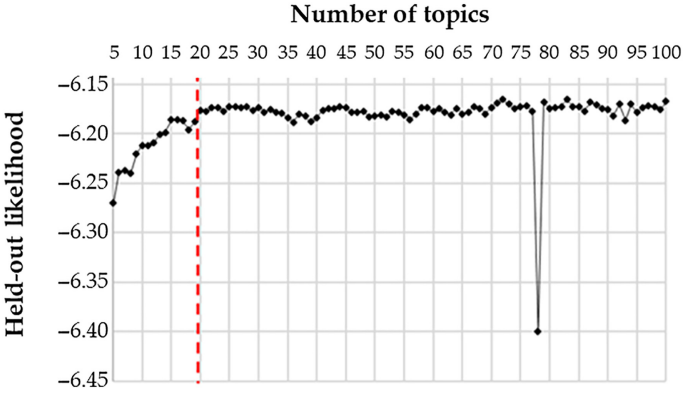


Figure 3.
Results of the held-out likelihood analysis to determine the optimal number of topics (ranging from 5 to 100)

	Criterium	Keywords	Label
1	Highest prob	Help, phone, call, person, office, answer, number	CUSTOMER SERVICE (PHYSICAL OFFICE)
	FREX	Help, office, staff, answer, phone, person, question	
2	Highest prob	Damage, report, accident, fault, member, enterprise, claim	ACCIDENT AND DAMAGES MANAGEMENT
	FREX	Damage, accident, minor, claim, deduct, fault, scratch	
3	Highest prob	Sign, process, website, license, drive, driver, registration	REGISTRATION PROCESS
	FREX	License, application, registration, process, driver, website, sign	
4	Highest prob	Charge, fee, late, return, time, pay, hour	CHARGES AND FEES
	FREX	Fee, charge, late, return, dollar, extra, extend	
5	Highest prob	Park, lot, spot, find, ticket, street, space	PARKING AREAS
	FREX	Park, spot, ticket, space, street, lot, mete	
6	Highest prob	App, work, update, book, map, reserve, time	APP RELIABILITY
	FREX	App, map, crash, feature, load, slow, version	
7	Highest prob	Trip, end, time, make, actual, take, system	END TRIP ISSUES
	FREX	Trip, end, life, stuck, make, connect, actual	
8	Highest prob	Gas, dirty, rent, clean, tank, card, tire	CAR CONDITION
	FREX	Dirty, smell, smoke, tank, hair, seat, dog	
9	Highest prob	Need, convenient, quick, recommend, awesome, clean, perfect	CONVENIENCE
	FREX	Awesome, excel, amazing, perfect, quick, convenient, super	
10	Highest prob	Hour, price, rate, cost, expense, mile, cheaper	USE RATES
	FREX	Price, rate, expense, cost, daily, tax, rental	
11	Highest prob	Minute, reservation, walk, wait, home, time, away	CAR PROXIMITY
	FREX	Minute, walk, home, wait, figure, block, stand	
12	Highest prob	Car, available, location, vehicle, area, change, time	CAR AVAILABILITY
	FREX	Available, vehicle, car, select, search, choose, date	
13	Highest prob	Use, time, now, far, user, review, star	EFFICACY
	FREX	Use, user, far, happy, review, love, code	
14	Highest prob	City, year, insurance, member, gas, need, month	SHARING BENEFITS
	FREX	Errand, hybrid, live, city, SUV, insurance, variety	
15	Highest prob	Service, custom, issue, company, terrible, problem, experience	CUSTOMER SERVICE RESPONSIVENESS
	FREX	Service, custom, terrible, issue, support, resolve, company	
16	Highest prob	Way, drive, little, take, get, town, bus	INTERMODAL TRANSPORTATION
	FREX	Town, bus, airport, taxi, bike, store, run	
17	Highest prob	Time, start, location, turn, lock, pick, key	CAR START-UP ISSUES
	FREX	Key, turn, battery, lock, door, start, waste	
18	Highest prob	Call, member, cancel, ask, rep, refund, manage	CUSTOMER SERVICE COURTESY
	FREX	Representative, supervisor, agent, rude, manage, speak, conversation	
19	Highest prob	Account, card, email, credit, month, day, membership	BILLING AND MEMBERSHIP
	FREX	Account, email, payment, bank, credit, card, address	
20	Highest prob	Reservation, plan, time, need, book, cancel, advance	CAR RESERVATION
	FREX	Plan, advance, entire, reservation, chance, screw, ruin	

Note(s): Case study on car-sharing services

Table 3.
Topic keywords and topic labels

3.3 Automatic topic assignment

The topic modelling algorithm assigns each document a multinomial distribution that describes the probability that the identified topics are discussed within the document, the

ID	Review	Human topic assignment
1	Such a great idea! Such poor customer service! In theory, there is an office on Rhode Island Ave, but do not count on it being staffed. After I discovered that someone had left a bag filled with valuables in a car I rented and after the 24 h "help desk" said that their policy was to try to hide valuables in the car so that the owner could retrieve them if/when they discovered their loss, I went to the office on Rhode Island to leave the valuables there	Topic 1: customer service (physical office)
2	I have been a member of this company for about 6 months now and have had mixed experiences. For short trips in place of a cab, it is just fine, but when it comes to day-long rentals, there are hidden fees, and any time I have needed to use their customer service (twice), I have found it to be infuriating enough to want to cancel the membership altogether use them to take your groceries home and that is all	Topic 4: charges and fees Topic 9: convenience Topic 10: use rates Topic 15: customer service responsiveness
3	Disclaimer: this service does not work, at least not for me. I have used (attempted) this service three times, all three times I could not get into the car. The first two I had some spare time, so I called customer service who spent 15 min on the phone with me and finally got me in. The last time I was in a hurry and could not do that ended up getting an uber. 3 attempts 3 fails	Topic 15: customer service responsiveness Topic 17: car start-up issues
4	Reserved a car for a day. Less than 24 h before the rent period started, I got a weird email from them – they simply moved my reservation to another car in another city without asking me. When I tried to cancel, I was told that I cannot cancel 24 h before the rent period starts. I tried to call them, they did not pick up the phone. There is no other way to contact them There are weird emails that do not work. Oh and your monthly membership? You cannot cancel it The only way to cancel is through phone, which they do not pick up. I am not convinced this is not a giant scam. Stay away from this company!	Topic 17: car start-up issues Topic 18: customer service courtesy Topic 20: car reservation

Note(s): Case study on car-sharing services

Table 4.
Example of topic assignment performed by a human evaluator

so-called topical prevalence. Given the multinomial distribution produced by the algorithm, it is necessary to identify which topics are most likely to be discussed in the document (i.e. automatic topic assignment).

There are no reference standards for automatic topic assignment to rely on, so different applications may use different rules. Table 5 reports the multinomial probability distributions assigned to the reviews shown in Table 4. Relevant topics can be selected according to three different strategies: highest probability, static threshold or the proposed dynamic threshold.

A considerable number of applications consider the topic with the highest probability as the representative topic of the document (Jelodar *et al.*, 2019). However, this approach has some critical limitations since it contrasts with the underlying principle of topic modelling algorithms, according to which each document can discuss a mix of topics. Consider as an example the multinomial distribution related to Document 4 shown in Table 5. Topic 17 has a score of 0.21, while Topic 28 has a score of 0.20. The rule of maximum would only indicate Topic 17 as the only representative topic of the document.

An alternative approach is defining a static threshold above which the topic can be considered as representative. However, this approach also has shortcomings: relevant topics may not be identified or marginal topics may be indicated as relevant. Considering again the examples provided in Table 5, the fixed threshold set at 0.15 causes no topic to be identified in Document 2.

Table 5.
Reviews' topic proportion and review's topic assignment

ID	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	Highest probability	Static threshold *(0.15)*	Dynamic threshold $DT_i = Q3_i + (1.5 \cdot IQR_i)$
1	*0.47*	0.01	0.01	0.03	0.02	0.01	0.04	0.04	0.01	0.01	0.07	0.01	0.02	0.01	0.08	0.02	0.05	0.05	0.03	0.01	T1	T1	T1 ($DT_1 = 0.10$)
2	0.04	0.04	0.03	*0.14*	0.02	0.02	0.07	0.01	0.04	0.06	0.07	0.04	0.07	0.04	0.05	0.08	0.04	0.04	0.05	0.05	T4		T4 ($DT_2 = 0.10$)
3	0.03	0.01	0.01	0.03	0.02	0.04	0.08	0.01	0.01	0.01	*0.14*	0.02	0.04	0	0.08	0.03	*0.36*	0.03	0.03	0.02	T17	T17	T11 + T17 ($DT_3 = 0.09$)
4	0.02	0.01	0.02	0.04	0.01	0.01	0.01	0.02	0.01	0.04	0.03	0.01	0.02	0.02	0.02	0.08	*0.21*	*0.2*	*0.14*	0.07	T1	T1	T17 + T18 + 19 ($DT_4 = 0.12$)

Note(s): T1, T2, . . . , T20 means Topic 1, Topic 2, . . . , Topic 20

To overcome these limitations, we propose the adoption of a *dynamic thresholds* to identify the representative topics of a document. The dynamic threshold is better fitted to the distribution under consideration and consequently provides a more accurate identification of relevant topics.

The problem of identifying relevant topics can be traced back to the problem of identifying outliers in a distribution. A variety of methodologies are available to detect outliers in multinomial distribution (García-Heras *et al.*, 1993). Given a non-normal multinomial distribution resulting from the application of topic modelling algorithms, we suggest the use of the Tukey fence non-parametric outlier detection method (Rousseeuw *et al.*, 1999; Zijlstra *et al.*, 2007). Values outside the upper Turkey fence were considered as outliers and such outliers were identified as topics discussed within the document.

For each multinomial distribution associated with the i-th document, data are sorted into ascending sequences to obtain $Q1_i$ and $Q3_i$, i.e. the lower and upper quartile points. The difference between $Q1_i$ and $Q3_i$, namely, inter-quartile range (IQR_i), can be used to assess the dynamic threshold as follows:

$$DT_i = Q3_i + (1.5 \cdot IQR_i) \quad (1)$$

3.4 Comparison of results

It is then possible to compare the evaluations obtained in the previous step with the results generated by a topic modelling algorithm. The human topic assignment can act as a "golden" reference for the result of the topic modelling: for each review and topic, the four possible cases reported in the confusion matrix represented in Figure 4 can occur.

Table 6 shows some examples of comparisons between the evaluation by humans and by the topic model. For example, in Document 1, the two evaluations agree to identify topic 1 as the only topic discussed in the document. This provides one true positive and 19 true negatives. In Document 3, the human evaluation identified two topics, Topics 15 and 17, while the topic modelling algorithm identified Topics 11 and 17. This provides one true positive for Topic 17, one false positive for Topic 11, one false negative for Topic 15 and 16 true negatives. This operation is repeated for all the documents included in the extracted sample. Figure 5 reports the confusion matrix for the analysed case study (sample of four reviews).

		Human topic assignment ***(true condition)***	
		T_i existence	*T_i non-existence*
Automatic topic assignment	*T_i existence*	**True Positive (tp)** Correct inference Agreement between human and automatic assignment. Both recognise the presence of a topic in a review.	**False Positive (fp)** Type I error Misalignment between human and automatic assignment. The automatic assignment recognises the presence of a topic, while the human assignment does not.
	T_i non-existence	**False Negative (fn)** Type II error Misalignment between human and automatic assignment. The human assignment recognises the presence of a topic, while the automatic assignment does not.	**True Negative (tn)** Correct inference Agreement between human and automatic assignment. Both do not recognise the presence of a topic in a review.

Figure 4.
Confusion matrix for topic modelling assignments

3.5 Calculation of the validation metrics

Based on the comparison of the results obtained by the two methods of topic assignment (automatic and human), it is possible to calculate some validation metrics presented below and summarised in Figure 6 (Costa *et al.*, 2007; Franceschini *et al.*, 2019; Maria Navin and Pankaja, 2016; Zaki and McColl-Kennedy, 2020).

Accuracy evaluates the effectiveness of the algorithm by its percentage of correct predictions. It is defined as the ratio of correct predictions related to both the presence and non-presence of topics compared to the total observations. Accuracy can be calculated as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \text{tp}_i + \sum_{i=1}^{n} \text{tn}_i}{\sum_{i=1}^{n} \text{tp}_i + \sum_{i=1}^{n} \text{tn}_i + \sum_{i=1}^{n} \text{fp}_i + \sum_{i=1}^{n} \text{fn}_i} \quad (2)$$

where:

- $\sum_{i=1}^{n} \text{tp}_i$ is the total number of true positives observed comparing human and automatic topic assignments,
- $\sum_{i=1}^{n} \text{tn}_i$ is the total number of true negatives,
- $\sum_{i=1}^{n} \text{fp}_i$ is the total number of false positives,
- $\sum_{i=1}^{n} \text{fn}_i$ is the total number of false negatives.
- n is the sample size of the analysed records.

ID	Human topic assignment	Automatic topic assignment	tp	fp	fn	tn
1	T1	T1	1	0	0	19
2	T4, T9, T10 and T15	T4	1	0	3	16
3	T15 and T17	T11 and T17	1	1	1	17
4	T17, T18 and T20	T17, T18 and 19	2	1	1	16

Note(s): tp = true negatives; fp = false positives; fn = false negatives; tn = true negatives. T1, T2, . . . , T20 means Topic 1, Topic 2, . . . , Topic 20

Table 6.
Example of results comparison

		Human topic assignment *(true condition)*	
		T_i existence	*T_i non-existence*
Automatic topic assignment	*T_i existence*	$\sum_{i=1}^{n} tp_i = 5$	$\sum_{i=1}^{n} fp_i = 2$
	T_i non-existence	$\sum_{i=1}^{n} fn_i = 5$	$\sum_{i=1}^{n} tn_i = 68$

Note(s): $\sum_{(i\,=\,1)}^{n} tp_i$ is the total number of true positives observed comparing human and automatic topic assignments. $\sum_{(i\,=\,1)}^{n} tn_i$ is the total number of true negatives. $\sum_{(i\,=\,1)}^{n} fp_i$ is the total number of false positives. $\sum_{(i\,=\,1)}^{n} fn_i$ is the total number of false negatives. Values refer to the sample (four reviews) used to exemplify the proposed procedure

Figure 5.
Confusion matrix example

		Human topic assignment ***(true condition)***			
		T_i *existence*	T_i *non-existence*	$\frac{Accuracy}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i + \sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} fn_i}$	
Automatic topic assignment	T_i *existence*	**True Positive** (tp) Correct inference	**False Positive** (fp) Type I error	$\frac{Precision}{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fp_i}$	$\frac{False\ discovery\ rate}{\sum_{i=1}^{n} fp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fp_i}$
	T_i *non-existence*	**False Negative** (fn) Type II error	**True Negative** (tn) Correct inference	$\frac{False\ omission\ rate}{\sum_{i=1}^{n} fn_i}{\sum_{i=1}^{n} fn_i + \sum_{i=1}^{n} tn_i}$	$\frac{Negative\ predictive\ value}{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} fn_i + \sum_{i=1}^{n} tn_i}$
		$\frac{Recall}{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i}$	$\frac{Fall\text{-}out}{\sum_{i=1}^{n} fp_i}{\sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} tn_i}$	$2 \cdot \frac{F_1\ Score}{Precision \cdot Recall}{Precision + Recall}$	
		$\frac{Miss\ rate}{\sum_{i=1}^{n} fn_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fn_i}$	$\frac{Specificity}{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} tn_i}$		

Note(s): $\sum_{(i=1)}^{n} tp_i, \sum_{(i=1)}^{n} tn_i, \sum_{(i=1)}^{n} fp_i, \sum_{(i=1)}^{n} fn_i$ indicate respectively the total amount of true positives, true negatives, false positives, and false negatives observed when comparing human and automatic topic assignments. n is the sample size of the analysed records

Figure 6.
Quality metrics for topic model validation

Precision (also called positive predictive value) is an estimate of the probability that a positive prediction is correct. It is the ratio between the correctly predicted positive observations (i.e. correctly predicted presence of topics) and the total predicted positive observations (i.e. correctly predicted presence and non-presence of topics). Precision can be calculated as:

$$\text{Precision} = \frac{\sum_{i=1}^{n} \text{tp}_i}{\sum_{i=1}^{n} \text{tp}_i + \sum_{i=1}^{n} \text{fp}_i} \tag{3}$$

Recall (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved. In topic modelling analysis, recall represents the ratio between the correctly predicted topic and the total amount of predicted topics, either correctly and incorrectly. Recall can be calculated as:

$$\text{Recall} = \frac{\sum_{i=1}^{n} \text{tp}_i}{\sum_{i=1}^{n} \text{tp}_i + \sum_{i=1}^{n} \text{tn}_i} \tag{4}$$

More generally, recall is the complement to unity of the Type II error rate (β):

$$\text{Recall} = 1 - \beta \tag{5}$$

The F_1 *score* is a measure of a test's accuracy, and it is calculated as the harmonic mean of precision and recall. This score takes both false positives and false negatives into account. The F_1 score can be calculated as:

$$F_1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

The *fall-out* (or false positive rate) is the proportion of all negatives yielding to positive test outcomes, i.e. the conditional probability of detecting a topic that in reality is not present. The fall-out can be calculated as:

$$\text{Fall} - \text{out} = \frac{\sum_{i=1}^{n} \text{fp}_i}{\sum_{i=1}^{n} \text{fp}_i + \sum_{i=1}^{n} \text{tn}_i} \tag{7}$$

The fall-out is equal to the significance level (α):

$$\text{Fall} - \text{out} = \alpha \tag{8}$$

Complementarily, the *Miss rate* (or false-negative rate) is the proportion of positives that yield negative test outcomes, i.e. the conditional probability of not identifying a topic when it is present. Miss rate is equal to the Type II error rate (β). The Miss rate can be calculated as:

$$\text{Miss rate} = \frac{\sum_{i=1}^{n} \text{fn}_i}{\sum_{i=1}^{n} \text{tp}_i + \sum_{i=1}^{n} \text{fn}_i} \tag{9}$$

Specificity measures the proportion of true negatives. In topic modelling, it can be interpreted as the proportion of topics that are not identified because they are not actually present. Specificity can be calculated as:

$$\text{Specificity} = \frac{\sum_{i=1}^{n} \text{tn}_i}{\sum_{i=1}^{n} \text{fp}_i + \sum_{i=1}^{n} \text{tn}_i} \tag{10}$$

The *negative predictive value* is the probability that the topic modelling algorithm will not detect a topic when it is not actually present. The negative predictive value can be calculated as:

$$\text{Negative predictive value} = \frac{\sum_{i=1}^{n} \text{tn}_i}{\sum_{i=1}^{n} \text{fn}_i + \sum_{i=1}^{n} \text{tn}_i} \tag{11}$$

The complement to the unity of the negative predictive value is the *false omission rate* that is the proportion of topics not detected when the topic was instead present. The false omission rate can be calculated as:

$$\text{False omission rate} = \frac{\sum_{i=1}^{n} \text{fn}_i}{\sum_{i=1}^{n} \text{fn}_i + \sum_{i=1}^{n} \text{tn}_i} \tag{12}$$

Finally, the *false discovery rate* is the proportion of erroneously identified topics compared to all identified topics. The false discovery rate can be calculated as:

$$\text{False discovery rate} = \frac{\sum_{i=1}^{n} \text{fp}_i}{\sum_{i=1}^{n} \text{tp}_i + \sum_{i=1}^{n} \text{fp}_i} \tag{13}$$

Table 7 shows the values of the indicators related to the case study under analysis.

Metrics	Example (4 reviews)	Case study
Validation metrics		
Accuracy	0.91	0.96
Recall	0.50	0.68
Precision	0.71	0.91
F_1 score	0.59	0.78
Miss rate	0.50	0.32
Fall-out	0.03	0.01
Specificity	0.97	0.99
False omission rate	0.07	0.04
False discovery rate	0.29	0.09
Negative predictive value	0.93	0.96
Automatic performance metrics		
Held-out likelihood	−6.17	
Range $\epsilon(-\infty, 0]$		
Semantic coherence	−101.70	
Range $\epsilon(-\infty, 0]$		
Exclusivity	9.77	
Range $\epsilon[0, \infty)$		

Note(s): Column "Example" refers to the sample (four reviews) used to exemplify the proposed procedure (confusion matrix in Figure 5). Column "Case study" refers to the entire validation sample (100 reviews)

Table 7.
Example of metrics for topic model validation

The goodness of validation results can be obtained by comparing the values of the validation metrics with desired target values (Table 8).

In addition to the supervised validation metrics, Table 7 reports the values of the automatic performance metrics calculated for the case study. Since no reference targets are defined, it is difficult to understand whether these values refer to a good performance. Held-out likelihood, for example, ranges between $\vdash(-\infty, 0]$. Even though we know that the higher the held-out likelihood, the better the behaviour is, in the absence of an absolute reference, we cannot easily discern whether that value is good or not. The same holds for the other automatic metrics. For these reasons, despite the wide use of these metrics as tools to calibrate some parameters of topic model algorithms, their application as validation metrics appears to be inadequate (Chang *et al.*, 2009).

By contrast, the proposed metrics have a well-defined co-domain and specific target values (Table 8). This allows an immediate evaluation of the quality of the results obtained without the need for comparisons with other results or models. Moreover, the variety of the proposed metrics allows to evidence strengths and weaknesses of the developed model (e.g. errors resulting from lack of sensitivity or due to false topic detection).

Indicators	Range	Direction	Target values
Accuracy	[0;1]	High is good	>0.95
Recall	[0;1]	High is good	>0.70
Precision	[0;1]	High is good	>0.70
F_1 score	[0;1]	High is good	>0.70
Miss-rate	[0;1]	Low is good	<0.20
Fall-out	[0;1]	Low is good	<0.05
Specificity	[0;1]	High is good	>0.90
False omission rate	[0;1]	Low is good	<0.05
False discovery rate	[0;1]	Low is good	<0.05
Negative predictive value	[0;1]	High is good	>0.90

Table 8.
Reference values for topic model validation indicators

3.6 Further consideration about validation metrics

Although the literature does not provide a gold standard concerning target values for these metrics, their calculation allows a preliminary assessment of the goodness of results. Table 8 shows the ranges, direction and target values for the proposed metrics.

Values distant from the targets listed in Table 8 generally indicate that the generated topic model does not adequately describe the semantic content of the analysed set of documents.

In the following, we report some corrective to be taken when the values of these metrics are not acceptable:

(1) *Review of the labels assigned to the identified quality determinants.* They are not representative of the identified topic.

(2) *Review of the input parameters of the topic model algorithm.* A non-optimal choice of the model input parameters (e.g. the number of topics) may lead to a bad categorisation of the content of documents that does not reflect the actual semantic content;

(3) *Review of the considered covariates.* It becomes important when topic modelling involves the joint analysis of text and metadata (e.g. in STM);

(4) *Pertinence and adequacy of the analysed database.* If the collection of digital VoC is highly heterogeneous in terms of typology and content, topic modelling can produce low-quality results.

4. Conclusions

The application of topic modelling algorithms in quality management is experiencing growing success. From their intrinsic design, topic modelling algorithms can be thought of as black boxes that receive as input textual data and produce as output a model capable of describing the semantic content (topics) of the analysed documents. The quality of the results is strongly influenced by the model parameters defined by the users, often using empirical procedures.

Automatic criteria generally applied to evaluate the performance of topic models (e.g. held-out likelihood) are not exhaustive and present several limitations (e.g. they do not consider the comprehensibility and the semantic of the topics). To overcome these limitations, it is necessary to rely on supervised criteria. Supervised approaches, based on human evaluations, are preferred whenever robust evidence of the quality of the results obtained is required.

This paper provides a first practical and structured procedure to validate the results of topic modelling algorithms specifically employed for quality-related applications. The proposed method is based on a comparison of the outputs produced automatically by topic modelling algorithms with those generated by human evaluators on a limited random sample of documents.

From a theoretical point of view, this paper proposes: (1) a practical method for compiling a confusion matrix helpful in comparing the results of automatic and human topic assignments; (2) the adoption of a dynamic threshold to automatically and rigorously determine which topics the algorithm identifies as relevant within each document; (3) a comprehensive set of metrics to allow comparison between the outputs of topic modelling algorithms and human-supervised classification.

From a practical point of view, this paper highlights the need for a supervised validation and proposes a structured procedure. This procedure can become a reference for all practitioners who must face the problem of empirical validating the results of an analysis of the digital VoC.

Future research steps will include implementing further experiments and applications to provide more evidence on the effectiveness of the proposed procedure.

References

Aggarwal, C.C. and Zhai, C. (2012), *Mining Text Data*, Springer, New York.

Barravecchia, F., Mastrogiacomo, L. and Franceschini, F. (2020a), "Categorizing quality determinants mining user-generated contents", *Sustainability*, Vol. 12 No. 23, p. 9944.

Barravecchia, F., Mastrogiacomo, L. and Franceschini, F. (2020b), "Identifying car-sharing quality determinants: a data-driven approach to improve engineering design", in *International Conference on Quality Engineering and Management*, pp. 125-140.

Bi, J.-W., Liu, Y., Fan, Z.-P. and Cambria, E. (2019), "Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model", *International Journal of Production Research*, Vol. 57 No. 22, pp. 7068-7088.

Bischof, J. and Airoldi, E.M. (2012), "Summarizing topical content with word frequency and exclusivity", in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 201-208.

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

Carnerud, D. (2017), "Exploring research on quality and reliability management through text mining methodology", *International Journal of Quality and Reliability Management*, Vol. 34 No. 7, pp. 975-1014.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L. and Blei, D.M. (2009), "Reading tea leaves: how humans interpret topic models", *Advances in Neural Information Processing Systems*, pp. 288-296.

Costa, E., Lorena, A., Carvalho, A. and Freitas, A. (2007), "A review of performance evaluation measures for hierarchical classifiers", in *Evaluation Methods for Machine Learning II: Papers from the AAAI-2007 Workshop*, pp. 1-6.

DeVellis, R.F. (2016), *Scale Development: Theory and Applications*, SAGE Publications, Newbury CA.

Franceschini, F., Galetto, M. and Maisano, D. (2019), *Designing Performance Measurement Systems. Management for Professionals*, Springer Nature, Cham.

García-Heras, J., Muñoz-García, J. and Pascual-Acosta, A. (1993), "Criteria for the detection of outliers in the multinomial model", *Journal of Applied Statistics*, Vol. 20 No. 1, pp. 137-142.

Groves, R.M. (2006), "Nonresponse rates and nonresponse bias in household surveys", *Public Opinion Quarterly*, Vol. 70 No. 5, pp. 646-675.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. (2019), "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", *Multimedia Tools and Applications*, Vol. 78 No. 11, pp. 15169-15211.

Kannan, K.S.P.N. and Garad, A. (2020), "Competencies of quality professionals in the era of industry 4.0: a case study of electronics manufacturer from Malaysia", *International Journal of Quality and Reliability Management*, Vol. 38 No. 3, pp. 839-871.

Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihók, G. and Den Hartog, D.N. (2018), "Text mining in organizational research", *Organizational Research Methods*, Vol. 21 No. 3, pp. 733-765.

Liu, B. (2012), "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, Vol. 5 No. 1, pp. 1-167.

Maria Navin, J.R. and Pankaja, R. (2016), "Performance analysis of text classification algorithms using confusion matrix", *International Journal of Engineering and Technical Research (IJETR)*, pp. 869-2321.

Mastrogiacomo, L., Barravecchia, F., Franceschini, F. and Marimon, F. (2021), "Mining quality determinants of Product-Service Systems from unstructured User-Generated Contents: the case of car-sharing", *Quality Engineering*, Vol. 33 No. 3, pp. 425-442.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and McCallum, A. (2011), "Optimizing semantic coherence in topic models", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262-272.

Newman, D., Lau, J.H., Grieser, K. and Baldwin, T. (2010), "Automatic evaluation of topic coherence", in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100-108.

Özdağoğlu, G., Kapucugil-İkiz, A. and Çelik, A.F. (2018), "Topic modelling-based decision framework for analysing digital voice of the customer", *Total Quality Management and Business Excellence*, Vol. 29 Nos 13-14, pp. 1545-1562.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G. (2014), "Structural topic models for open-ended survey responses", *American Journal of Political Science*, Vol. 58 No. 4, pp. 1064-1082.

Roberts, M.E., Stewart, B.M. and Tingley, D. (2019), "STM: r package for structural topic models", *Journal of Statistical Software*, Vol. 91 No. 2, pp. 1-40.

Rousseeuw, P.J., Ruts, I. and Tukey, J.W. (1999), "The bagplot: a bivariate boxplot", *The American Statistician*, Vol. 53 No. 4, pp. 382-387.

Sony, M., Antony, J. and Douglas, J.A. (2020), "Essential ingredients for the implementation of Quality 4.0", *TQM Journal*, Vol. 32 No. 4, pp. 779-793.

Teh, Y., Jordan, M., Beal, M. and Blei, D. (2004), "Sharing clusters among related groups: hierarchical Dirichlet processes", *Advances in Neural Information Processing Systems*, Vol. 17, pp. 1385-1392.

Tirunillai, S. and Tellis, G.J. (2014), "Mining marketing meaning from online chatter: strategic brand analysis of big data using latent Dirichlet allocation", *Journal of Marketing Research*, Vol. 51 No. 4, pp. 463-479.

Wallach, H.M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009), "Evaluation methods for topic models", in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105-1112.

Wang, Y., Agichtein, E. and Benzi, M. (2012), "TM-LDA: efficient online modeling of latent topic transitions in social media", in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 123-131.

Zaki, M. and McColl-Kennedy, J.R. (2020), "Text mining analysis roadmap (TMAR) for service", *Journal of Services Marketing*, Vol. 34 No. 1, pp. 30-47, doi: 10.1108/JSM-02-2019-0074.

Zijlstra, W.P., Van Der Ark, L.A. and Sijtsma, K. (2007), "Outlier detection in test and questionnaire data", *Multivariate Behavioral Research*, Vol. 42 No. 3, pp. 531-555.

Zonnenshain, A. and Kenett, R.S. (2020), "Quality 4.0—the challenging future of quality engineering", *Quality Engineering*, Vol. 32 No. 4, pp. 614-626.

About the authors

Federico Barravecchia is an Assistant Professor at the Department of Management and Production Engineering at Politecnico di Torino. His main scientific interests currently concern the areas of service quality and quality engineering.

Luca Mastrogiacomo is an Associate Professor at the Department of Management and Production Engineering at Politecnico di Torino. His main scientific interests currently concern the areas of service quality and quality engineering.

Fiorenzo Franceschini is a Professor of Quality Engineering at Politecnico di Torino (Italy) – Department of Management and Production Engineering. He is author or co-author of nine books and more than 270 published papers in prestigious scientific journals and international conference proceedings. His current research interests are in the areas of quality engineering, performance measurements systems and service quality. Fiorenzo Franceschini is the corresponding author and can be contacted at: fiorenzo.franceschini@polito.it