Highlights

- Measuring quality is challenging due to preference heterogeneity among experts
- Online reviews offer a solution through quality features extracted from review text
- Using structural topic modeling we couple review text with numerical ratings
- An experimental application to airline passengers' reviews is demonstrated
- Quality features better predict variations in passenger preferences and competition

Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews

Nikolaos Korfiatis

(n.korfiatis@uea.ac.uk)

Norwich Business School and Centre for Competition Policy (CCP), University of East Anglia Norwich, United Kingdom.

Panagiotis Stamolampros[1]

(p.stamolampros@uea.ac.uk)

Norwich Business School, University of East Anglia, Norwich, United Kingdom.

Panos Kourouthanassis

(pkour@ionio.gr)

Department of Informatics, Ionian University, Corfu, Greece.

Vasileios Sagiadinos

(vsagiadinos@aueb.gr)

Athens University of Economics and Business, Athens, Greece.

[1] Corresponding author: Thomas Paine Study Centre, 0.15, Norwich Research Park, NR4 7TJ, Norwich, United Kingdom. Tel: +44(0) 1603 59 1041. E-mail: p.stamolampros@uea.ac.uk

Abstract

Service quality is a multi-dimensional construct which is not accurately measured by aspects deriving from numerical ratings and their associated weights. Extant literature in the expert and intelligent systems examines this issue by relying mainly on such constrained information sets. In this study, we utilize online reviews to show the information gains from the consideration of factors identified from topic modeling of unstructured data which provide a flexible extension to numerical scores to understand customer satisfaction and subsequently service quality. When numerical and textual features are combined, the explained variation in overall satisfaction improves significantly. We further present how such information can be of value for firms for corporate strategy decision-making when incorporated in an expert system that acts as a tool to perform market analysis and assess their competitive performance. We apply our methodology on airline passengers' online reviews using Structural Topic Models (STM), a recent probabilistic extension to Latent Dirichlet Allocation (LDA) that allows the incorporation of document level covariates. This innovation allows us to capture dominant drivers of satisfaction along with their dynamics and interdependencies. Results unveil the orthogonality of the low-cost aspect of airline competition when all other service quality dimensions are considered, thus explaining the success of low-cost carriers in the airline market.

Keywords: Electronic WOM, Unstructured Data, Service Quality, Correspondence Analysis, Structural Topic Model

1. Introduction

Quality measurement is an area of research where expert and intelligent systems have contributed substantially in the past (Lin, 2010). Product/ Service quality has been identified as a multidimensional construct which has significant challenges on its measurement due to the heterogeneity in consumer preferences for various quality dimensions (Hjorth-Andersen, 1984; Kamakura, Ratchford, & Agrawal, 1988). A generally accepted view in the literature is that *"…no single expert can possibly rate the quality of products unambiguously because he or she would not be able to come up with a composite scale that would appeal to all consumers…"* (Tellis & Johnson, 2007 p. 760). Due to their intangible, heterogeneous, and inseparable nature (Parasuraman, Zeithaml, & Berry, 1985), services exhibit a higher rate of complexity especially in cases where service failure occurs, and service quality is interdependent on different dimensions of a company's service offering.

The primary approach outlined in most of the studies is to either gauge expert input (Büyüközkan, Çifçi, & Güleryüz, 2011; W.-C. Chou & Cheng, 2012) or to sample and aggregate consumer responses in a structured form (de Oña, de Oña, & Calvo, 2012; Kuo & Liang, 2011). However, developing decision-making models utilizing either expert input or a sample of consumers is still criticized as inefficient (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). Online reviews tend to offer an alternative solution to alleviate this issue by acting as an important source of information for firms to gain a better understanding not only for their product/ service characteristics (e.g., sales performance), but also the conditions of the market they operate. Firms tend to leverage the information content of online reviews to improve their knowledge and understanding of their clientele preferences and the competitiveness of their service/ product offerings (Melo, Hernández-Maestro, & Muñoz-Gallego, 2017; Rodríguez-Díaz & Espino-Rodríguez, 2017). The primary driver for this trend is the abundance of data in unstructured form (e.g., open-ended responses in customer

complaints) which, if not extracted, analyzed and converted to actionable interventions, may not lead to improved decision making and, eventually, corporate performance (Xu, Frankwick, & Ramirez, 2016).

This study focuses on the problem of measuring service quality using unstructured data and converting them into managerial insights, thus paving the ground for developing more effective intelligent systems. Specifically, we apply a topic modeling variant that allows us to model the dependence between individual topics with response level variables (in terms of metadata) through the reliable extraction of service quality indicators. Moreover, we investigate how these information gains can further explain the variation of customers overall satisfaction. We also present how firms may make use of unstructured data in order to have a better understanding of their market positioning and comparison with their main competitors in the eyes of customers. Specifically, our study aims to answer the following set of questions:

a. Do service quality features extracted from unstructured data add predictability to customer satisfaction models that use numerical sub-indexes?
b. How do these features capture time-related aspects of service quality taking into account the dynamics and interdependencies between service aspects?

To answer these questions, we consider an experimental scenario using a large dataset of airline passengers' online reviews retrieved from TripAdvisor. Our rationale is guided from the following reasons. First, the nature of airline service offerings is interdependent in several factors related to different aspects of the service quality (e.g., onboard service, check-in service, delays, baggage handling, etc.). Second, this service offering is highly dependent on seasonal variations and, as such, a large sample containing several monthly observations would be able to demonstrate how service quality dimensions fluctuate across a time continuum. Third, considering the competition between airlines, as evinced from the

heterogeneity between their service offerings, this application domain provides an ideal application to demonstrate how valuable insights regarding the drivers of passengers' satisfaction (dissatisfaction) can be extracted at the firm (airline) as well as the market level (segment). From a managerial viewpoint, this is a problem of high economic significance since differentiation of service quality aspects provides a distinctive advantage that drives revenue growth and profitability (Tellis & Johnson, 2007).

Taking the above context in an experimental application, we demonstrate how the perceptual mapping of service quality indicators can be converted into managerial insights, namely: (a) the identification of latent quality dimensions, which are not captured by incumbent measurement instruments or (b) identification of opportunities for entering a particular market segment, which may contribute to increased satisfaction with an airline's quality offerings. The latter is of particular importance since not all service factors influence customer satisfaction equivalently with some of them tend to be closely related and others more orthogonal. In these cases, irrespective of the firm's overall performance, consumers will value those service factors as the key drivers for the formation of their post-service evaluations. While our study focuses on the airline industry we argue that our findings may also apply on other types of services (e.g., Hotel services, Hospital services, etc.) to identify which service factors lead to higher levels of customer satisfaction.

To this end, the remainder of the paper is organized as follows: A discussion about the information content of online reviews and its use in the context of service quality measurement along, the methodological assumptions behind structural topic models (STM), as well as a comparison with existing methods are provided in Section 2. The description of the data used, the steps involved in the STM method and the topic solutions are discussed in Section 3. In Section 4 we present the benefits of the inclusion of the topic solution in a regression framework for better explaining the variation of passengers' satisfaction, how the

identified topics capture temporal dynamics, and how this can be utilized in an expert system scenario that evaluates airline competition in terms of service quality offerings. Section 5 discusses theoretical and managerial implications, and our study concludes in Section 6 with a discussion of limitations and future research directions.

2. Background and Related work

2.1 Information Content of Online Reviews

Online reviews offer firms a low-cost vehicle for the acquisition and retention of customers. Managers may also take advantage of such mechanisms to identify their own as well as their competitors' strengths and weaknesses, while the power and the velocity of the dissemination of information through online communities also serve as a tool that necessitates the fulfillment of contractual agreements (Dellarocas, 2003). The primary motivation for customers to resort to online articulations from others is to reduce their risk and information asymmetries about a product or service (Yan, Xing, Zhang, & Ma, 2015). As such, their purchase decisions are significantly influenced by opinions shared by other customers in review aggregators.

The advent of review aggregators such as Yelp! and TripAdvisor has increased the availability of reviews to consumers and managers. Such aggregators enable decision-makers to assess the performance of product/ service offerings through a consolidated and comparative manner. Nevertheless, literature has identified several drawbacks inherent to their nature that may be summarized to the existence of response bias (N. Hu, Zhang, & Pavlou, 2009), self-selection bias (X. Li & Hitt, 2008), sequential bias (Eryarsoy & Piramuthu, 2014) or psychological biases (Stamolampros & Korfiatis, 2018). Even under the presence of such biases, given the long-established effect of eWOM on sales elasticities (See

for example Floyd et al., 2014), firms should incorporate this information into their business strategies.

A substantial body of literature has so far evaluated the characteristics of online reviews and their effect on consumer decision making. Extant studies in this stream explore, among others, consumer motivation to participate on eWOM (Tong, Wang, Tan, & Teo, 2013), review characteristics that affect the credibility of eWOM (Luo, Luo, Xu, Warkentin, & Sia, 2015) or their helpfulness (Korfiatis, García-Barriocanal, & Sánchez–Alonso, 2012; Krishnamoorthy, 2015; Ngo-Ye, Sinha, & Sen, 2017). However, the pertinent question on how firms can utilize the information content of online reviews is focused on either extracting information that reflects customer preferences in order to improve their products (Law, Gruss, & Abrahams, 2017; Zhang, Xu, & Wan, 2012) or understand the demographic or cultural characteristics of their clients (Stamolampros, Korfiatis, Kourouthanassis, & Symitsi, 2018).

Our study focuses on the latter stream of literature and in particular how the unstructured data of online reviews can be utilized by firms delivering services which are characterized by high complexity. Service quality is a multidimensional construct that captures individuals' perceptions pertaining their experiences with a service encounter (Atilgan, Akinci, & Aksoy, 2003; Brady & Cronin Jr, 2001) and particular service attributes (Bigne, Sanchez, & Sanchez, 2001; Fick & Brent Ritchie, 1991; Ladhari, 2009).

Although the relation between review scores and customer satisfaction (and its subsequent impact on sales) is widely explored in the context of eWOM literature (Radojevic, Stanisic, & Stanic, 2017; Viglia, Minazzi, & Buhalis, 2016), its information content is restrained by the aggregation of the dimensions of satisfaction in a generic score as well as the constraints imposed by the preselection of the individual categories designed from

platforms that allow users to score individual service quality dimensions. Consequently, as a service quality signal, online reviews carry some of the limitations found on traditional surveys where the multidimensionality of the quality (Tellis & Johnson, 2007) makes their use problematic because they are designed to reflect specific predefined dimensions leading to significant information loss.

Furthermore, since customer preferences are dynamic, established service quality measurement frameworks, such as SERVQUAL, fail to capture changing expectations and outcome beliefs (Buttle, 1996; Verhoef et al., 2009). Unstructured online reviews deal with those limitations since they reflect up-to-date preferences and provide an open forum for customers to pinpoint specific service offering dimensions that positively (or negatively) influenced the overall service experience. Therefore, under the reasonable assumption that reviewers discuss the primary drivers of their satisfaction (or dissatisfaction) in the reviews textual justification, this form of data can be used by firms to extract valuable information that is not gauged either by the overall score, or the predesigned individual rating categories.

2.2 Online Reviews and Service Quality in Travel Research

The relationship between service quality and satisfaction in the travel context has been a topic extensively discussed by scholars (Augustyn & Ho, 1998; Baker & Crompton, 2000; González, Comesaña, & Brea, 2007).

In the context of airline service encounters, service quality is typically assessed using performance metrics such as flight delays, customer complaints, mishandled baggage, airline safety records and consumer satisfaction indices that are captured through structured research instruments (See for example, Chang and Yeh, 2002; Keiningham et al., 2014; Suzuki et al., 2001). However, structured research instruments have two significant disadvantages. First, their investigation scope is limited to a predefined array of measurement items, which as

discussed before, leads to information loss. Second, response style biases may exist driven by the fact that some consumers tend to be orthogonal on a particular factor (e.g., price) and may hinder the importance of residual factors (De Jong, Steenkamp, Fox, & Baumgartner, 2008).

In a closely related stream of literature, that of hospitality services, analysis of online reviews has recently been regarded as an essential method for apprehending individuals' overall satisfaction. One stream of research is involved in determining the emotional polarity of individuals against a service encounter by employing sentiment analysis techniques (Park, Ok, & Chae, 2016; Ye, Zhang, & Law, 2009) and relating emotions with service performance (Phillips, Barnes, Zigan, & Schegg, 2017). Another stream of research follows an explanatory stance and probes for specific cues or aspects that elaborate on the service features or properties that contributed to the formation of travelers' perceptions following the service encounter (Marrese-Taylor, Velásquez, & Bravo-Marquez, 2014; Sotiriadis & Van Zyl, 2013) through content analysis and opinion mining.

These features reflect the perceived dimensions of service quality and are explicitly extracted or implicitly ascertained from the review comments. Unstructured responses elicited with free-form textboxes tend to contain additional information that helps overcome the abovementioned issues. Predominant opinion mining techniques include aspect-based mining with unsupervised learning, such as the Latent Dirichlet Allocation method (Blei et al., 2003) and Correlated Topic Models (Blei & Lafferty, 2006). Topic modeling has recently gained attention as a useful tool for analyzing customer provided information using unstructured textual data (Guo, Barnes, & Jia, 2017; Tirunillai & Tellis, 2014). In principle, topic models are unsupervised techniques which self-organize textual corpora in groups of topics. These topics are formed based on how specific groups of words appear together using both volume and context as inputs.

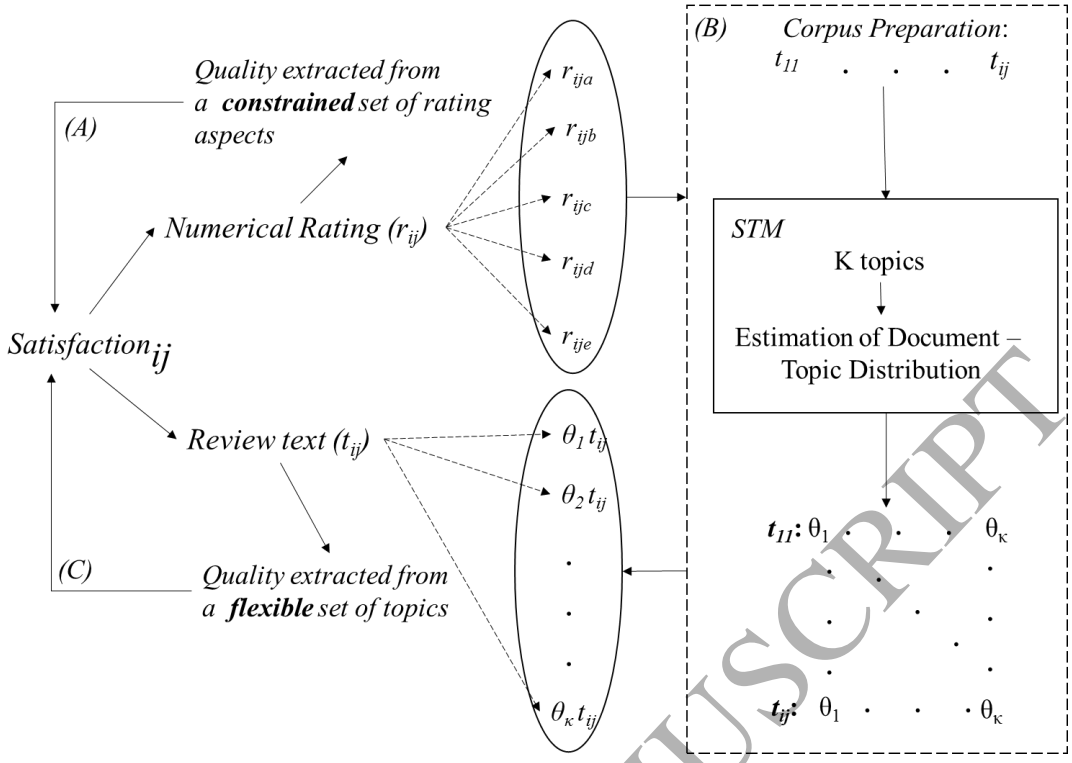

Figure 1: Identification and decomposition of the service aspects by incorporating structured data in the feature extraction process.

Figure 1 outlines the framework for the feature extraction process of service quality dimensions followed in this study. We consider three significant stages where structured data in numerical form (ratings of customer *i for airline j*) are combined with a corpus of unstructured data consisting of the textual justifications (t_{ij}) of these ratings (document format) in order to extract topics. In Stage B the numerical ratings are used as an input to derive the document-topic distribution, and a document topic matrix is generated where each topic representing a quality dimension is represented proportionally for each textual justification (Stage C). We outline the estimation process in the section that follows.

2.3 Extracting Service Quality dimensions from Structural Topic Models

In this study, we employ a probabilistic topic modeling method, coined as Structural Topic Models (STM), where topic coverage and word distribution are approximated through Bayesian inference (Roberts et al., 2016, 2014). This approach extends established

probabilistic topic models, such as Latent Dirichlet Allocation –LDA (Blei et al., 2003) or Correlated Topic Models-CTP (Blei & Lafferty, 2006), since documents (which in our case are the textual justifications that accompany review ratings) represent a mixture of latent topics and a distribution of words describes each topic. A significant advantage of STM, when compared to LDA and CTP, is that it allows the connection of arbitrary information, in the form of covariates, with the degree of association of a document with a topic (topic prevalence) as well as the degree of association of a word with a topic (content prevalence).

The primary principle behind topic modeling is the concept of exchangeability, which assumes that all authors are equally likely to write a document and the topics within this document are drawn from a prior distribution (Blei et al., 2003). In the case of STM, the probability of topic prevalence can be modeled with other covariates making it a more suitable methodology for online reviews allowing for example to connect the numerical rating of a review with the topics derived from the topic solution. Roberts et al. (2016) displayed that structural topic models outperform other approaches on topic modeling because of the possibility of including document level covariates (or metadata) as the primary factor driving the topic distribution. Applications of the STM model have gained momentum in the literature. Roberts et al. (2014) analyzed open-ended survey responses in political science; Light & Odden, (2017) focused on the dynamics of critics valuation on music albums; most recently, Kuhn (2018) elaborated on the discourse elements in accident investigation reports.

The STM process for review text is graphically depicted in Figure 2 using plate notation. The steps are described as follows:

Let us assume a corpus of R reviews with each review r indexed as $r_i \in (1, \ldots . R)$ containing $\boldsymbol{w}$ observed words within each review which are indexed as $\mathrm{n} \in (1, \ldots, N_{\mathrm{r}})$. Each

word is part of the general vocabulary of the review corpus with each term denoted by $v \in (1, \ldots, V)$ a word in the vocabulary is denoted by $w_{r,n}$. The primary input variable in a topic model is the number of topics k $\in$ (1, ... K) drawn from a distribution and this should be defined at the beginning of the estimation process. There are various ways to identify the number of topics in a corpus such as using the concepts of heldout likelihood and a combination of qualitative criteria which may require input by domain experts.

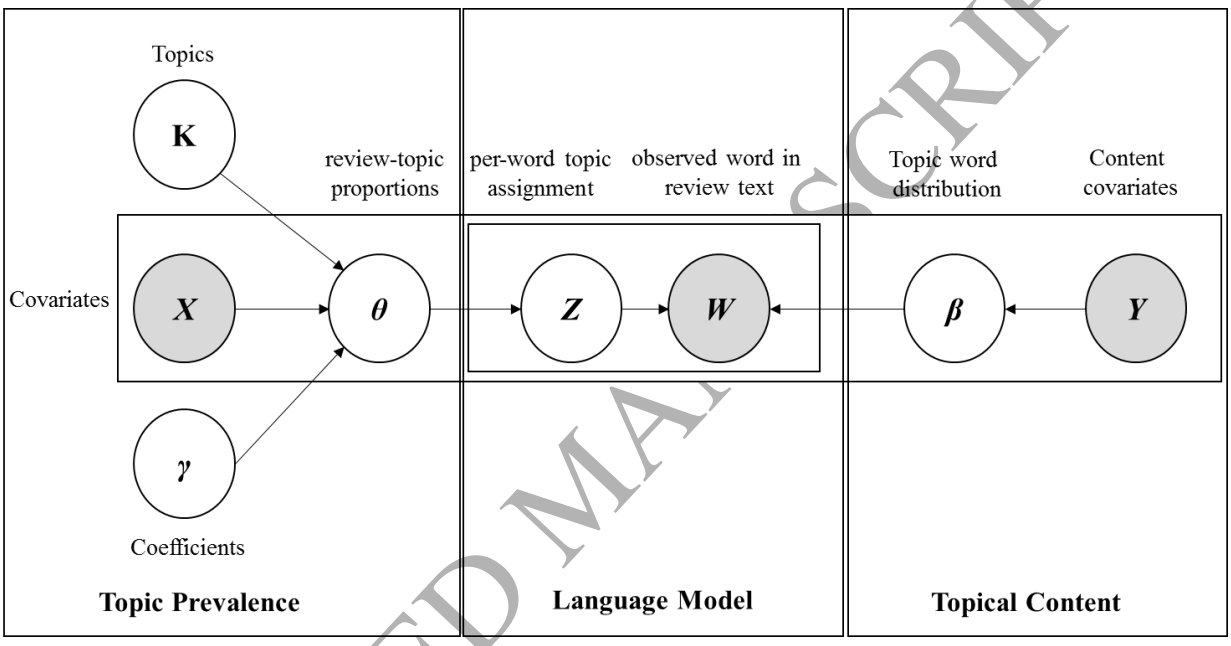

Figure 2: Structural topic model process using plate notation (Adopted by Roberts et al., 2016)

The distribution is affected by topic prevalence covariates which are specified in a $p \times 1$ vector X_r. When no topic prevalence covariates are defined, then the STM process works in the same way as Latent Dirichlet allocation by using Gibbs sampling to draw the topics from a Dirichlet distribution. X_r contains the review-based covariates that affect the dominance of a topic k_i for each review r_i, such as its rating score, reviewer's metadata, as well as the time the review was posted. The process runs in three steps as follows:

First, the review-level relation to each topic k is drawn from a logistic normal generalized linear model based on covariates and a set of priors as shown in Equation (1).

$$\vec{\theta}_{\gamma} \mid X_{r\gamma}, \sum \sim LogisticNormal(\mu = X_{r\gamma}, \sum) \quad (1)$$

, where γ represents a $p \times (K-1)$ matrix of coefficients drawn from a Normal distribution for each k $(k = 1, \dots, K-1)$ with the other *K-1* topics to provide bivariate dependence between topics. $\sum$ is a $(K-1) \times (K-1)$ covariance matrix.

Second, using the review-specific distribution over words initially attributed to each topic (k) by the log frequency distribution (m) of the vocabulary vector, a topic-specific deviation from the initial stage κ_k as well as a covariate for group deviation κ_g and an interaction term κ_i between them can be modeled as:

$$\beta_{r,k,v} \propto exp\left(m + \kappa_{k,v} + \kappa_{g,v} + \kappa_{i=(k,g_r,v)}\right) \quad (2)$$

Each of $m, \kappa_{k,v}, \kappa_{g,v,} \kappa_i$ are vectors (V-length) that contain one input per word in the vocabulary.

Finally, for each word $n \in (1, \dots, N_r)$ in a review text r_i the word-specific topic assignment $z_{r,n}$ can be modelled based on the review-specific distribution over the given finite set of topics as:

$$z_{r,n} \mid \vec{\theta}_{\gamma} \sim Multinomial(\vec{\theta}_{\gamma}) \quad (3)$$

The probability of an observed word to be attributed on this topic is given by:

$$w_{r,n} \mid z_{r,n}, \beta_{r,k} = Z_{r,n} \sim Multinomial(\beta_{r,k} = Z_{r,n}) \quad (4)$$

The model is then fit using a semiparametric estimation from a semi Expectation – Maximization algorithm (Blei & Lafferty, 2007; Wang & Blei, 2013) which upon convergence identifies the topic-specific proportions $\theta_{r,k|X}$ of a review using information from the vector of covariates provided in the initial stages of the estimation.

2.4 Comparison with other approaches

Several approaches for extracting service quality dimensions have been proposed in the literature. Table 1 provides an overview of the pros and cons of these approaches in terms of (a) data sources and input and (b) the set of predefined parameters that need to be defined.

Table 1: Overview of existing approaches in measuring service quality

Approach	Data Source	Pros	Cons	Indicative Studies
Survey Instruments (SERVQUAL)	Customer perceptions (structured questionnaire)	– Pre-defined set of dimensions of service quality – Known validity and reliability of the survey instrument. – Allows for comparable benchmarks when information is available.	– Constrained information set based on the constructs available in the survey instrument. – No possibility for comparison with other companies due to non-publicly available data. – Prone to sampling issues and response style biases	(Basfirinci & Mitra, 2015; Rajaguru, 2016; Suki, 2014)
Post-service feedback (company administered)	Customer Perceptions (structured questionnaire)	– High response rate based on single item constructs. – Immediate feedback which allows for the incorporation in performance dashboards	– Limited information set based on single-question attributes. – Prone to sampling issues and response style biases – Cross-sectional in nature	(Liou, Hsu, Yeh, & Lin, 2011; Liou, Tsai, Lin, & Tzeng, 2011)
Consumer Surveys	Customer Perceptions (Structured questionnaire administered by third-parties, e.g., Skytrax)	– Benchmark data available for comparison – Continuous information flow with repeated measurements (longitudinal) – Carefully selected panel	– Limited information set based on single-question attributes – Sample of the whole population and limits of panel composition	(Y.-W. Chang & Chang, 2010; Han, Ham, Yang, & Baek, 2012; Jiang & Zhang, 2016)
Expert input (TOPSIS)	Customer Perceptions (Sample) / Expert	– Small sample – Optimal configuration available	– Pre-defined parameters on the information set – Requires	(Awasthi, Chauhan, Omrani, & Panahi, 2011;

	Opinion (qualitative interpretation based on interviews)	– Rank-ordering of service quality dimensions	recruitment of experts (customers or service managers).	Büyüközkan et al., 2011; Sun, 2010)
Fuzzy input (FSQCA)	Customer Perceptions (Sample) / Expert Opinion (qualitative interpretation based on anchor classification scores)	– Small sample – Configuration options based on Fuzzy sets	– Difficulty of interpretation – Sampling bias – Sensitive to the membership function of the fuzzy set.	(S.-Y. Chou, Shen, Chiu, & Chou, 2016; Hsiao, Chen, Chang, & Chiu, 2016; Wu, Yeh, Woodside, & others, 2014)
Consumer reviews – Sentiment Analysis	Customer Reviews (unstructured text)	– Identification of positive and negative aspects of satisfaction using rating segmentation – Publicly available data – Continuous information flow	– Bag-of-words based approach using opinion dictionaries is prone to word length bias. – Self-selection bias (overly satisfied and overly dissatisfied consumers)	(Calheiros, Moro, & Rita, 2017; Geetha, Singha, & Sinha, 2017; Kim, Park, Yun, & Yun, 2017)
Consumer Reviews – LDA	Customer Reviews (unstructured text, e.g. online reviews)	– Topic models built from textual feedback – Flexible membership of each document to a topic	– No reliable way to estimate the predefined number of topics – No possibility to model the dependence between topics and review metadata (rating and service characteristics).	(Guo et al., 2017; Y.-H. Hu, Chen, & Lee, 2017; Rossetti, Stella, & Zanker, 2016)

Approaches for measuring service quality consider established instruments, however the multi-dimensionality of service quality as a construct leads to information loss along with other negative aspects, such as sampling and responder bias. Sentiment classification-based approaches are also reported in the literature; yet, this type of analysis uses a bag-of-words model where each word in the review is indexed using a unigram model. As such, it carries the limitation of not considering the order of words, but only associations between terms

based on the collocation of each word in the same document. Moreover, the reliability of scoring dictionaries for sentiment detection is established in the literature when considering single word frequencies, which can be hindered by context-specific limitations that are abundant in open-ended response text such as negations (e.g., '*not* great service').

Topic-model based approaches address most of the limitations of the approaches discussed so far, but they hinder two significant drawbacks. First, the definition of the number of topics can be arbitrary and is subject to the interpretability of the topic solution by the researchers. Second, the consideration that responses are documents with no exogenous covariates (as in the standard LDA and correlated LDA models) can hinder the reliability of the topic solution. In the case of online reviews, a critical meta-information - that of the numerical rating, can be of a significant factor for the document-topic as well as the word-topic distribution.

Structural topic models provide distinct advantages that address these issues. First, the interdependence of topics is explicitly modeled in terms of topic correlation. This allows for latent factors to be described by a combination of topics that are closer together in higher–order dimensions of the central latent concept. Second, STM allows for the topic vocabulary (the collection of words for a particular topic) to be associated with covariates in addition to the topic itself. Third, by explicitly including the covariate relationships in the model, we can include measurement uncertainty from the estimation of the latent topics into regression analysis, thus the number of topics can be accurately estimated. Fourth, based on open-ended responses there is significant information gain since predesigned aspects often come with missing ratings. Using the closest associated topic from the review text, we can impute missing information and increase the predictive accuracy. Based on the above, STM comprises a suitable method to extrapolate pertinent service quality dimensions from online reviews.

3. Data and Method

3.1 Dataset Description

Our data are sourced from TripAdvisor, the dominant online travel intermediary when it comes to booking hotels and evaluating hospitality experiences. More recently, TripAdvisor allows its members also to book flights and evaluate their experiences with a carrier. We collected all available airline passenger reviews (N= *557,208*) with information about flight date, name of airline, route (start and destination airport), cabin class (*first class*, *business class*, *premium economy*, and *economy class*) and reviewers' level of contribution to the platform (computed from the number of review posts). Passengers provide an overall score for their total experience (in an ordinal categorical scale from 1 to 5), which is accompanied by an individual rating for 8 specific service aspects of the flight namely: (a) Seat Comfort, (b) Customer Service, (c) Cleanliness, (d) Food and Beverage, (e) Legroom, (f) In-Flight Entertainment (g) Value for Money and (h) Check-in and Boarding.

We also computed the flight distance for each flight by geocoding the longitude and latitude of the departure and destination cities using the Haversine method. We initially collected 557,208 reviews written from 376,519 passengers. Approximately more than half of the reviews in our sample are written in English (254,424) with an approximate length of 560 characters for the review text. The average rating for all reviews in our sample is relatively good (*M=3.68, SD = 1.29*) and comparable to the rating score of the individual aspects. The ratings regarding the cleanliness/state of aircraft cabin (*M=3.94, SD=1.03*), customer service (*M=3.75, SD=1.34*) and check-in experience (*M=3.81, SD=1.25*) were above the average of the overall score. On the other hand, ratings for Seat Comfort (*M=3.46, SD=1.11*), Food and Beverages (*M=3.32, SD=1.27*) and Inflight Entertainment/Wi-Fi (*M=3.01, SD=1.47*) were below the overall rating, while Value for Money was very close to the Overall Score (*M=3.66, SD=1.23*).

3.2 Topic Prevalence Parameters

The primary variable that influences topic prevalence (the assignment of a review in a particular topic) in our STM model is the overall satisfaction of passengers with the service they received by an airline during a flight. Cabin class, flight distance and reviewers' (*passengers'*) level of contribution on TripAdvisor, which also acts as a proxy for passenger experience, are considered as additional controls that may influence topic prevalence.

3.3 Applying STM to Evaluate Quality Dimensions

To perform our analysis, we followed a three-step process: First, we applied established pre-processing techniques to extract the corpus used in the analysis; Second, we identified the number of topics that best describe the variability of the corpus; Third, we estimated how the topics change for different review ratings and additional controls. In the following parts we describe these steps in more details.

3.3.1 Text Preparation for Analysis

From the total pool of reviews, we selected only those written in English. In addition to the availability of stop-word lists, topic model approaches work best with English corpora due to the availability of part-of-speech taggers that assist with text preparation and estimation of marginal frequencies. The pre-processing workflow was based on previous studies (Guo et al., 2017; Tirunillai & Tellis, 2014) and involved the following steps: (i) word text tokenization, (ii) elimination of numbers and punctuation marks, (iii) exclusion of language stop-words (using the SMART list) as well as context-specific stopwords such as names of airlines, airports, and routes, and words with a length under a specific threshold (set to three characters), (iv) filtering the remaining words to keep only adverbs adjectives and nouns as these words have information about the product and product quality (Guo et al., 2017; Tirunillai & Tellis, 2014). This was done using part-of-speech (POS) tagging and the Python Natural Language Processing Toolkit (NLTK).

After pre-processing, we stemmed and lemmatized each word to derive groups with the same root form, excluding those that didn't appear in at least 1% of the initial corpus. This step reduced our final corpus to 184,502 online reviews which were used for the estimation of the topic solution.

3.3.2 Estimating the Number of Topics

Our analysis was performed in R (R Core Team, 2017) using the STM package. The basic STM considers the assignment of a document vector *(*with each document corresponding to each review *r)* containing a vocabulary of size *(V)* into *K* topics. As aforementioned, the underlying assumption of STM is that document level covariates influence the assignment of documents to topics (e.g., a review which is written by a business class passenger). Following Roberts et al., (2017), the number of topics was selected using three criteria: (i) Heldout likelihood (a measure on how the candidate number of topics is able to explain the overall variability in the review corpus) (ii) Semantic Coherence of the words to each topic and (iii) Exclusivity of topic words to the topic.

In order to generate a candidate number of topics for evaluation, we began with a stepwise estimation for an initial number of eight topics, (parameter *K* in the STM process as shown in Figure 2) similar to the number of rating categories that are offered by TripAdvisor, and evaluated the heldout likelihood until a maximum of 40 topics using an increment of two topics in each step. Then, the topic solutions with the highest heldout likelihood were evaluated against the FREX criterion (Roberts et al., 2016) as follows:

$$FREX_{k,v} = \left(\frac{\omega}{ECDF\left(\beta_{k,v} / \sum_{j=1}^{k} \beta_{j,v}\right)} + \frac{1-\omega}{ECDF\left(\beta_{k,v}\right)} \right)^{-1} \quad (5)$$

, where *k* is the *k*-th topic, v is the word under consideration, β is the word distribution for the *k*-th topic and ω a prior used to impose exclusivity (the default is 0.7). FREX is estimated as a weighted harmonic mean of a word's rank in terms of exclusivity and semantic coherence.

Semantic coherence, developed by Mimno et al. (2011), uses the frequency of co-occurrence of the most probable words in each topic of the topic solution, while exclusivity considers the mutual appearance of the most probable words in more than one topics.

Table 2: Labels, distribution and FREX score for the top 7 keywords in the topic solution.

#	Topic Label	Prop. (%)	Top 7 FREX words
1	Business Class	3.24	flat, lounge, business, class, bed, access, wine
2	Value for Money	6.84	good, food, value, overall, experience, money, airplane
3	Baggage Policy	3.02	bag, carry, charge, line, checked, fee, item
4	Low Cost	4.94	low, budget, price, cheap, cost, fare, carrier
5	Legroom (Critique)	2.02	room, space, enough, tall, foot, extra, amount
6	Delays	11.03	delay, hotel, hour, late, due, minute, connection
7	Staff (Praise)	10.11	friendly, helpful, clean, efficient, professional, courteous, smooth
8	Premium Economy	2.23	economy, premium, comfort, upgrade, difference, section, haul
9	Staff (Critique)	5.50	water, stewardess, toilet, poor, terrible, light, steward
10	Passenger Experience	6.01	best, many, domestic, world, past, travel, frequent
11	Frequent Flyer Status	1.91	flyer, member, traveller, group, point, aircraft
12	Mode of Travel	5.49	trip, return, direct, stop, home, round, non
13	Seating (Critique)	6.70	row, front, uncomfortable, seat, window, aisle, exit
14	Refund/Cancelation	5.42	phone, card, credit, email, call, agent, ticket
15	Food/In-flight entertainment	6.24	entertainment, movie, inflight, selection, screen, meal, average
16	Staff Assistance	5.22	child, holiday, nothing, much, special, cabin, crew
17	Legroom (Praise)	1.71	leg, lot, plenty, journey, second, extra, bit
18	Check-in	4.75	hand, luggage, check, queue, case, baggage, online
19	Airport Experience	4.98	free, snack, board, terminal, early, boarding, WIFI
20	Onboard Service	2.65	short, tea, full, usual, bit, etc, available

Using the FREX criterion, a 20-topic solution has the best relationship between heldout likelihood, semantic coherence, and exclusivity. We assigned labels to topics by recruiting two experts with airline customer service experience to help us evaluate each topic using a sample of the top 10 loading reviews and the top 7 FREX words. Both experts agreed that the selected topic solution had a high degree of coherence in terms of the top loading reviews and assigned mutually agreed labels. The estimated topic solution with the words having the highest FREX score and the assigned labels is provided in Table 2.

4. Analysis and Results

In this part, we summarize the results of the experimental application of STM in our corpus and evaluate our results in terms of (a) information gain from the inclusion of the topic solution to a regression analysis framework and (b) the temporal dynamics of the service factors identified from the topic solution. Our primary goal is to examine whether the features extracted from the STM process can add predictive ability to the overall satisfaction by minimizing information loss due to scale design as well as capture the temporal nature of service quality offerings.

4.1 Assessing the Information Gain from Unstructured Data

We begin our analysis using the information derived from review ratings. An ordinal logistic regression was performed, which is the suitable method based on the nature of our Likert-scaled response variable (Wooldridge, 2010). For covariates, we included the eight service categories that TripAdvisor allows passengers to vote along with the overall satisfaction of the travel experience controlling for cabin class, flight distance and reviewer's level of contribution to the platform. Our econometric specification is estimated as follows:

$$RevScore_i = \sum_{j=1}^{8} \beta_{1j} R_i^j + \sum_{c=1}^{3} \beta_{2c} Cabin_i^c + \beta_3 \mathrm{RevLevelContr}_i + \beta_4 \mathrm{FlightD}_i + \varepsilon_i \quad (6)$$

, where: $RevScore_i$ is the overall score given for this review. R_i^j is the individual rating score for the sub-rating category j of the review i, with j corresponding to one of the following eight categories: Seat Comfort, Customer Service, Cleanliness, Food & Beverages, Legroom, Inflight Entertainment, Value for Money, Check-in /Boarding. $Cabin_i^c$ is the cabin class of the passenger for the specific flight he wrote the review with c representing one of the following: Premium Economy, Business Class, First Class (Economy is used as a baseline). $RevLevelContr_i$ provides the number of reviews that the passenger who wrote the review i

has written and is also used as a proxy of passenger experience. $FlightD_i$ is the flight distance for the flight that corresponds to the review *i*.

To illustrate the effect of the incorporation of the estimated topic proportions to the explanatory power of the specified model, we estimated additional variations of the baseline model by incrementally adding one topic variables at a time. To conserve space and since studying all the topics found from the topic solution is beyond the scope of this analysis, we select a more parsimonious model including only the three topics with the biggest (positive or negative) marginal effects in their prevalence (Topics: 6, 14, and 7). As such, we considered three additional models: Model 2 with Topic #6 (issues with delays), Model 3 with Topic #14 (Refunds / Cancelation issues) and Model 4 with Topic #7 (Staff Praise). For each review the estimated θ parameter for the topic from the topic solution (Figure 1) was included as a covariate in the form of a dichotomous variable, denoting whether this topic was the dominant topic for this review or not. The results are shown in Table 3 along with the likelihood ratio tests that are provided to evaluate the increased information gain (in the form of χ^2 difference tests). Considering that not all passengers rate all the eight aspects of the service provided during the flight, the final number of observations was truncated to N = 173,481.

Table 3: Baseline and additional models for assessing the predictive relevance of topic proportions on the overall rating

	(1)	(2)	(3)	(4)
Rating Categories				
Seat Comfort	0.460***	0.483***	0.506***	0.449***
	(0.009)	(0.009)	(0.009)	(0.009)
Customer Service	0.905***	0.886***	0.800***	0.720***
	(0.007)	(0.007)	(0.007)	(0.007)
Cleanliness	0.200***	0.217***	0.242***	0.222***

	(0.008)	(0.008)	(0.008)	(0.008)
Food & Beverages	0.334*** (0.006)	0.346*** (0.006)	0.372*** (0.006)	0.340*** (0.006)
Legroom	0.184*** (0.008)	0.187*** (0.008)	0.182*** (0.008)	0.186*** (0.008)
In-flight Entertainment	0.175*** (0.005)	0.169*** (0.005)	0.177*** (0.005)	0.183*** (0.005)
Value for Money	0.756*** (0.007)	0.751*** (0.007)	0.737*** (0.007)	0.729*** (0.007)
Check-in /Boarding	0.507*** (0.006)	0.434*** (0.006)	0.389*** (0.006)	0.356*** (0.006)
Control Variables				
Premium Economy	-0.085** (0.029)	-0.230*** (0.029)	-0.295*** (0.029)	-0.035 (0.030)
Business Class	-0.433*** (0.018)	-0.553*** (0.018)	-0.644*** (0.019)	-0.336*** (0.019)
First Class	-0.378*** (0.033)	-0.444*** (0.034)	-0.390*** (0.034)	0.009 (0.035)
Reviewer Level of Contribution	0.046*** (0.003)	0.040*** (0.003)	0.023*** (0.003)	0.046*** (0.003)
Flight Distance.	-0.123*** (0.006)	-0.172*** (0.006)	-0.222*** (0.006)	-0.134*** (0.006)
Topic Membership (θ)				
Delays (Topic: 6)		-4.413*** (0.055)	-4.655*** (0.056)	-3.667*** (0.056)
Refund/Cancelation (Topic: 14)			-9.122*** (0.111)	-7.576*** (0.110)
Staff Praise (Topic: 7)				10.042*** (0.100)
AIC	270040.4	263413.5	255508.6	243023.7
LL	-135,003	-131,689	-127,735	-121,492
χ^2		6628.9***	7906.8***	12486.9***

Note: N=173,481 observations for all models after case wise deletion for those reviews where no service category ratings were available. Bootstrap standard errors in parenthesis. $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Using this specification and considering the nature of our response variable (review score) an ordered logistic regression model was estimated. This was assigned as the baseline model (Model 1 in Table 3) to examine how the information about the topic distribution (extracted from the STM process) could increase the predictability of the review score. We evaluated the prevalence of each topic for the continuum of the values of the review rating (Low rating: 1 to High Rating: 5) to decide which topics will augment the baseline model.

Figure 3 displays the marginal effects of the overall rating to the topics discussed. The longer the distance of the topic from the dotted line (that depicts the zero effect), the more prominent the change in the proportion of the topic as part of the overall corpus. Customer service is the critical factor that is highly connected with increased satisfaction since Staff praise (topic #7) has the most significant change per rating score increase (a 4% increase per unit score). This is in line with the previous studies regarding drivers of service satisfaction (Anderson, Pearo, & Widener, 2008) or loyalty (Vlachos & Lin, 2014) in airlines where interaction with personnel is described as a critical factor. The second more significant positive effect per unit increase of rating score is related to Topic #2 that corresponds to Value for Money. The stronger negative effects are observed for the topics related to delays (Topic #6), staff critique (Topic #9) as well as the refund/cancellation topic (Topic #14) which refers to service recovery failures.

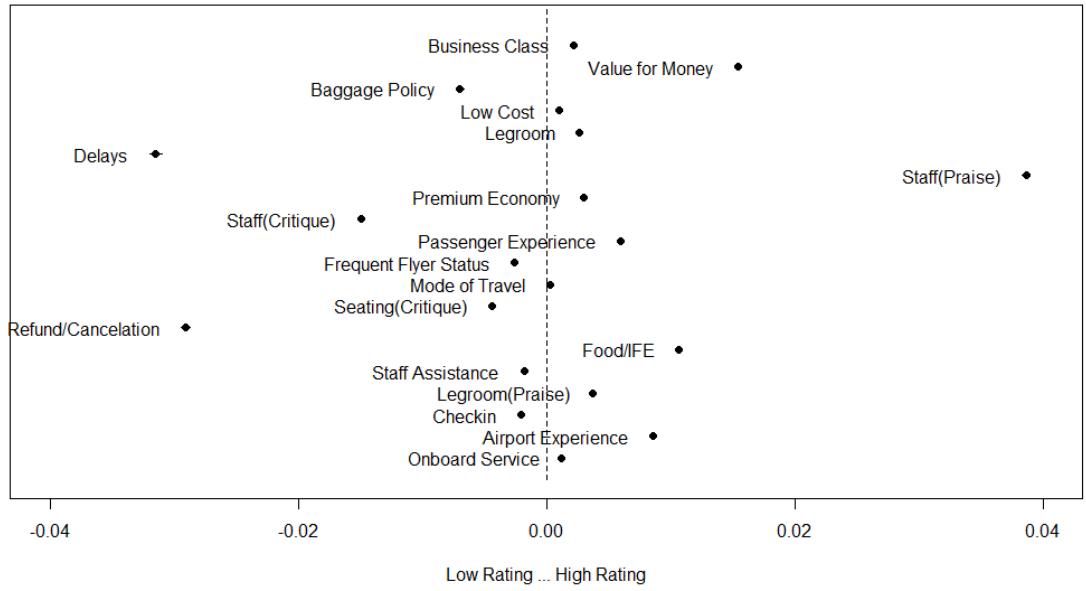

Figure 3: Marginal effects in the change of the expected proportions of topic prevalence based on low and high review score. The dotted line represents the zero effect. The interpretation of the graph for example in the case of the topic Staff (Praise) is that per unit increase in overall satisfaction there is an increase of 4% in the discussion about the specific topic.

Comparing the marginal effects obtained from our STM solution (Figure 3), with the size of the covariates of the rating aspects from the baseline regression (Model 1), we found an agreement regarding the importance of customer service and value for money service factors. As the distance from the dashed axis increases (decreases), the probability of specific topics become more dominant. For instance, Delays and Carriers' responses to service failures (refunds/cancellations) are critical factors that lead to customer dissatisfaction when occurring and if not appropriately addressed. On the other hand, staff praise by consumers is the primary driver of high ratings as the topic is becoming dominant for extremely satisfied customers. However, there are several dimensions, which are not directly captured from the current review interface in TripAdvisor and thus not measured. This indicates the value of the topic modeling approach on extracting additional quality dimensions.

4.2 Temporal Variations in Service Quality Dimensions

In addition to the information gain analysis demonstrated in a regression framework, we turn our attention to the second limitation that cross-sectional service quality surveys exhibit, namely capturing the temporal variation of service levels, which may be over or under-represented in a sample at the time of data collection. We do so by considering the effect of seasonal variations on the main topics in our topic solution and estimating the effect of review time on the topic prevalence. As such we consider the topic membership function *(θ)* to fluctuate by time as: $Y_k = \{Y_{kt} : t \in T\}$ where T is the bandwidth of the period (in terms of days) that is available in our dataset is (T=584). To test the impact of time on the prevalence of the k-th topic prevalence) we consider the following model:

$$Y_{kt} = \begin{bmatrix} Prevalence_1 \\ \vdots \\ Prevalence_k \end{bmatrix} = \alpha_k + \beta_{ik} \begin{bmatrix} \text{rating}_{1t} \\ \vdots \\ \text{rating}_{iT} \end{bmatrix} + \gamma_m A \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} + \delta_c C \begin{bmatrix} C_1 \\ \vdots \\ C_i \end{bmatrix} + d_{1t} \quad (7)$$

Where β_{ik} indexes the influence of the time-related rating covariate on the topic prevalence, γ_m and δ_c indexes the airline and cabin class respectively and d is a smoothing covariate to account for time effects. For each topic we estimated the influence of these covariates on the topic distribution and plotted the marginal effects for the bandwidth of 574 days that were available in our dataset.

The plot in Figure 4 shows the temporal dynamics during periods of high load for airlines (summer period and Christmas). This is in line with the literature that observes similar dynamics in online ratings (X. Li & Hitt, 2008). However, in our case temporal dynamics exhibited upon the dominance of some topics in the corpus can be explained by specific conditions (in the case of airlines – high load) where some topics (e.g., Delays) become more prevalent than others. Thus, the analysis in that level better captures the current service provisioning conditions than the variations exhibited on online ratings. That result signifies that customer preferences change over time (Buttle, 1996; Verhoef et al., 2009) and as such firms which want to maximize the insights offered from online reviews should develop expert systems to monitor not only their service offerings but the overall market conditions. We discuss the application of such an expert system for airlines in the section that follows.

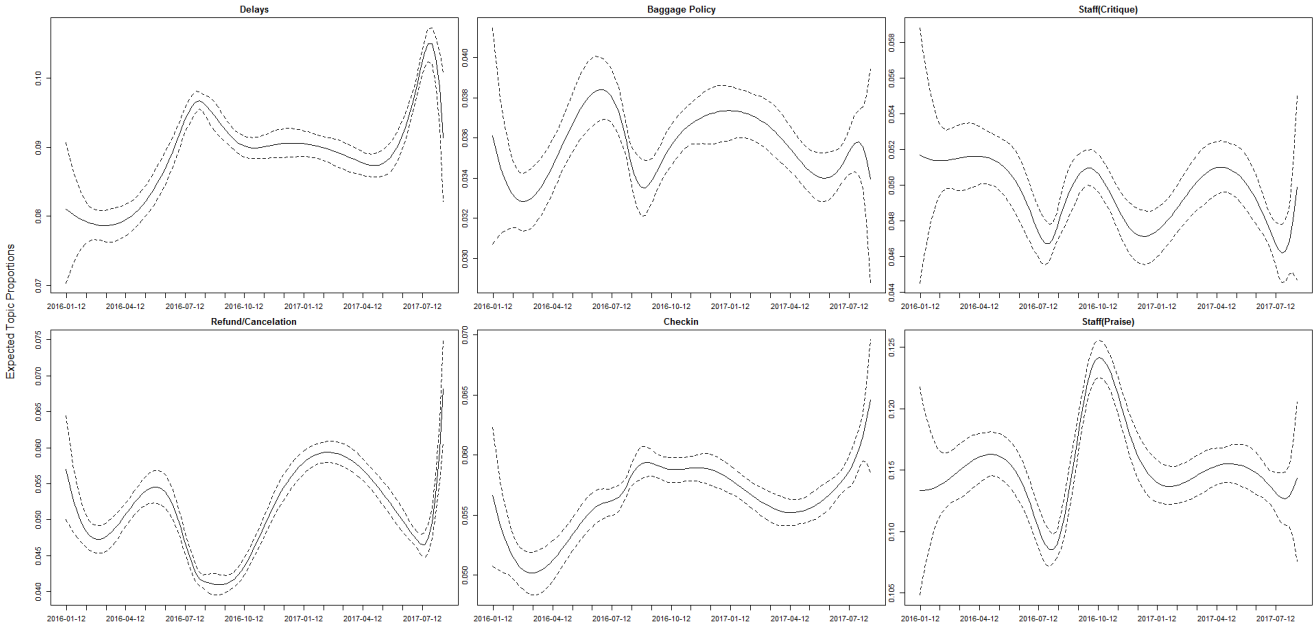

Figure 4: Seasonality patterns in the growth and decline of topics in our corpus (Bandwidth: 574 days)

4.3 Mapping Airline Service Quality Competition

In order to demonstrate how airlines can exploit the textual content of online reviews for evaluating their competitive performance, a perceptual mapping using correspondence analysis for service quality dimensions extracted from the STM topic solution was performed by selecting the seven most prevalent topics from our corpus (Table 2). Service quality dimensions are multifaceted, and the goal of each airline is to be more or sometimes less associated with them based on their strategy. For example, airlines that build their model on pricing (Low cost) such as Ryanair and EasyJet, would like to evaluate their position towards other quality factors as well as their relative position with other airlines, especially considering the introduction of low-cost subsidiaries by legacy carriers.

We consider a bivariate frequency table *F* having *n* rows indexing an airline (*i*) such as *i*=1,…,*n* and *m* columns indexing the service quality dimensions (*j*) extracted from our STM solution such as *j=1, ...,m* conditioning $n \gg m$ (Suggesting that the number of airlines to be evaluated is significantly larger than the number of passengers). Following Greenacre (2017) we consider the total number of observations indexed in the table as $n_c = \sum_{i=1}^{n}\sum_{j=1}^{j} F_{ij}$. The vectors of marginal probabilities for airline *i* to belong to the service quality dimension w_m and for the service quality dimension to load for that particular airline w_n are given by $w_m = \frac{1}{n_c} C d_1$ and $w_n = \frac{1}{n_c} d^T{}_1 C$ where d_1 is the column vector of the maximum probability for each dimension. In estimating the deviation from independence between service quality dimensions for each airline we assume that all service quality dimensions are orthogonal, suggesting that each airline has an equal probability to exhibit a service quality dimension. As such the deviation from independence *(M)* can be estimated in a matrix form as $M = \frac{1}{n_c} C - w_m w_n$. Using generalized value decomposition (VD), *M* can be decomposed and factor scores for the airlines (*n*) and service quality dimensions (*m*) can be

extracted as $F_n = W_n U \sum VD$ and $= W_m U \sum$VD. With W_n and W_m representing the diagonals of w_n and w_m respectively.

Table 4: Chi-squared (χ^2) *decomposition of the factors obtained from the model solution. For both segments row- and column-based Benzecri Root Mean Squared Error (RMSE) < 0.001*

Factor								Proportion	Cum. Proportion
Segment 1 – Intra-European (total χ^2 = 0.846)									
Factor 1	0.513							0.607	60.7
Factor 2		0.189						0.223	83.0
Factor 3			0.075					0.089	91.9
Factor 4				0.029				0.035	95.4
Factor 5					0.021			0.024	97.8
Factor 6						0.015		0.018	99.6
Factor 7							0.004	0.004	100
Segment 2 – Intercontinental (χ^2 = 0.592)									
Factor 1	0.407							0.688	68.8
Factor 2		0.088						0.148	83.5
Factor 3			0.046					0.077	91.3
Factor 4				0.034				0.058	97.1
Factor 5					0.010			0.016	98.7
Factor 6						0.004		0.007	99.4
Factor 7							0.004	0.006	100

We extracted two subsets from our topic solution considering the different nature of service offerings for airlines and the associated aircrafts used (narrow-body vs. wide-body). The first is for flights within Europe (Intra-European segment) and the airlines that are active in this market while the second subset considers those airlines which compete on intercontinental routes (Intercontinental segment). For each passenger review in both subsets,

we estimated the topic membership *(θ)* extracted from the topic solution and aggregated it on airline level.

As can be obtained from Table 4, for both subsets more than 80% of the variation in service quality dimensions can be explained by two dominant components. Considering the first subset (Europe) correspondence analysis resulted in two principal components who jointly explained 83% of the variance in the topic prevalence which were used for the visualization if of the service components in Figure 5. Several interesting observations can be extracted regarding the positioning of airlines to dimensions of service quality from topic solution. For example, the dimension of low cost is relatively orthogonal to all the other dimensions of focus. In that aspect, we can see that both EasyJet and Ryanair have the same loading for this dimension. However, EasyJet and other low-cost carriers (e.g., Transavia) are closer to the dimension of Staff praise which was the most dominant factor for high ratings given the results reported in the previous section. Considering recent events in the airline market (e.g., Ryanair's cancelations due to pilot shortage and Air Berlin's Bankruptcy) it is no surprise that Ryanair is closer to the delay aspects, and Air Berlin was the top loading airline for issues related with refund and cancellation. An interesting observation is that despite our dataset contains observations until July 2017 the aggregation of topic prevalence reveals some predictability of future events. From the perspective of non-low-cost airlines, we can see that Aegean airlines has the best combination of staff praise and price levels which confirms its status as the best regional airline in Europe according to ERA Awards – Skytrax[a].

Traditional carriers, such as British Airways, score higher for staff critique which is also expected considering the strike actions and the cabin crew disputes with the airline[b]. On the other hand, Lufthansa scores highly for value for money and legroom praise and manages to distance itself from all the negative aspects of quality dimensions extracted from our topic

solution. This comes as no surprise as in the past three years the company has reported increased profitability[c].

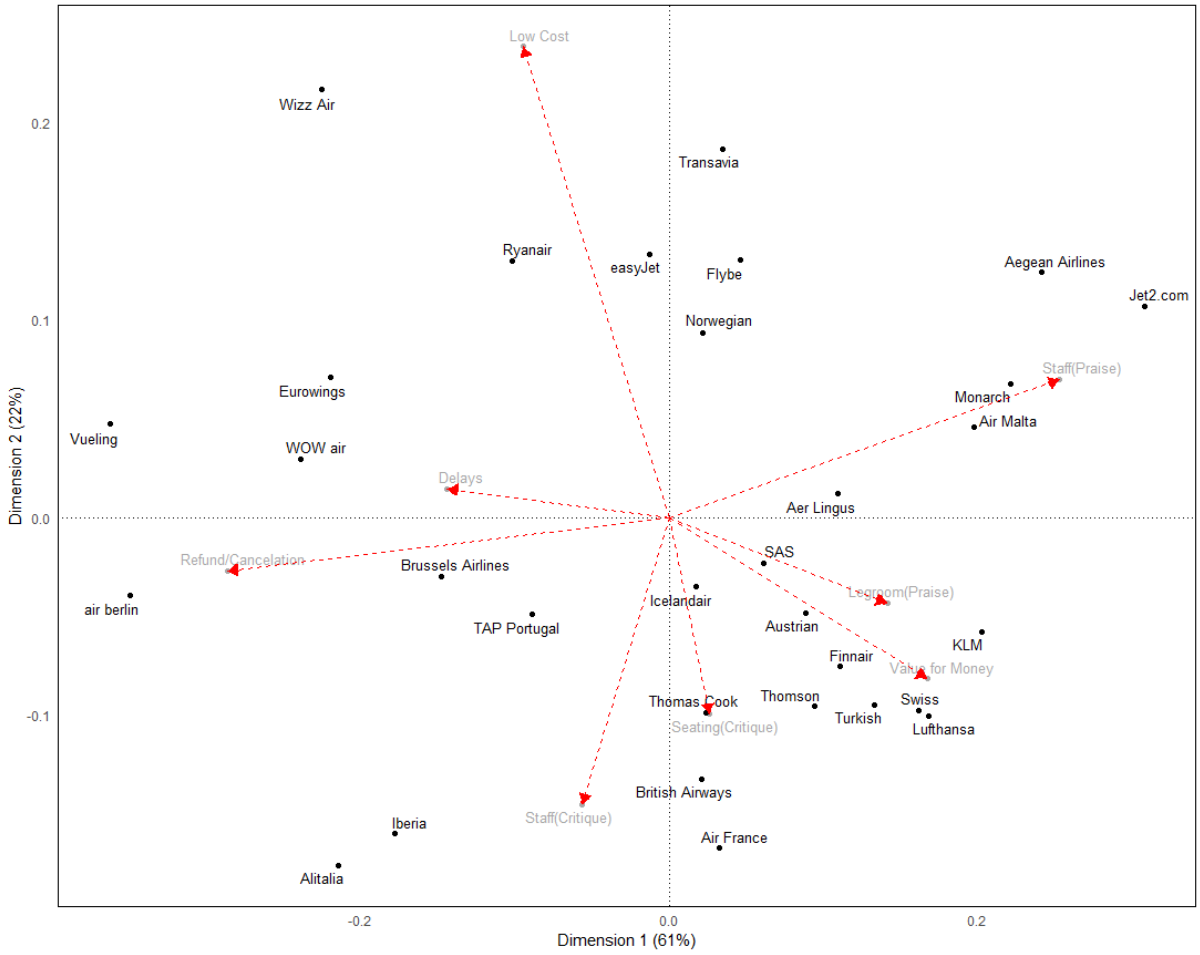

Figure 5: Competition of service quality dimensions for Intra-European segment.

For the second subset considering the service offerings in intercontinental routes (Figure 6), we similarly observe the orthogonality of the low-cost dimension with Norwegian currently being the winner on that segment while Air Lingus and Jet Airways being more praised for their staff. The best combination of value for money and staff praise is given to Singapore Airlines and Emirates closely followed by Qatar Airways, Quantas, and Cathay

Pacific. This is expected considering that all these airlines are currently ranked as 5-star and 4-star airlines by Skytrax[d].

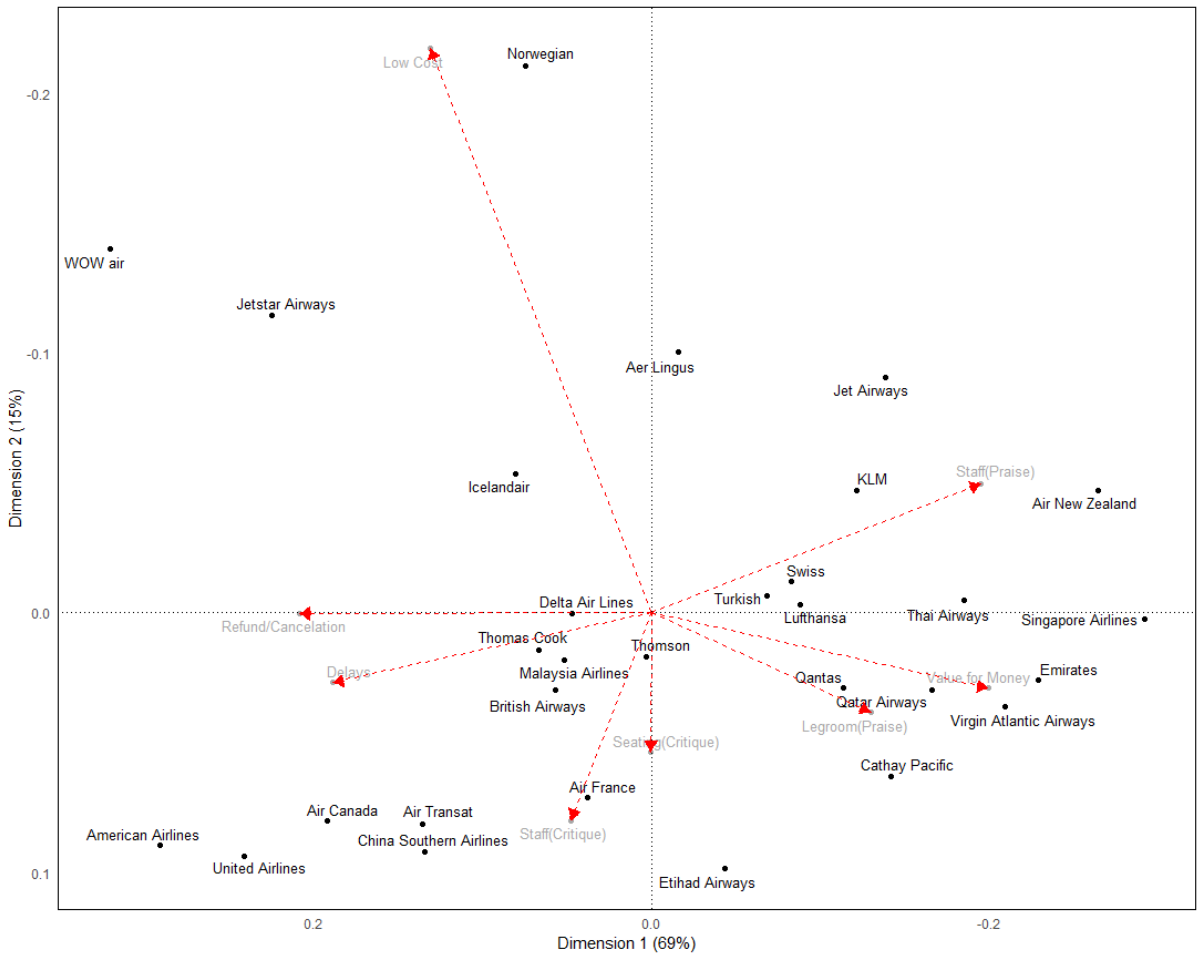

Figure 6: Competition in service quality dimensions for the intercontinental segment.

5. Discussion and Implications

5.1 Implications for Theory

Compared with extant approaches in the literature when applied on service quality measurement our study has significant advantages. First, rank-based approaches, such as TOPSIS, typically rely on a limited pool of experts for evaluating a set of pre-defined criteria (Liu, Bi, & Fan, 2017). Our approach relaxes the requirement of pre-defining the service quality dimensions as well as expert recruitment and directs its attention on open-ended

responses from customers. Second, by modeling the dependence between the extracted dimensions, our model shows significant qualitative performance and interpretability when compared with factual information. As such, we show that the inclusion of topic models increases significantly the variation of the customer satisfaction explained in our model, while at the same time it allows us to capture factors that are not measured by predefined sub-indexes. In that way we respond to several calls by researchers in the expert and intelligent systems literature (J. Li, Xu, Tang, Wang, & Li, 2018; Tsui, Wang, Cai, Cheung, & Lee, 2014) related to the value of unstructured data, by establishing information gains that can be achieved when textual features from open-ended responses are incorporated with structured information from measurement scales.

Our study also provides a methodological contribution by demonstrating the suitability of a particular class of topic models, namely the Structural Topic Model – STM (Roberts, Stewart, & Airoldi, 2016) for mapping service quality dimensions, thus allowing a more effective measurement of service quality. This class of topic models, when compared to established probabilistic topic models, and in particular the widely used Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), offers distinctive advantages stemming mainly from the inclusion of covariates of interest into the prior distributions of the document-topic proportion and topic-word distribution. This text mining method also presents increased predictive power and qualitative interpretability in determining the dominant topics of online reviews compared to extant text analytics methods based on the bag-of-words approach (e.g., dictionary scoring).

5.2 Implications for Practice

The study has valuable implications for managerial practice. Specifically, it enables airline managers to pinpoint salient dimensions of service quality from user-generated data that influence customer satisfaction and relate these dimensions to ascertain the competitive

positioning of each airline company within the industry. Interestingly, our findings are in line with the literature discussing the importance of Big Data as a mean to capture forward-looking information and evolving customer preferences (Lambrecht & Tucker, 2017) and may inform the identification of different passenger clusters based on the salient service quality dimensions (e.g., price-sensitive passengers and convenience passengers). Since the STM method may be employed dynamically on changing customer perceptions over time, this study paves the ground for the specification of expert systems that analyze such unstructured data to reshape a company's strategy.

In the context of the airline industry, managers may identify up-to-date drivers of customer satisfaction (or dissatisfaction) and their dynamics across time and position their service offerings accordingly. Our study identifies such clusters of airline companies that share common perceptions on specific service quality properties (i.e., cost, cabin comfort, staff behavior, etc.) and reveal the importance that passengers attribute to these dimensions. For example, our study confirms the importance of such dimensions as cost and comfort as proxies of customer satisfaction and, ultimately, loyalty (Koklic, Kukar-Kinney, & Vegelj, 2017; Lee & Yu, 2018). This is more apparent with legacy carriers who operate a low-cost operation under another brand but would like to keep or change some distinctive characteristics in quality offering that stem from their own brand (e.g., KLM and Transavia, British Airways and Vueling, etc.).

An additional managerial implication concerns review aggregators and in particular those aggregators that provide specialized review interfaces to measure specific dimensions of the overall rating. The design of review interfaces is a challenging task and requires informed selection of the pertinent dimensions. Such dimensions may vary depending on the culture or past experiences of passengers. Therefore, an intelligent system that dynamically probes for such dimensions may contribute to the design of personalized review interfaces

based on inherent passenger features. Interestingly, our findings from the topic modeling method application unraveled essential dimensions, such as delays and baggage policy, which are not included in the rating dimensions specified by contemporary review aggregators, such as TripAdvisor.

6. Limitations and Future Research

While our study has provided an innovative way to capture service quality indicators from unstructured data and showcased their predictive ability on customer satisfaction and airline competition, it has a set of limitations which are pertinent to the use of online reviews as we outlined in Section 2.4. Indeed, response biases based on consumer demographics and culture can alter the textual content as well as the rating of online reviews (Korfiatis & Poulos, 2013; Stamolampros et al., 2018) and as such more control variables are needed in the estimation of document-topic and word-topic distributions.

Nevertheless, several promising avenues for future research can be initiated. First, the study can be extended to accommodate the effect of temporal shifts in the service quality offering under periods of high load (e.g., summer months). Airlines tend to function with different schedules (winter and summer) where additional staff is brought in to increase capacity (either in the form of ground crew or additional aircraft leases), and this may influence the service offering (something that we are not able to control). Second, the identified service quality factors may be further analyzed by extending the unit of analysis to other stakeholders, such as airports (Kuo & Liang, 2011). While the service level in our study could be proxied by the cabin class, additional information about the respondents, such as demographics, could add further predictive validity on our study. Along with this line, our investigation lens could be extended through other forms of online information that have

gained momentum in the literature such as for example employee reviews (Symitsi, Stamolampros, & Daskalakis, 2018).

Author Contribution Statement

Nikolaos Korfiatis as the lead author of the paper was responsible for the conceptualization of the study, the identification of the relevant literature analysis of the data and write-up of the various stages of the manuscript.

Panagiotis Stamolampros was jointly responsible for the conceptualization of the study and performed the data collection as well as write-up of the various stages of the manuscript.

Panos Kourouthanassis contributed to the identification of the literature, the positioning of the study with the relevant literature and the conceptualization of practical and theoretical implications.

Vasileios Sagiadinos contributed to the positioning of the study with the related work and identification of the relevant literature.

References

Anderson, S., Pearo, L. K., & Widener, S. K. (2008). Drivers of Service Satisfaction: Linking Customer Satisfaction to the Service Concept and Customer Characteristics. *Journal of Service Research*, *10*(4), 365–381.

Atilgan, E., Akinci, S., & Aksoy, S. (2003). Mapping service quality in the tourism industry. *Managing Service Quality: An International Journal*, *13*(5), 412–422.

Augustyn, M., & Ho, S. K. (1998). Service quality and tourism. *Journal of Travel Research*, *37*(1), 71–75.

Awasthi, A., Chauhan, S. S., Omrani, H., & Panahi, A. (2011). A hybrid approach based on SERVQUAL and fuzzy TOPSIS for evaluating transportation service quality. *Computers & Industrial Engineering*, *61*(3), 637–646.

Baker, D. A., & Crompton, J. L. (2000). Quality, satisfaction and behavioral intentions. *Annals of Tourism Research*, *27*(3), 785–804.

Basfirinci, C., & Mitra, A. (2015). A cross cultural investigation of airlines service quality through integration of Servqual and the Kano model. *Journal of Air Transport Management*, *42*, 239–248.

Bigne, J. E., Sanchez, M. I., & Sanchez, J. (2001). Tourism image, evaluation variables and after purchase behaviour: inter-relationship. *Tourism Management*, *22*(6), 607–616.

Blei, D. M., & Lafferty, J. D. (2006). Correlated Topic Models. In *In Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). MIT Press.

Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Brady, M. K., & Cronin Jr, J. J. (2001). Some new thoughts on conceptualizing perceived service quality: a hierarchical approach. *Journal of Marketing*, *65*(3), 34–49.

Buttle, F. (1996). SERVQUAL: review, critique, research agenda. *European Journal of Marketing*, *30*(1), 8–32.

Büyüközkan, G., Çifçi, G., & Güleryüz, S. (2011). Strategic analysis of healthcare service quality using fuzzy AHP methodology. *Expert Systems with Applications*, *38*(8), 9407–9424.

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

Chang, Y.-H., & Yeh, C.-H. (2002). A survey analysis of service quality for domestic airlines. *European Journal of Operational Research*, *139*(1), 166–177.

Chang, Y.-W., & Chang, Y.-H. (2010). Does service recovery affect satisfaction and customer loyalty? An empirical study of airline services. *Journal of Air Transport Management*, *16*(6), 340–342.

Chou, S.-Y., Shen, G. C., Chiu, H.-C., & Chou, Y.-T. (2016). Multichannel service providers' strategy: Understanding customers' switching and free-riding behavior. *Journal of Business Research*, *69*(6), 2226–2232.

Chou, W.-C., & Cheng, Y.-P. (2012). A hybrid fuzzy MCDM approach for evaluating website quality of professional accounting firms. *Expert Systems with Applications*, *39*(3), 2783–2793.

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research*, *45*(1), 104–115.

de Oña, J., de Oña, R., & Calvo, F. J. (2012). A classification tree approach to identify key factors of transit service quality. *Expert Systems with Applications*, *39*(12), 11164–11171.

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–1424.

Eryarsoy, E., & Piramuthu, S. (2014). Experimental evaluation of sequential bias in online customer reviews. *Information & Management*, *51*(8), 964–971.

Fick, G. R., & Brent Ritchie, J. (1991). Measuring service quality in the travel and tourism industry. *Journal of Travel Research*, *30*(2), 2–9.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How Online Product Reviews Affect Retail Sales: A Meta-analysis. *Journal of Retailing*, *90*(2), 217–232.

Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. *Tourism Management*, *61*, 43–54.

González, M. E. A., Comesaña, L. R., & Brea, J. A. F. (2007). Assessing tourist behavioral intentions through perceived service quality and customer satisfaction. *Journal of Business Research*, *60*(2), 153–160.

Greenacre, M. (2017). *Correspondence analysis in practice*. CRC press.

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

Han, S., Ham, S. S., Yang, I., & Baek, S. (2012). Passengers' perceptions of airline lounges: Importance of attributes that determine usage and service quality measurement. *Tourism Management*, *33*(5), 1103–1111.

Hjorth-Andersen, C. (1984). The concept of quality and the efficiency of markets for consumer products. *Journal of Consumer Research*, *11*(2), 708–718.

Hsiao, Y.-H., Chen, L.-F., Chang, C.-C., & Chiu, F.-H. (2016). Configurational path to customer satisfaction and stickiness for a restaurant chain using fuzzy set qualitative comparative analysis. *Journal of Business Research*, *69*(8), 2939–2949.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144–147.

Hu, Y.-H., Chen, K., & Lee, P.-J. (2017). The effect of user-controllable filters on the prediction of online hotel reviews. *Information & Management*, *54*(6), 728–744.

Jiang, H., & Zhang, Y. (2016). An investigation of service quality, customer satisfaction and loyalty in China's airline market. *Journal of Air Transport Management*, *57*, 80–88.

Kamakura, W. A., Ratchford, B. T., & Agrawal, J. (1988). Measuring market efficiency and welfare loss. *Journal of Consumer Research*, *15*(3), 289–302.

Keiningham, T. L., Morgeson, F. V., Aksoy, L., & Williams, L. (2014). Service Failure Severity, Customer Satisfaction, and Market Share An Examination of the Airline Industry. *Journal of Service Research*, *17*(4), 415–431.

Kim, K., Park, O., Yun, S., & Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change*, *123*, 362–369.

Korfiatis, N., García-Barriocanal, E., & Sánchez–Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205–217.

Korfiatis, N., & Poulos, M. (2013). Using online consumer reviews as a source for demographic recommendations: A case study using online travel reviews. *Expert Systems with Applications*, *40*(14), 5507–5515.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, *42*(7), 3751–3759.

Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, *87*, 105–122.

Kuo, M.-S., & Liang, G.-S. (2011). Combining VIKOR with GRA techniques to evaluate service quality of airports under fuzzy environment. *Expert Systems with Applications*, *38*(3), 1304–1312.

Ladhari, R. (2009). Service quality, emotional satisfaction, and behavioural intentions: A study in the hotel industry. *Managing Service Quality: An International Journal*, *19*(3), 308–331.

Lambrecht, A., & Tucker, C. E. (2017). Can big data protect a firm from competition. *Competition Policy International*, *17*.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21.

Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, *67*, 84–94.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323.

Li, X., & Hitt, L. M. (2008). Self-Selection and Information Role of Online Product Reviews. *Information Systems Research*, *19*(4), 456–474.

Light, R., & Odden, C. (2017). Managing the boundaries of taste: culture, valuation, and computational social science. *Social Forces*, *96*(2), 877–908.

Lin, H.-T. (2010). Fuzzy application in service quality analysis: An empirical study. *Expert Systems with Applications*, *37*(1), 517–526.

Liou, J. J., Hsu, C.-C., Yeh, W.-C., & Lin, R.-H. (2011). Using a modified grey relation method for improving airline service quality. *Tourism Management*, *32*(6), 1381–1388.

Liou, J. J., Tsai, C.-Y., Lin, R.-H., & Tzeng, G.-H. (2011). A modified VIKOR multiple-criteria decision method for improving domestic airlines service quality. *Journal of Air Transport Management*, *17*(2), 57–61.

Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). A method for ranking products through online reviews based on sentiment classification and interval-valued intuitionistic fuzzy TOPSIS. *International Journal of Information Technology & Decision Making*, *16*(06), 1497–1522.

Luo, C., Luo, X. R., Xu, Y., Warkentin, M., & Sia, C. L. (2015). Examining the moderating role of sense of membership in online review evaluations. *Information & Management*, *52*(3), 305–316.

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, *41*(17), 7764–7775.

Melo, A. J., Hernández-Maestro, R. M., & Muñoz-Gallego, P. A. (2017). Service quality perceptions, online visibility, and business performance in rural lodging establishments. *Journal of Travel Research*, *56*(2), 250–262.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Association for Computational Linguistics.

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

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *The Journal of Marketing*, 41–50.

Park, S., Ok, C., & Chae, B. (2016). Using Twitter data for cruise tourism marketing and research. *Journal of Travel & Tourism Marketing*, *33*(6), 885–898.

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2017). Understanding the impact of online reviews on hotel performance: an empirical analysis. *Journal of Travel Research*, *56*(2), 235–249.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rajaguru, R. (2016). Role of value for money and service quality on behavioural intention: A study of full service and low cost airlines. *Journal of Air Transport Management*, *53*, 114–122.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2017). stm: R Package for Structural Topic Models. Retrieved from http://www.structuraltopicmodel.com

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Rodríguez-Díaz, M., & Espino-Rodríguez, T. F. (2017). A methodology for a comparative analysis of the lodging offer of tourism destinations based on online customer reviews. *Journal of Destination Marketing & Management*.

Rossetti, M., Stella, F., & Zanker, M. (2016). Analyzing user reviews in tourism with topic models. *Information Technology & Tourism*, *16*(1), 5–21.

Sotiriadis, M. D., & Van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: the use of twitter by tourists. *Electronic Commerce Research*, *13*(1), 103–124.

Stamolampros, P., & Korfiatis, N. (2018). Exploring the Behavioral Drivers of Review Valence: The Direct and Indirect Effects of Multiple Psychological Distances. *International Journal of Contemporary Hospitality Management*, *30*(8), Forthcoming.

Stamolampros, P., Korfiatis, N., Kourouthanassis, P., & Symitsi, E. (2018). Flying to Quality: Cultural Influences on Online Reviews. *Journal of Travel Research, Forthcoming*.

Suki, N. M. (2014). Passenger satisfaction with airline service quality in Malaysia: A structural equation modeling approach. *Research in Transportation Business & Management*, *10*, 26–32.

Sun, C.-C. (2010). A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods. *Expert Systems with Applications*, *37*(12), 7745–7754.

Suzuki, Y., Tyworth, J. E., & Novack, R. A. (2001). Airline market share and customer service quality: a reference-dependent model. *Transportation Research Part A: Policy and Practice*, *35*(9), 773–788.

Symitsi, E., Stamolampros, P., & Daskalakis, G. (2018). Employees' online reviews and equity prices. *Economics Letters*, *162*, 53–55.

Tellis, G. J., & Johnson, J. (2007). The Value of Quality. *Marketing Science*, *26*(6), 758–773.

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–479.

Tong, Y., Wang, X., Tan, C.-H., & Teo, H.-H. (2013). An empirical study of information contribution to online feedback systems: A motivation perspective. *Information & Management*, *50*(7), 562–570.

Tsui, E., Wang, W. M., Cai, L., Cheung, C., & Lee, W. (2014). Knowledge-based extraction of intellectual capital-related information from unstructured data. *Expert Systems with Applications*, *41*(4), 1315–1325.

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. *Journal of Retailing*, *85*(1), 31–41.

Vlachos, I., & Lin, Z. (2014). Drivers of airline loyalty: Evidence from the business travelers in China. *Transportation Research Part E: Logistics and Transportation Review*, *71*, 1–17.

Wang, C., & Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, *14*(Apr), 1005–1031.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). Cambridge, Massachusetts: MIT press.

Wu, P.-L., Yeh, S.-S., Woodside, A. G., & others. (2014). Applying complexity theory to deepen service dominant logic: Configural analysis of customer experience-and-

outcome assessments of professional services for personal transformations. *Journal of Business Research*, *67*(8), 1647–1670.

Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, *69*(5), 1562–1566.

Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, *52*(7), 850–858.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3), 6527–6535.

Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*, *39*(11), 10283–10291.

[a] Aegean Airlines Awards. http://en.aegeanair.com/all-about-us/awards/
[b] The Guardian (03/08/2017) British Airways Cabin Crew extend strike for further two weeks. https://www.theguardian.com/business/2017/aug/03/british-airways-cabin-crew-extend-strike-further-two-weeks
[c] Reuters (16.05.2017). Lufthansa Shares boosted by upbeat profit http://www.reuters.com/article/us-lufthansa-results/lufthansa-shares-boosted-by-upbeat-profit-target-idUSKBN16N0LC
[d] Skytrax Certified Airline Ratings. http://www.airlinequality.com/ratings/