Journal of Air Transport Management 83 (2020) 101760

Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: http://www.elsevier.com/locate/jairtraman

Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews

Filipe R. Lucini [a,*], Leandro M. Tonetto [b], Flavio S. Fogliatto [a], Michel J. Anzanello [a]

[a] Department of Industrial and Transportation Engineering, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99 – 5° Andar, Porto Alegre, RS, Brazil, 90035-190

[b] Universidade do Vale do Rio dos Sinos, Av. Dr. Nilo Peçanha, 1600, Porto Alegre, RS, Brazil, 91330-002

ARTICLE INFO

Keywords:
Airline industry
Competitive advantages
Text mining
Online customer reviews

ABSTRACT

The airline industry operates in a highly competitive market, in which achieving and maintaining a high level of passenger satisfaction is seen as a key competitive advantage. This study presents a novel framework for measuring customer satisfaction in the airline industry. Using text mining methods we explore Online Customer Reviews (OCRs) to provide guidelines for airlines companies to improve in competitiveness. We analyze a database of more than 55,000 OCRs, covering over 400 airlines and passengers from 170 countries. Using a Latent Dirichlet Allocation model we identified 27 dimensions of satisfaction described by 882 adjectives. Dimensions and adjectives were used to predict airline recommendation by customers, resulting in an accuracy of 79.95%. The most relevant dimensions for airlines' recommendation prediction were calculated. OCRs were stratified according to several variables. Of those, type of passenger impacted the least on the number of dimensions of customer satisfaction, while type of cabin flown impacted the most. Observing results in different publication years we showed airline customer trends through time. Our method showed sensitiveness to identify variations in dimensions distribution according to different passenger characteristics and preferences. Practical implications are that airline service providers aiming at maximizing customer satisfaction should focus their efforts on (i) customer service to first class passengers, (ii) comfort to premium economy passengers, and (iii) checking luggage and waiting time to economy class travelers. Regression analysis revealed cabin staff, onboard service and value for money as top three dimensions of satisfaction to predict the recommendation of airlines. Designing services that excel in those dimensions is likely to improve the company's performance with customers.

1. Introduction

The airline industry operates in a highly competitive market, in which companies have to deal with various challenges to succeed (Calisir et al., 2016; Dolnicar et al., 2011). Examples include oscillating fuel prices, fluctuating demands for service, economic crises, natural disasters, strikes, personnel shortage, restrictive government regulations, and increased security precautions (Calisir et al., 2016; Dolnicar et al., 2011). In such a challenging economic environment achieving and maintaining a high level of passenger satisfaction is seen as a key competitive advantage (Chen, 2008; Li et al., 2017). In that sense, it is important not only to understand how passengers evaluate airlines' services, but also to identify their most valued dimensions of satisfaction (Park et al., 2004).

Customer satisfaction may be measured by the gap between perceived quality of the product or service, and pre-purchase quality expectations (Chow, 2015; Forgas et al., 2010; Guo et al., 2017). Customers tend to be satisfied with an airline when service quality attributes deemed most important are met or surpassed (Chow, 2015); such attributes represent dimensions of satisfaction (Guo et al., 2017). Several studies state that customer satisfaction plays an important role in motivating customers' behavioral loyalty, which translates into giving positive reviews, returning as customers, or recommending the product or service to others (Forgas et al., 2010; Guo et al., 2017; Mattila, 2004; Morgan and Hunt, 1994). On the other hand, unsatisfied passengers may reconsider using the same airline in future flights (Namukasa, 2013), or

* Corresponding author.
E-mail addresses: filipe.lucini@producao.ufrgs.br (F.R. Lucini), ltonetto@unisinos.br (L.M. Tonetto), ffogliatto@producao.ufrgs.br (F.S. Fogliatto), anzanello@producao.ufrgs.br (M.J. Anzanello).

https://doi.org/10.1016/j.jairtraman.2019.101760
Received 18 June 2019; Received in revised form 18 October 2019; Accepted 19 December 2019
Available online 27 December 2019

start a negative word-of-mouth campaign (which may be electronic) that could cause damage to the company's reputation and image (Blodgett and Li, 2007).

Measuring customer satisfaction in the airline industry is becoming more frequent and relevant (Park et al., 2005); there are at least five reasons for that (Grigoroudis and Siskos, 2009). First, customer satisfaction measurement programs could improve the communication with the clientele. Second, companies may examine whether their services fulfill customers' expectations. Moreover, companies may analyze the impact of specific efforts and actions on the clientele. Third, key satisfaction dimensions that should be improved may be identified. Fourth, companies could identify their strengths and weaknesses against competition, based on customer perceptions and judgments. Finally, companies' personnel may be motivated to increase their productivity given that all improvement efforts promoted on services are evaluated by customers themselves.

However, since the relationship between service quality dimensions and customer satisfaction may show a nonlinear pattern, assessing customer satisfaction may become a challenging task (Basfirinci and Mitra, 2015). In recent years, different approaches were proposed to identify and measure airline service quality and passenger satisfaction (Tsafarakis et al., 2017). One research stream uses statistical techniques such as regression (e.g. logistic or ordinal) to model the relationship between quality dimensions and customer satisfaction (Ali et al., 2016; Eboli and Mazzulla, 2009; Josephat and Ismail, 2012); another uses Multi-Criteria Decision Making (MCDM) methods and tools to evaluate airlines' service levels and generate propositions for improvements (Li et al., 2017; Liou et al., 2011; Liou and Tzeng, 2007).

Conventional qualitative, quantitative or mixed methods (e.g. focus groups, questionnaire surveys, or a combination of both) are generally used by researchers and practitioners as an information source to identify and measure customers' dimensions of satisfaction (Guo et al., 2017). However, these methods are time consuming and often inaccurate (Wan and Gao, 2015). For example, biased conclusions may result from limited sample sizes or inconsistent measurement items and questions (Chow, 2015). Not all customers take questionnaires seriously, and many customers fill them out randomly, bringing noise into analyses (Wan and Gao, 2015). In addition, questions are normally set based on previous research (Guo et al., 2017), compromising the uncovering of subjects not contemplated in questionnaires; e.g. emerging customer preferences (Wan and Gao, 2015).

As an alternative to conventional methods, studies from different fields have shown that user-generated content (UGC) can be used as an information source to understand customer preferences and demands (Chau and Xu, 2012). UGC results from the widespread diffusion of Web 2.0 technologies (Guo et al., 2017), which enabled customers, including those from the airline industry, to share their experiences and opinions. Thousands of customers may share spontaneous, insightful, and passionate online feedback, thereby creating the "wisdom of crowds" (Surowiecki and Silverman, 2007). More important, such feedback is widely available online and easily accessible for free or at low-cost (Guo et al., 2017). One popular form of UGC is online customer reviews (OCRs) (Felbermayr and Nanopoulos, 2016). Usually, OCRs are provided by customers who had direct (and usually recent) experience with a product or service; they are reported in multiple forms including online ratings (e.g. number of stars) and online reviews (e.g. personal opinion in text format) (Flanagin and Metzger, 2013; Jin et al., 2016; Sparks et al., 2016).

The present study aims to explore the "wisdom of crowds" (Surowiecki and Silverman, 2007) contained in thousands of OCRs to provide directions for airlines companies to improve in competitiveness. More specifically, this study aims to: (*i*) identify and extract dimensions of customer satisfaction expressed in OCRs; (*ii*) verify the distribution and importance of those dimensions in OCRs from different groups of airline customers; (*iii*) identify and extract adjectives used to describe perceptions in those dimensions and calculate the adjectives' sentiment scores; and (*iv*) test and validate the dimensions and adjectives in (*iii*) through regression analysis. By exploring OCRs, we believe that the "voice of the customer" (Gaskin et al., 1993) will be heard, since those records are written in their own language. OCRs are also organized and prioritized in the way customers think about, use, and interact with airline services.

This article contributes to the state-of-the art on the use of spontaneous generated content to explore dimensions of customer satisfaction in at least three ways. First, to the best of our knowledge, this is the first attempt to use Latent Dirichlet Allocation (LDA) to identify and extract dimensions of customer satisfaction using airline company data. Despite the fact that LDA is considered the principal method to extract latent topics from unstructured texts (Zhao et al., 2015), there is no evidence of using it to analyze customer satisfaction in the airline industry. Second, we propose to expand the scope of data stratification, normally limited to a specific airline and/or a particular online rating system, to include passenger characteristics and preferences (e.g. nationality, type of traveler, and cabin flown). We also propose a temporal stratification of data to enable the analysis of changes in customer demands over time. Third, this study uses a database with over 55,000 OCRs, referring to more than 400 airlines, and passengers from over 170 countries. Compared to previous studies, our data base is much larger allowing more reliable generalizations [e.g. 2,105 tweets in Mostafa (2013), 7, 466 OCRs in Yao et al. (2015), 9,238 tweets in Sreenivasan et al. (2012), 10,895 tweets in Liau and Tan (2014), and 12,864 tweets in Wan and Gao (2015)].

2. Background

Despite considerable attention devoted to analyzing UGCs in different industries, there is a gap in the literature regarding retrieval and information mining from such source in the airline industry. Previous studies are mostly related to sentiment analysis (Liau and Tan, 2014; Mostafa, 2013; Wan and Gao, 2015) and topic detection (Liau and Tan, 2014; Sreenivasan et al., 2012), and have used UGCs from the online social networking service Twitter. On Twitter, users produce online content in the form of micro-messages called "tweets" (Liau and Tan, 2014).

Sreenivasan et al. (2012) investigated how airlines and airline customers communicate through Twitter analyzing 9,238 tweets related to three airlines, using content analysis. Results suggested that customers use Twitter mainly to give compliments, share information, offer help, and provide personal updates; in opposition, airlines primarily use Twitter for marketing purposes. Based on their sample, the authors concluded that airlines did not appear to be responsive to users' concerns and issues.

Using a different approach, Yao et al. (2015) analyzed 7,466 OCRs from SkyTrax, a website specialized in reviews of the airline industry. They identified and compared the most representative terms used to refer to 25 different companies, which were compiled in 25 sets of reviews. Analysis of the 50 most frequent terms in each set displayed no relevant differences across airlines. The authors also analyzed similarities between reviews made by users giving different star levels (1–5) to companies. Results showed that users wrote reviews using similar terms regardless of the airline, but that differences in the occurrence of terms in reviews increased with the difference in star levels.

Mostafa (2013) analyzed 2,105 tweets expressing customers' sentiments towards the service provided by 16 airlines. For that, the author analyzed the occurrence of 6,800 seed adjectives proposed by Hu and Liu (2004) with known orientation. Negative sentiments in tweets were given a score -1, while positive sentiments were given a score $+1$. Results suggested that most airline services are sub-optimal since sentiment scores were overall negative. However, it is important to note that the analysis did not consider the context in which the adjectives were used. The author suggests the use of topic detection techniques such as LDA to identify most relevant topics responsible for generating positive and negative sentiments, as well as a larger dataset for more reliable results.

Another study by Liau and Tan (2014) examined 10,895 tweets from customers of a low-cost airlines operating in Malaysia. To study customers' opinions on the airline, two algorithms were tested for topic detection and two algorithms for sentiment analysis. Results showed that customers mainly discussed four topics on Twitter: customer service, ticket promotions, flight cancellations and delays, and post-booking management. Regarding sentiment analysis, customers displayed more positive than negative sentiments across the four main topics. Customers tended to write positive tweets regarding customer service, booking management and ticket promotions, and negative tweets regarding flight cancelations and delays.

More recently, another study (Wan and Gao, 2015) tested seven different classification algorithms for sentiment analysis, one of which was an ensemble algorithm based on majority voting on other six classifiers. Tests used a sample of 12,864 tweets on 16 airlines, which were manually labeled as carrying positive, neutral, and negative sentiments. The algorithms' performances were evaluated for two-class (positive and negative sentiments) and three-class (positive, neutral, and negative sentiments) datasets using a 10-fold cross-validation method. The ensemble algorithm presented the best performance for both two-class and three-class datasets, with an *F*-measure (or F1-score, which gives the harmonic mean of precision and recall in classifying the sample) of 84.2% and 91.7%, respectively.

One study from the hospitality industry explores UGCs through the combined use of LDA and statistical modelling, resembling some of the propositions in this paper. Guo et al. (2017) analyzed a dataset of 266, 544 online reviews for 25,670 hotels located in 16 countries made available through TripAdvisor, a popular online review community of travel consumers. In their study, dimensions of customer satisfaction, importance of different dimensions, and words related to dimensions were extracted through LDA. They tested for differences in dimensions' importance weights stratifying the sample of respondents by gender and age, and ran a stepwise regression analysis using ratings on five dimensions (location, cleanliness, room experience, service quality, and value for money) as independent variables and overall satisfaction as the dependent variable. Although using LDA for dimension extraction from textual records, the remaining method steps in Guo et al. (2017) differ from the ones proposed here.

A second study (Xu and Li, 2016), also from the hospitality industry, analyses customer reviews from Booking.com to obtain customer satisfaction dimensions using latent semantic analysis, a least generalizable version of LDA. Their study however is restricted to the identification of dimensions and analysis of high-loading terms within each dimension.

3. Methods

3.1. Database

We used data extracted from the website Air Travel Review (ATR), created and currently managed by SkyTrax, and available at airlinequality.com. ATR is an independent customer forum, which has become the leading review website for airline, airport, and associated air travel passenger reviews. It provides comprehensive ratings and reviews contributed by customers for over 681 airlines and 725 airports worldwide, being used in previous studies (e.g. Yao et al., 2015; Xu et al., 2018). OCRs in ATR are individually registered, enabling to parse individual reviews and aggregate them for our analytical purposes. For that, we built a web crawler that collected all available OCRs in ATR on August 11th, 2016. The web crawler, programmed in Python, accessed the webpage and downloaded the html code. Based on the html structure code, the crawler separated all information needed to run the analysis presented in the paper. In total, 55,775 OCRs were collected and saved in a comma-separated value (.csv) file. All OCRs were comprised of publication date, a free form text review written in English, and a yes or no answer to the question "would you recommend the airline to other passengers?". Table 1 presents statistics for length, in terms of number of words, of the free form text reviews.

Some OCRs also exhibited contextual information, such as nationality, type of passenger, and cabin flown. A general score, ranging from one to ten, was also available in some OCRs. Table 2 summarizes frequencies of the available contextual information. The database refers to 419 airlines evaluated by passengers of 171 nationalities, thus representing a wide spectrum of points of view.

3.2. Pre-processing

According to Tirunillai and Tellis (2014) the analysis of free form text from OCRs is difficult for three main reasons. First, reviews do not have a standard structure, and most texts tend to be casual in their word and grammar usage. Second, only few words are informative regarding the product and its dimensions of quality, calling for cleaning and selection procedures. Third, text must be transformed to a format that enables numerical analysis. The main steps needed to overcome those difficulties and prepare reviews for the statistical analysis are presented next.

The pre-processing steps implemented here are similar to those adopted in previous studies (Guo et al., 2017; Lee and Bradlow, 2007; Tirunillai and Tellis, 2014). First, words containing non-English characters were converted to ASCII code (e.g. "Ljubljana Jože Pučnik Airport" was converted to "Ljubljana Joze Pucnik Airport"). Synonyms were substituted by their most frequent version in reviews. Text illustrating time, periods of time, weights, sizes, prices, dates, flight codes, airline names, airplane models, routes, countries, languages, airport names and codes, and city names and codes were replaced by key words. For example, airport names and codes were replaced by the word "_airport_" (the character "_" at the beginning and end of key words enabled the identification of added words). Table 3 presents the key words used and some examples of replaced texts.

Next, reviews were broken into individual sentences delimited by specific punctuation (".", "?" or "!"). The continuous string of characters of each sentence was broken down into linguistic units called tokens, which were delimited by blank spaces in the string (Lucini et al., 2017). Part-of-speech (POS) tagging was applied to identify the syntax function of each token in sentences. For example, the use of POS tagging in the sentence "Staff was very nice" resulted in "Staff" (noun) "was" (verb) "very" (adverb) "nice" (adjective). Only adjectives and nouns were kept. Although there is evidence in favor of using adjectives and adverbs to enhance results of sentiment analyses (see, for example, Benamara et al., 2005), that was not verified in our dataset and we chose to use only adjectives and nouns in the determination of customer satisfaction dimensions. Compound nouns were adjusted to become single tokens (e.g. "cabin_staff"). All capital letters were replaced by lowercases, stop words were removed (Lucini et al., 2017), and words were reduced to their base form, excluding inflectional endings (lemmatization) (Feldmann and Sanger, 2007). Finally, low frequency tokens (i.e. below 2%) were excluded (Guo et al., 2017). After that, an original review that read:

> "Two short flights with Adria Airways (Ljubljana - Zurich and Munich -Ljubljana), both on time, aircrafts new, cabin bright and clean. Nice and professional service on both flights. On the morning flight out of Ljubljana I did not take any free newspapers and just said

Table 1
Statistics for length of reviews – number of words.

Minimum	Q1	Median	Q3	Maximum	Mean
14	76	114	170	988	135

Q1 first quartile, Q3: third quartile.

Table 2
Contextual information and frequencies.

Variable	Frequency	%
Airline	55,775	100.00
British Airways	1,541	2.76
Spirit Airlines	1,498	2.69
United Airlines	1,361	2.44
American Airlines	1,199	2.15
Other (415 airlines)	50,176	89.96
Year of Review Publication	55,775	100.00
2002–2012	11,135	19.96
2013–2014	22,492	40.33
2015–2016	22,148	39.71
Passenger Nationality	54,094	96.99
United Kingdom	13,435	24.09
United States	12,054	21.61
Australia	6,333	11.35
Canada	4,126	7.40
Other (167 nationalities)	18,146	32.54
Type of Passenger	17,105	30.67
Solo Leisure	5,991	10.74
Couple Leisure	4,861	8.72
Family Leisure	3,554	6.37
Business	2,699	4.84
Cabin Flown	52,756	94.59
Economy Class	40,719	73.01
Business Class	8,667	15.54
Premium Economy Class	2,111	3.78
First Class	1,259	2.26
General Score	51,302	91.98
1	8,944	16.04
2	7,643	13.70
3	4,357	7.81
4	3,330	5.97
5	2,390	4.29
6	3,241	5.81
7	2,537	4.55
8	4,445	7.97
9	7,114	12.75
10	7,301	13.09
Passenger Recommends Airline?	55,775	100.00
Yes	29,491	52.87
No	26,284	47.13

that I will sleep. The chief flight attendant before take-off brought me a pillow for a more comfortable sleeps. Nice touch."

Became:

"short flight _airline_ _route_ _route_ on_time aircraft new cabin bright clean nice professional flight morning flight _city_ free newspaper chief cabin_staff pillow comfortable sleep nice touch"

The information of the syntax function of each token was kept allowing the selection of tokens to attend the different objectives of the study. For example, to identify and extract dimensions of customer satisfaction only nouns were kept in records. On the other hand, only adjectives were kept to calculate sentiment scores of terms used to describe perceptions of satisfaction dimensions. Finally, nouns and adjectives were used to validate findings through regression analysis.

Once all reviews had been pre-processed, they were converted to set-of-words representation. For that, a matrix was created indicating the occurrence of tokens in reviews. Matrix columns corresponded to tokens and rows corresponded to reviews. Matrix cells were filled using the term frequency – inverse document frequency (TF-IDF) value (Lucini et al., 2017). All pre-processing steps were carried out using scikit-learn (Hackeling, 2014) and NLTK (Bird et al., 2009) modules in Python 3.6 (Python Software Foundation, 2017).

3.3. Identification and extraction of dimensions of customer satisfaction

One objective in this study is to identify and extract potential dimensions influencing customer satisfaction in the airline industry. To achieve that goal we used Latent Dirichlet Allocation (LDA), a popular topic detection method from the field of machine learning and natural language processing. LDA assumes that each document (i.e. review) may be represented as a probabilistic distribution over latent topics (i.e. dimensions of customer satisfaction), and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is represented as a probabilistic distribution over words, and the word distributions in topics also share a common Dirichlet prior. Topics can potentially be shared by all reviews, and every review has its own mixing proportion of topics. Formally, LDA assumes the following generative process for each document in a corpus (Blei et al., 2003):

1 Choose $N \sim \mathit{Poisson}(\xi)$.
2 Choose $\theta \sim \mathit{Dirichlet}(\alpha)$.
3 For each w_n in $w = \{w_1, w_2, \ldots, w_N\}$:
  a. Choose $z_n \sim \mathit{Multinomial}(\theta)$.
  b Choose w_n from $p(w_n | z_n, \beta)$

Table 3
Key words and examples of replaced texts.

Key word	Examples of replaced text	Key word	Examples of replaced text	Key word	Examples of replaced text
time	10 am	_weight_	1 kg	_price_	$ 500
	10:00 am		20,5 g		£ 20
	14:30		2.3 lbs		€ 300
	10.00 pm	_airplane_	A380		12 dollars
			Boeing 787-800		30 euros
_period_time_	10–15 min		ERJ-190		20 CAD
	1–2 days		Dash 8-400		AUD$660
	3 and a half hours	_route_	POA-CGH		
	2.5 months		POA to CGH	_language_	English
	2 weeks		London – New York		German
			POA-GRU-MIA		Portuguese
_flight_code_	AC4601		POA to MIA via GRU		
	DL 471	_date_	on march 27	_airport_	Heathrow
airline	British Airways		on june 4th		John F Kennedy
	Air Canada		march 27		JFK
	Iberia		on the 13th october	_size_	2 cm
country	Brazil		30th december 2015		
	USA		5th may		1.5 m
	Australia		26 april		2,3 in
			26 april 2017		4 inches
city	London		march 16 2015		2 ft
	New York		05/03/2011		4 feet
	NYC		on june 4th		6′ 2″

where N is the number of words in the document, ξ is the parameter of the Poisson distribution that shows the length of the reviews in each document, θ is the topic distribution of the document, α is the parameter of the Dirichlet prior on the per-document topic distribution, w_n is the n-th word used in the document $\boldsymbol{w}$, z_n is the topic for the n-th word, $\boldsymbol{z}$ is a set of topics, $p(w_n|z_n, \beta)$ is a multinomial probability conditioned on the topic z_n, and β is the parameter of the Dirichlet prior on the per-topic word distribution.

The key inferential problem that need to be solved in order to use LDA is that of computing the posterior distribution of the hidden variables (θ and $\boldsymbol{z}|\boldsymbol{w}$) given a document (equation (1)):

$$p(\theta, z|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, z, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \tag{1}$$

However, the exact inference of this distribution is intractable due to the coupling between θ and β in the summation over latent topics. So, approximate algorithms such as the variational inference (Airoldi et al., 2008; Blei and Jordan, 2003; Hoffman et al., 2010) or the Markov Chain Monte Carlo (Griffiths and Steyvers, 2004; Shivashankar et al., 2011; Tirunillai and Tellis, 2014) are typically used for that. In this study we used the Batch Variational Bayes Inference (BVBI) algorithm (Hoffman et al., 2010), available in the package scikit-learn (Hackeling, 2014). For more detailed descriptions of LDA and BVBI see Blei et al. (2003) and Hoffman et al. (2010), respectively.

Determining the ideal number of topics in a LDA model is a challenging task when the number of dimensions of satisfaction is not known *a priori* (Zhao et al., 2015; Zhao et al., 2014). An insufficient number of topics could render a model that is too coarse to identify accurate dimensions. On the other hand, an excessive number of topics could result in a model that is too complex, making interpretation and subjective validation difficult (Zhao et al., 2015). We followed the perplexity-based method (Zhao et al., 2015) to determinate the ideal number of latent topics in our dataset. This is an iterative approach in which different models with different number of topics are tested and compared using the perplexity measure. Perplexity is a popular measurement in information theory, used to evaluate how well a statistical model describes a dataset (Zhao et al., 2015); lower perplexity denotes a better probabilistic model. Formally, for a test set of documents the perplexity measure is defined as:

$$perplexity_{D_{test}} = exp\left\{ -\frac{\sum_{d=1}^{M} \log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{2}$$

where D_{test} is the test set, M is the number of documents in D_{test}, $\boldsymbol{w}_d$ is the d-th document of D_{test}, and N_d is the number of words in $\boldsymbol{w}_d$.

The dataset was divided in train and test sets to allow k-fold cross-validation considering models fitted to a number of topics varying from 2 to 100. The dataset was divided in k mutually exclusive subsets of equal size, such that one subset is used for perplexity measurement and $k-1$ subsets are used for parameter estimation. This process is carried out k times alternating the test subset; performance statistics are calculated from the results (Hastie et al., 2008). In this study, k was set to 5. Of all models tested pick the one with smallest mean perplexity and build a confidence interval (CI) for the perplexity values (say, reference CI) using the cross validation mean and standard-deviation statistics. Now list all models with complexity CIs overlapping the reference CI, and identify the one fitted to the smallest number of topics, which will be considered the ideal number of topics in remaining steps.

Once the number of topics is identified, two steps are needed to extract and identify dimensions of customer satisfaction expressed in OCRs (Tirunillai and Tellis, 2014). First, the LDA model is fitted using the entire dataset and the established number of topics. Then, important attributes of each dimension of satisfaction are identified (i.e. words closely related to each dimension). In that sense, words that have the highest likelihood to belong to a specific topic are considered important. The top ten words of each topic are listed according to the likelihood scores of the fitted model. In the second step each topic is named with a dimension of satisfaction, based on its list of important words.

Following (Guo et al., 2017), the naming of dimensions is conducted by two researchers individually, and results are compared and discussed until they reach consensus. Each researcher subjectively tries to identify logical connections between the top important words (e.g. the top three) for a topic, assigning it a candidate name. The candidate name is further tested subjectively via logical connection with other words in the top-10 list. Whenever a connection is found the candidate name is considered permanent; if a word that does not fit the topic candidate name is found, the naming process restarts and is repeated until a name logically connected to all words is devised. Final names assigned to topics are considered as dimensions of customer satisfaction expressed in the OCRs.

3.4. Distribution of customer satisfaction dimensions

Once the LDA model is fitted and dimensions of customer satisfaction are identified, it is possible to evaluate the distribution of probabilities of those dimensions across different groups of reviews. As a first step, all reviews are analyzed individually in order to calculate the distribution of probabilities of the identified dimensions. Next, reviews are grouped according to five different criteria: airline, publication year, passenger nationality, type of passenger, and type of cabin flown. Table 2 presents the strata used for each criterion. The probability of occurrence of a given dimension (e.g. onboard service) within a group (e.g. first-class passengers) is given by the mean of the probability values for that dimension across all reviews belonging to that group. Standard-deviations for each dimension within a group were also obtained using the review-topic pairs' probabilities. Next, a confidence interval referring to the probability of occurrence of each dimension within the group is obtained using the respective calculated mean and standard-deviation values. Finally, analysis of confidence intervals is carried out to identify differences and similarities among strata.

3.5. Sentiment analysis

To identify, extract and calculate sentiment strengths of adjectives that are normally used by airlines' customers, the pre-processed dataset was adjusted to keep only adjectives in reviews. We propose a method to analyze the sentiments of adjectives based on the coefficients of a Naïve Bayes Classifier (Duda et al., 2001). The Naïve Bayes Classifier is a supervised method based on Bayes' theorem. The sentiment strength of an adjective is calculated by the probability of a class happening (e.g. positive or negative sentiment) given that the adjective has being used in a review. Equation (3) presents its formulation. The higher the value of $P(Class|Adjective)$, the more informative the adjective is to classify a review as belonging to a specific class (i.e. the stronger is the sentiment, negative or positive, associated with the adjective).

$$P(Class|Adjective) = \frac{P(Adjective/Class)P(Class)}{P(Adjective)} \tag{3}$$

To carry out this analysis, some reviews were classified as presenting negative or positive sentiments. We considered negative reviews those with a general score of one (the lowest score available, occurring in 8,944 OCRs). On the other hand, reviews with a general score of ten (the highest score available, occurring in 7,301 OCRs) were considered positive reviews. All other reviews with general scores ranging from two to nine were not considered for classification. In addition, adjectives that appeared in less than 0.1% of the documents were disregarded from the analysis.

The Naïve Bayes Classifier was fitted resulting in sentiment strength coefficients for each of the adjectives, which were normalized (i.e. scaled) to facilitate interpretation. The normalization strategy linearly transforms a sentiment strength coefficient (X) according to equation

(4), such that normalized strength coefficients range from -1 to $+1$, with -1 and $+1$ denoting a very negative and very positive sentiment, respectively.

$$\left(\frac{X-\min(X)}{\max(X)-\min(X)}\right)\times 2-1 \tag{4}$$

As a result, the method provided a list of most frequent and important adjectives used by airline customers when describing their sentiment in reviews, as well as their sentiment scores.

3.6. Regression analysis

To test and validate the dimensions of satisfaction and adjectives extracted from OCRs, we propose the use of a Logistic Regression Classifier (Zhang et al., 2016) to predict the recommendation of airlines. Given the binary nature of the dependent variable to be predicted, Logistic Regression may be deemed a logical choice of modeling technique. In the regression, we used the entire pre-processed dataset: independent variables were the satisfaction dimensions identified using LDA (their importance and sentiment scores), and the dependent variable was the recommendation given by respondents regarding the airline reviewed (recommend/do not recommend). The dataset was divided in training and testing sets to allow a k-fold cross-validation, with k set to 10. The model's prediction performance was assessed through accuracy measurements. Applying the best regression model to the entire dataset also allowed us to identify the most relevant dimensions for predicting airline recommendation.

4. Results

4.1. Identification and extraction of dimensions of customer satisfaction

The best LDA model was obtained as follows. For a given number of topics t we partitioned the dataset in five parts (one for testing and four for training), fitted an LDA model to the training portion and calculated its perplexity using the testing portion; the partitioning was repeated five times in total. Using the five perplexity values obtained we calculated their mean, standard-deviation, and 95% CI. The procedure was then repeated for $t = 2, \ldots, 100$. The lowest perplexity mean of 72.65 (95% CI: 71.28–74.03) was achieved using the model with 85 topics. However, the model with smallest number of topics overlapping the 85-topic model perplexity CI had 67 topics and a perplexity mean score of 74.12 (95% CI: 72.46–75.77). Fig. 1 displays the mean perplexity as a function of the number of topics in the model. Once the number of topics was set to 67, the entire dataset was used to fit the LDA model to be used in what follows.

All topics were then named according to the procedure previously described. For example, in Table 4 the topic named "checking luggage" is based mainly on the word "luggage" (weight = 0.3964), which appears at the top of the list. Considering that all other words could be related to the act of checking luggage, the topic was named as such. Next, we grouped topics that had a similar content and meaning, and discarded three topics deemed not relevant (i.e. those with probability 0.05% or less over topics distribution). The final result was 27 dimensions of satisfaction, which are given in Fig. 2.

4.2. Distribution and importance of customer satisfaction dimensions

All reviews were analyzed individually and grouped according to five different criteria. 95%-confidence intervals were calculated for each criteria and dimension. Figs. 3–7 show graphs, with confidence intervals, for the five criteria (airline, publication year, passenger nationality, type of passenger, and type of cabin flown). In those figures, dimensions are listed in the horizontal axis; for a given dimension, we present their mean percentage participation, stratified according to one of the criteria, and the percentage CI.

Regarding the airline criterion, there are significant differences in the distribution of some dimensions (Fig. 3). For example, passengers of British Airways write more about food and drink, onboard services, problems, business class, airplane characteristics, couple flights, comfort and airport lounge, when compared to other companies. On the other

Table 4
Example of attributes and scores in topic "Checking luggage".

Attributes	Score
luggage	0.3964
time	0.1912
price	0.1745
person	0.0775
fee	0.0693
flight	0.0448
_period_time_	0.0215
time	0.0190
route	0.0045
city	0.0010

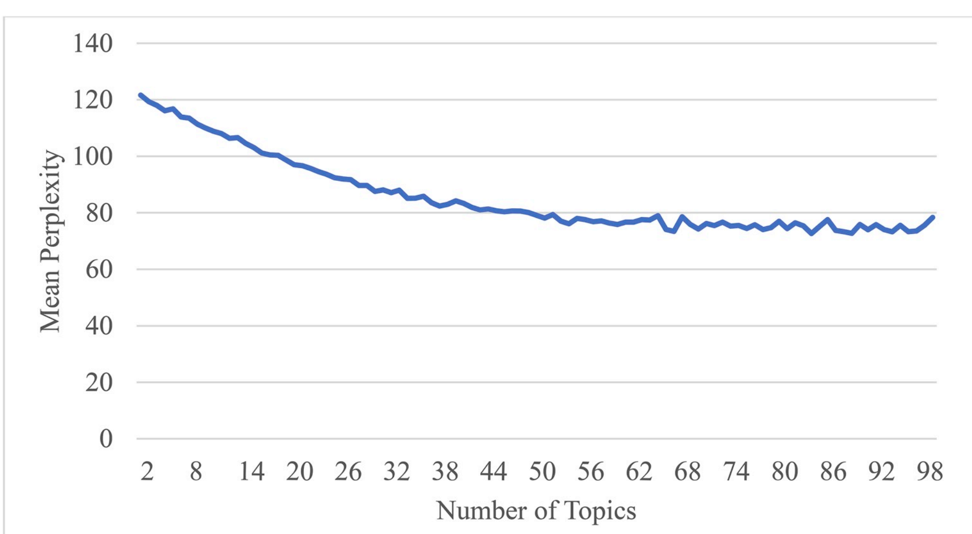

Fig. 1. Mean perplexity as a function of the number of topics.

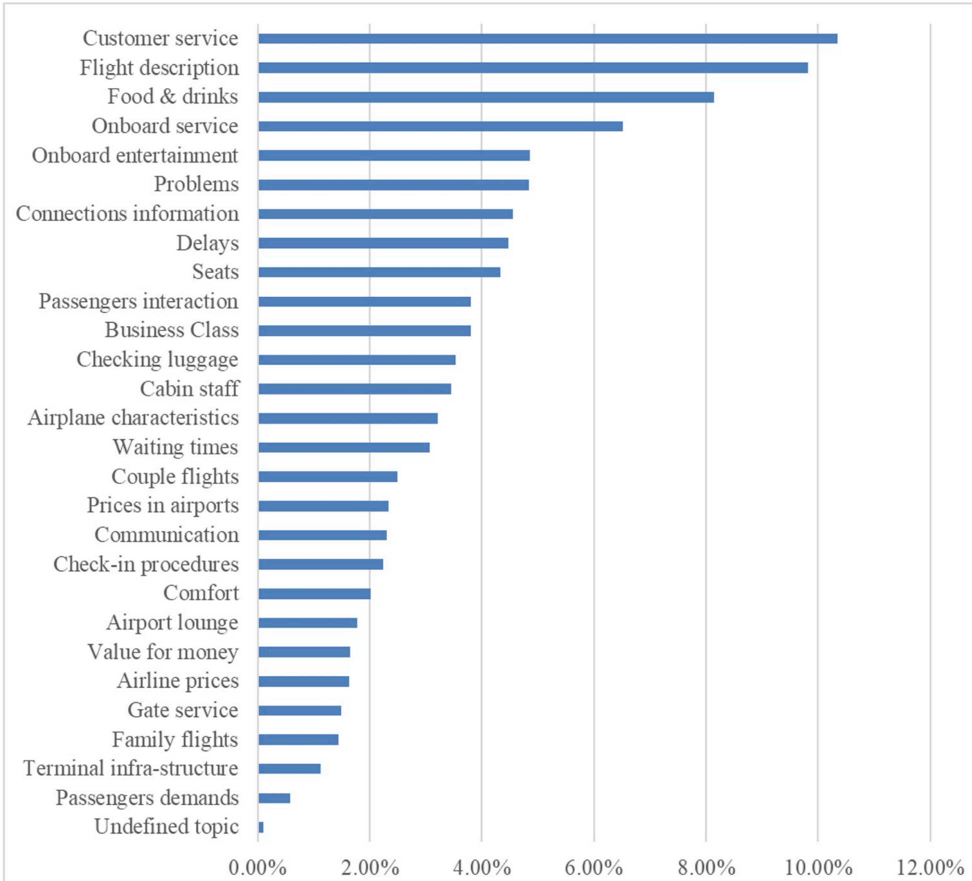

Fig. 2. Satisfaction dimensions and their distribution over the entire dataset.

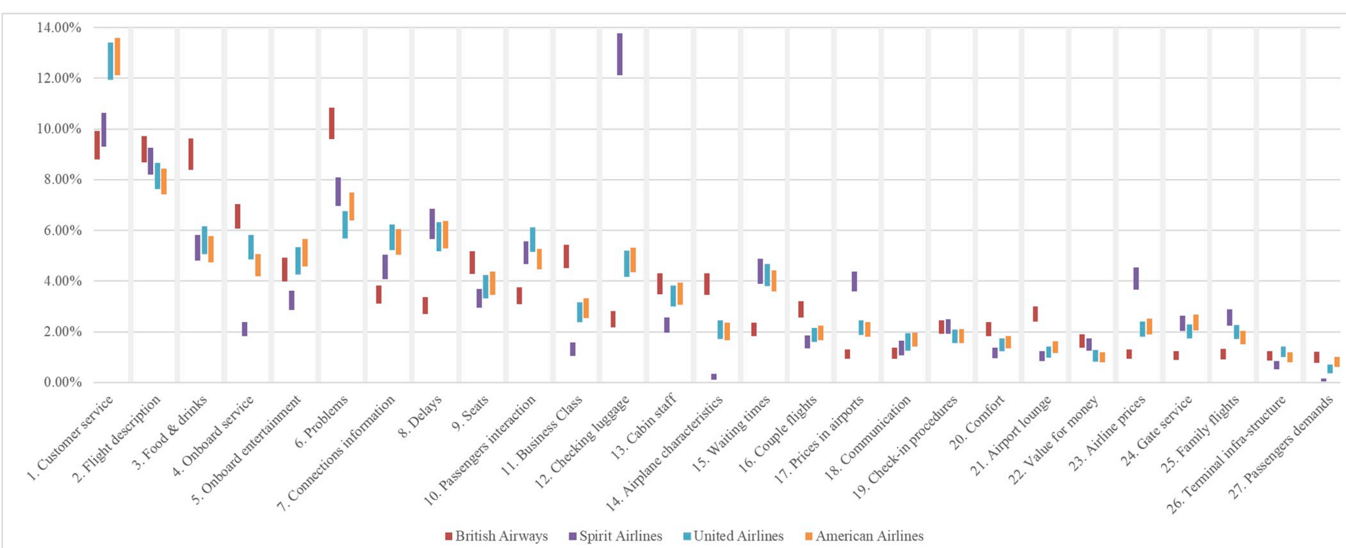

Fig. 3. Confidence intervals for airline criterion in data stratification.

hand, they write significantly less about connection information, delays, passenger interaction, checking luggage, waiting times, prices at airports, airline prices, gate services, and family flights. Spirit Airlines passengers prefer to write about checking luggage, prices in airports, and airline prices, and write less about onboard service, onboard entertainment, business class, cabin staff, airplane characteristics, and passengers demands (Spirit Airlines is a low cost company that operates only in the United States). Finally, American Airlines and United Airlines passengers have similar distributions across all dimensions.

When analyzing the distribution of dimensions in different periods of time (Fig. 4), it is possible to notice that many dimensions have similar proportions regardless of the year in which they were written. However, it is possible to verify some trends in the data. For example, passengers wrote less about flight descriptions, food and drinks, and airplane

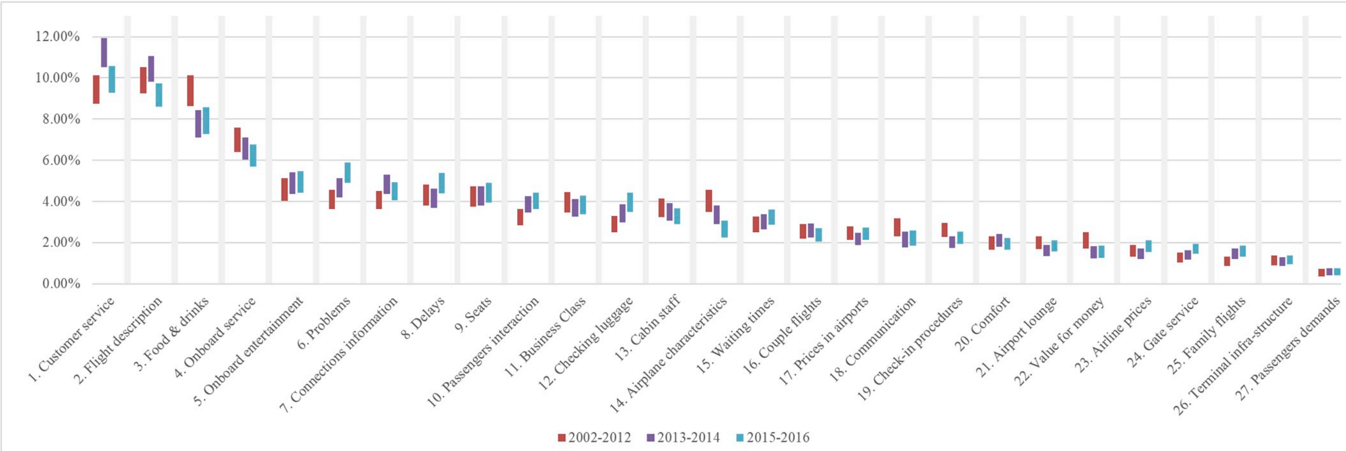

Fig. 4. Confidence intervals for publication year criterion in data stratification.

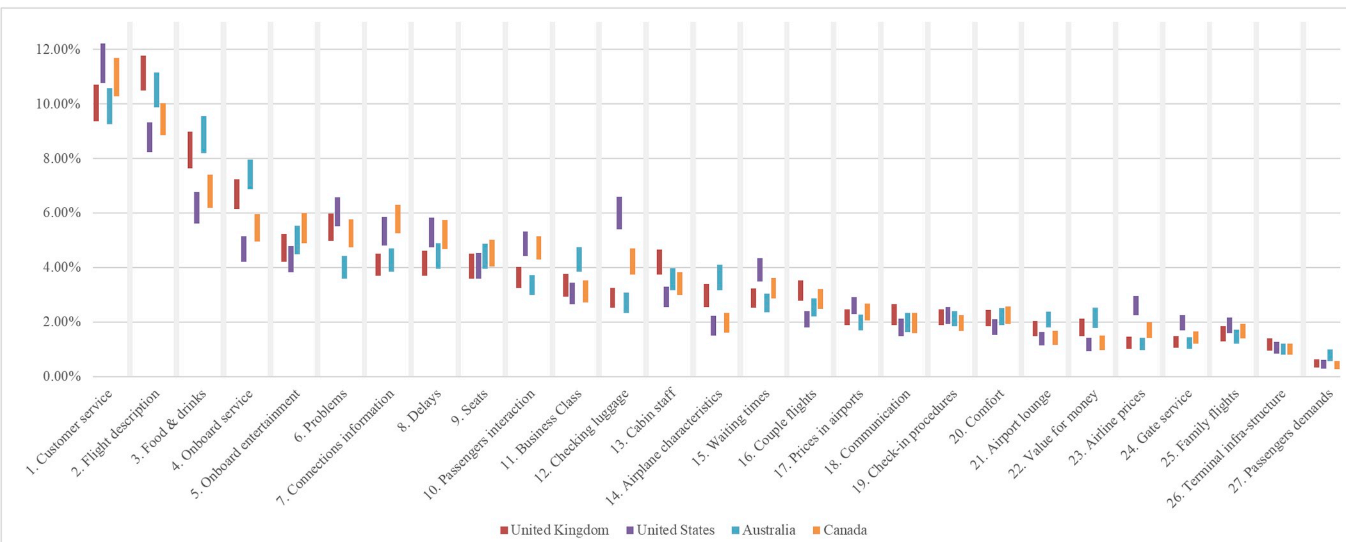

Fig. 5. Confidence intervals for passenger's nationality criterion in data stratification.

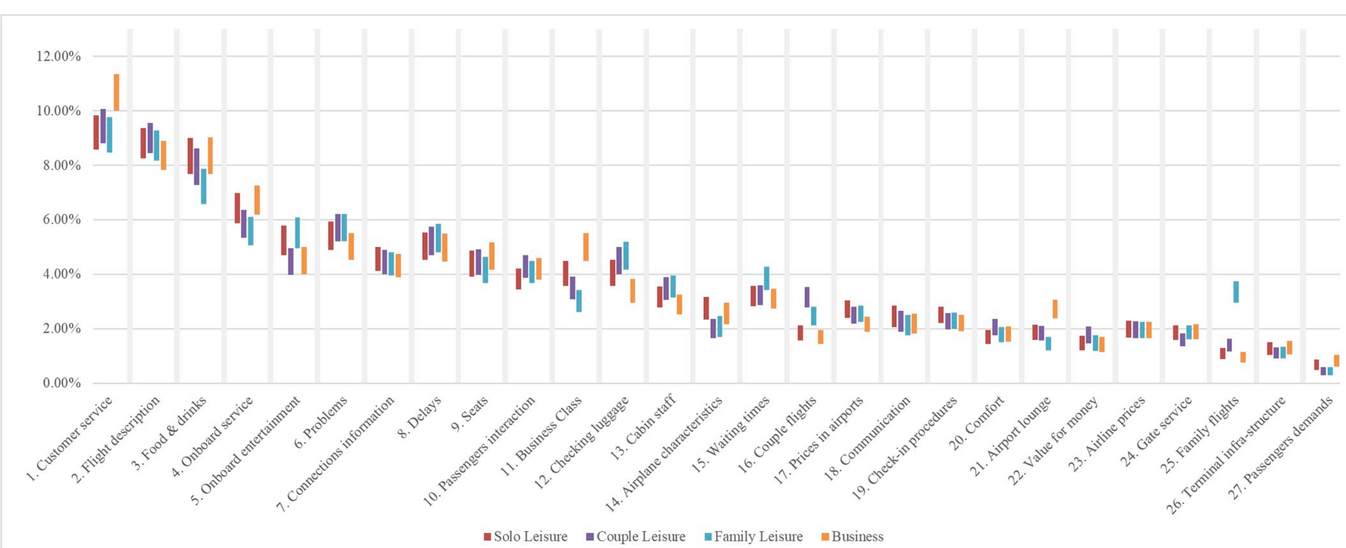

Fig. 6. Confidence intervals for type of passenger criterion in data stratification.

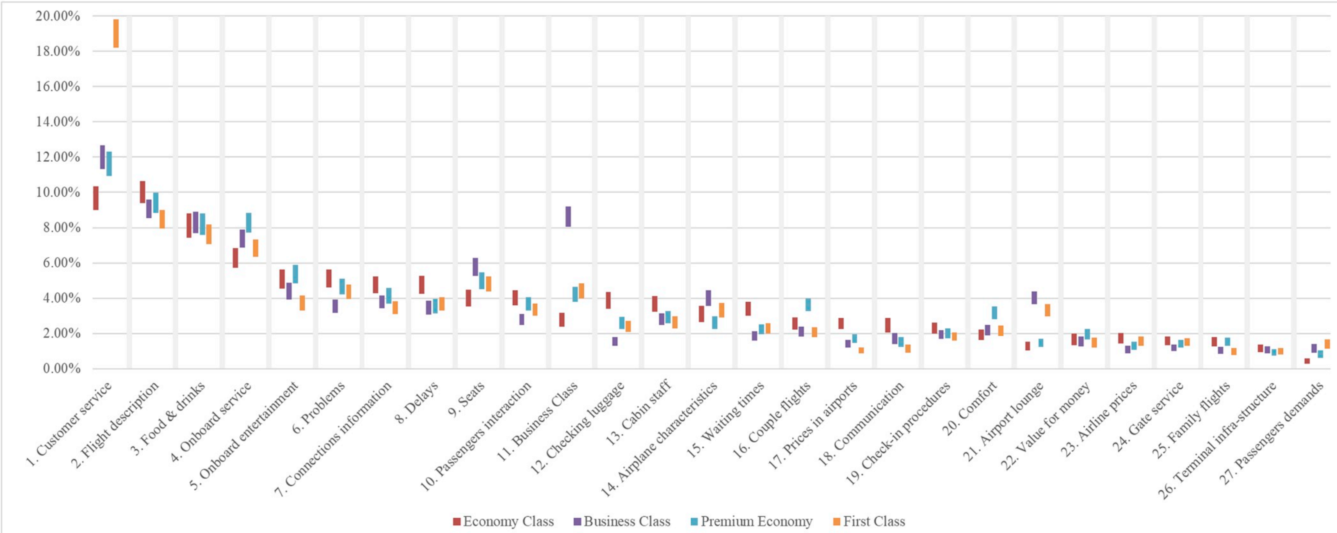

Fig. 7. Confidence intervals for cabin flown criterion in data stratification.

characteristics in recent years. In contrast, they are writing more about problems, passenger interaction, and checking luggage.

The distribution of dimensions considering passenger nationalities (Fig. 5) point to passengers giving attention to different dimensions, although some similarities among nationalities could be identified. For example, Americans and Canadians have similar behavior, which is significantly different from the British and Australians when writing about food and drink, onboard service, connection information, passenger interaction, airplane characteristic, and value for money. In contrast, all analyzed nationalities write similarly about seats, communication, check-in procedures, comfort, and terminal infra-structure.

Fig. 6 shows that most dimensions present similar importance to different types of passenger. However, compared to other types, passengers flying on business trips write more about business class and airport lounge dimensions. Also, passengers traveling with their families write more about family flights, as expected. Solo leisure and business passengers give similar attention to the couple flights dimension, writing significantly less about this dimension when compared to couple and family leisure passengers.

Regarding the type of cabin flown criterion, Fig. 7 shows that customer service is significantly more important to passengers flying in first class than to other passengers. On the other hand, economy class passengers write less about this dimension. Economy class passenger write significantly more about delays, checking luggage, waiting times, and prices in airports. In addition, economy premium class passengers give significantly more attention to couple flights and comfort dimensions.

4.3. Sentiment analysis

To identify, extract and calculate sentiment scores of adjectives that are normally used by airline customers, the dataset was reduced to contain only reviews with a general score of 1 (representing very negative sentiment) and 10 (representing very positive sentiment). A Naïve Bayes Classifier was fitted to the reduced dataset, yielding a training accuracy of 93.68%. Considering that only adjectives present in at least 0.1% of reviews were kept, the final number of adjectives analyzed was 882. To facilitate interpretations, each adjective's Naïve Bayes Classifier score was rescaled to vary from −1 (very negative sentiment) to +1 (very positive sentiment). Table 5 presents the top 20 adjectives for negative and positive sentiments, along with their respective sentiment scores. Note that the top 16 negative adjectives present a sentiment score of −1, corresponding to the most negative sentiment.

Table 5
Top 20 adjectives representing negative and positive sentiments.

Negative sentiments		Positive sentiments	
Term	Sentiment score	Term	Sentiment score
absurd	−1.00	good	1.00
abysmal	−1.00	excellent	0.93
dismissive	−1.00	great	0.91
disrespectful	−1.00	comfortable	0.91
incompetent	−1.00	friendly	0.88
incorrect	−1.00	nice	0.82
inexperienced	−1.00	new	0.82
not_alternative	−1.00	clean	0.79
not_clear	−1.00	best	0.77
stale	−1.00	helpful	0.76
stranded	−1.00	free	0.74
unapologetic	−1.00	attentive	0.74
uncaring	−1.00	professional	0.73
unfair	−1.00	pleasant	0.73
unorganized	−1.00	efficient	0.73
unsympathetic	−1.00	fantastic	0.67
unacceptable	−0.60	better	0.65
broken	−0.57	easy	0.65
inadequate	−0.55	quick	0.63
unwilling	−0.55	wonderful	0.61

4.4. Regression analysis

The dimensions of satisfaction and adjectives extracted from OCRs were tested and validated using a Logistic Regression Classifier to predict the recommendation of airlines. After data preparation, the dataset was divided in training and testing sets to allow k-fold cross-validation, with k set to 10. The mean accuracy of the testing set was 79.95% (95% CI: 79.93–79.97). The most relevant dimensions for the prediction of airline recommendation were calculated and are presented in Table 6. The top three dimensions with highest coefficients were “cabin staff” (8.58), “onboard service” (7.77), and “value for money” (6.24); the bottom three dimensions with lowest coefficients were “checking luggage” (0.46), “connections information” (1.39), and “flight description” (2.11).

5. Discussion

Identifying the dimensions of customer satisfaction is determinant to

Table 6
Coefficient of dimensions in the logistic regression decision function.

Dimension	Coefficient	p-value
Cabin staff	8.58	<0.000
Onboard service	7.77	<0.000
Value for money	6.24	<0.000
Seats	6.20	<0.000
Couple flights	5.85	<0.000
Passengers demands	5.51	<0.000
Airplane characteristics	5.20	<0.000
Airport lounge	5.14	<0.000
Customer service	4.24	<0.000
Food & drinks	4.21	<0.000
Onboard entertainment	4.17	<0.000
Communication	4.16	<0.000
Gate service	3.91	<0.000
Business Class	3.57	<0.000
Delays	3.54	<0.000
Comfort	3.19	<0.000
Family flights	3.12	<0.000
Problems	2.90	<0.000
Terminal infra-structure	2.77	<0.000
Waiting times	2.48	<0.000
Check-in procedures	2.46	<0.000
Prices in airports	2.30	<0.000
Passengers interaction	2.29	<0.000
Airline prices	2.16	<0.000
Flight description	2.11	<0.000
Connections information	1.39	<0.000
Checking luggage	0.46	0.006

accurately evaluate how passengers appraise airline services (Park et al., 2004). Using spontaneous user-generated content (UGC) is believed to lower the biases of artificial responses given by customers to traditional research tools, such as focus groups and questionnaire surveys.

To analyze UCGs from customers of the airline industry, we have employed a highly reliable method to identify the dimensions of their satisfaction – the Latent Dirichlet Allocation (LDA). LDA is a topic detection method from the field of machine learning and natural language processing. It is considered to be one of the main methods to extract latent topics from unstructured texts. There is no evidence, however, of its use to analyze customer satisfaction dimensions in the airline industry.

LDA determines the dimensions of customer satisfaction in a dataset. In order to identify the number of dimensions, which was not known *a priori*, we have used the perplexity-based method in our set of UCGs. Perplexity is a measure in information theory that is used to evaluate how well a statistical model describes a dataset. After identifying the number of dimensions, the LDA model has been fitted, based on the whole dataset. Important attributes of each dimension were then identified, grouping words related to them. The bias of this grouping process is low, since it is based on spontaneous experience reports from customers (UGCs). We used the Naïve Bayes Classifier to analyze the sentiment scores of adjectives, labeling them as positive or negative. Finally, we have used a Logistic Regression Classifier to predict the recommendation of airlines. The regression analysis allowed us to validate the dimensions of customer satisfaction and adjectives as predictors of recommendation.

Guo et al. (2017) carried out a similar analysis using UGCs from TripAdvisor. Reviewers rate hotels regarding five dimensions, in addition to an overall score, and produce a free text comment on their experience. A regression model relating overall score with scores from the five dimensions was obtained, yielding a 0.68 coefficient of determination; our logistic regression classifier yielded a 79.95% accuracy, although results from the two models are not directly comparable. The authors compare customer satisfaction dimensions obtained through LDA with those obtained through surveys and focus groups in similar studies, to conclude that their results are more comprehensive (i.e., a larger number of dimensions is found) and generalizable, due to the larger sample size. Based on the works reviewed in the background section, the same conclusions apply here.

In previous studies addressing airline passenger satisfaction reviewed in section 2, the context of use of adjectives was not evaluated. Our results represent a step forward in analyzing UGCs, since they are segmented by service characteristics and customer profile. Service characteristics (airline, publication year, and type of cabin flown) and customer profiles (type of passenger and passenger nationality) were related to users' levels of satisfaction and recommendation of airlines.

Comparing results by airline, differences in distribution of some dimensions of customer satisfaction were identified. Even though general dimensions of satisfaction were drawn from the dataset, this result suggests that our method is sensitive to detect dissimilarities between different services. Likewise, equivalent services (e.g., from American Airlines and United Airlines) are also identifiable through this method. The ability to detect similarities and differences in services indicates that our method does not only generate general results to be used in academic research, but it also works as a framework for airline service providers to apply in specific companies.

Our method is also sensitive to show customer trends through time, by observing results in different publication years. On one hand, decreasing attention has been given by customers to flight descriptions, food and drinks, and airplane characteristics in recent years. On the other hand, dimensions such as passenger interaction and checking in luggage are crescent focuses of the reviews observed in UGCs. These results indicate that companies are aware of and aligned with customers' needs regarding some of the dimensions of satisfaction. Nevertheless, important issues are being more and more neglected, such as luggage restrictions. Even if companies choose not to hear the "wisdom of crowds" (e.g., for profit purposes), our results can raise their awareness regarding the relevance of the neglected dimensions of customer satisfaction. Therefore, airline service providers may make informed decisions about their services.

The type of cabin flown has shown to influence the distribution of the dimensions of satisfaction. This result allows airline service providers to focus their efforts on dimensions that are relevant to segmented types of customers (e.g., customer service to passengers that fly in first class, comfort to those who fly premium economy, and checking luggage and waiting time to users flying economy class). Differently, type of passenger (users flying on business trips, solo leisure, and couples and families) was the segmentation variable which impacted the smallest number of dimensions of customer satisfaction.

Passenger nationality has also influenced the distribution of the dimensions of satisfaction. Apart from the evident use of this information to shape services locally, results have shown that our method was able to identify cultural differences. Since it is possible to map passengers' nationalities by route using purchase information available in the airlines' databases, it is possible to adjust service to meet specific preferences of passengers; e.g., in routes where Australian passengers prevail special attention should be given to the offer of foods and drinks (dimension #3 in our analysis), which is not predominant in routes mostly flown by American passengers. Even though some dimensions are culture-sensitive, as seen in Fig. 5, passengers write similarly about seats, communication, check-in procedures, comfort, and terminal infrastructure.

Analyzing Fig. 6 we conclude that passengers traveling on business trips demonstrate greater interest in the service dimension (dimension #1). Certain routes are known to be flown by business travelers, particularly at certain times; e.g. early morning and late evening flights from LGA to IAD. Companies may customize services to better meet expectations of those passengers, using technology to expedite the check-in of passengers typically traveling only with carry-on luggage, customizing food and beverage menus to meet their expectations, and assigning experienced staff to provide better in-flight services. Such initiatives would also be aligned with results from the regression analysis, in which Cabin Staff e Onboard Service are the variables with highest coefficients.

Regarding sentiment scores of adjectives, only “good” hit the highest level possible (+1) according to our measurement method (Naïve Bayes Classifier). This word is not specific of any domain, such as caring behavior or level of experience of cabin crew. Differently, many negative sentiments hit the lowest grade possible (−1) and they refer to specific themes: overall experience (absurd, abysmal, not_alternative, not_clear), uncaring behavior (dismissive, disrespectful, unapologetic, uncaring, unfair), lack of knowledge or experience of service providers (incompetent, incorrect, inexperienced, unorganized, unsympathetic) and general updateness (stale, stranded). The observation of the strength of these negative sentiments offer many inputs to design services to avoid negative experiences.

It is also possible to determine sentiment scores for individual LDA dimensions for each airline. That enables managers to (i) position the performance of their company with respect to direct competitors, (ii) establish benchmarks for each dimension, and (iii) investigate the reasons why competitors perform better in given dimensions, using that information to promote improvements. Analyzing regression coefficients, airline managers may also prioritize improvements in specific LDA dimensions.

Finally, positive reviews and recommendations from customers are commonly as results of customer satisfaction (Forgas et al., 2010; Guo et al., 2017; Mattila, 2004; Morgan and Hunt, 1994). Observing the UGCs in our dataset, the top three dimensions of satisfaction to predict the recommendation of airlines were cabin staff, onboard service, and value for money. Designing services that meet or surpass expectations within these dimensions is key to recommendation (Chow, 2015).

6. Delimitations

There are some delimitations to this study. They are presented as follows, including a future agenda for research in the field of airline services.

First, the analysis is restricted to UGCs and no relations between these reviews and data generated by conventional methods (e.g. survey questionnaires) are made. Therefore, future studies may relate both types of data to reinforce the belief that analyzing UGC provides more accurate results than other research tools to predict airline recommendation.

Second, we have adopted a single method of topic detection from the field of machine learning and natural language processing (Latent Dirichlet Allocation – LDA). The application of other methods, such as Paragraph Vectors (Hashimoto et al., 2016) and Hierarchical Recurrent Neural Network (Lu et al., 2017), offers an opportunity for future research.

Third, in our study the measure of satisfaction is the recommendation of airline services. Further measures may be adopted in future research, since we did not compare customers’ pre-purchase expectations and post-purchase evaluations (Chow, 2015; Forgas et al., 2010; Guo et al., 2017), and the recurrence of consumption was also not assessed (Forgas et al., 2010; Guo et al., 2017; Mattila, 2004; Morgan and Hunt, 1994).

Fourth, only reviews written in English were present in the dataset. That restricts the diversity of opinions gathered in the sample, since only respondents proficient in English are likely to provide information on airline experience. The analysis of a multilingual dataset may be implemented as future research.

Finally, we did not stratify the sample based on combinations of nationality and airline, which may be justified by the large number of nationalities (171) and airlines (419) in the sample. However, it is possible that national preferences create biases towards certain airlines; that is a topic for future investigation. Our sample also lacked detailed information on respondents’ demographics and there is a possibility that it does not properly represents the entire population. That could not be tested in this study, but we believe that the large sample size is likely to attenuate unbalances in the characterization of the population of respondents.

7. Concluding summary

In this paper we present a novel text mining-based approach to explore dimensions of airline customer satisfaction from the analysis of Online Customer Reviews (OCRs), which are an abundant and low-cost source of information. The proposed method is applied to a large database of airline reviews, comprised of 55,775 supervised OCRs, in which respondents recommended or do not recommended the airline after describing their experience. We designed the study such that respondents are stratified according to several meaningful variables.

In addition to identifying attributes that characterize positive and negative reviews, we also identify dimensions of satisfaction, using the text equivalent of a factor analysis – a Latent Dirichlet Allocation (LDA) model. We perform analyses using the entire dataset and some of its relevant strata. We were able to predict airline recommendation by customers with an accuracy of 79.95%.

Our paper explores an existent gap in the literature regarding retrieval and information mining from OCRs in the airline industry. This research represents a step forward in measuring customer satisfaction. Our method allows extracting dimensions of satisfaction from UGCs (text), providing results similar to factors that would be obtained through factor analysis (quantitative data). It is an innovative way of measuring customer satisfaction that is suitable for future academic and applied studies.

Referencesd

Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P., 2008. Mixed membership stochastic blockmodels. J. Mach. Learn. Res. 9, 1981–2014. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted, 2008.

Ali, F., Kim, W.G., Ryu, K., 2016. The effect of physical environment on passenger delight and satisfaction: moderating effect of national identity. Tour. Manag. 57, 213–224. https://doi.org/10.1016/j.tourman.2016.06.004.

Basfirinci, C., Mitra, A., 2015. A cross cultural investigation of airlines service quality through integration of Servqual and the Kano model. J. Air Transp. Manag. 42, 239–248. https://doi.org/10.1016/j.jairtraman.2014.11.005.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato Recupero, D., Subrahmanian, V.S., 2005. Sentiment analysis: adjectives and adverbs are better than adjectives alone. In: ICWSM. https://pdfs.semanticscholar.org/1133/455c8b743073160dd439361c4aaf8de0faad.pdf.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python, 1º edition. O’Reilly Media, Sebastopol, CA.

Blei, D.M., Jordan, M.I., 2003. Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM, pp. 127–134.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Blodgett, J.G., Li, H., 2007. Assessing the effects of post-purchase dissatisfaction and complaining behavior on profitability: a Monte Carlo simulation. J. Consumer Satisfaction, Dissatisfaction Complain. Behav. 20, 1–14.

Calisir, N., Basak, E., Calisir, F., 2016. Key drivers of passenger loyalty: a case of Frankfurt-Istanbul flights. J. Air Transp. Manag. 53, 211–217. https://doi.org/10.1016/j.jairtraman.2016.03.002.

Chau, M., Xu, J., 2012. Business intelligence in blogs: understanding consumer interactions and communities. MIS Q.: Manag. Inf. Syst. 36 (4), 1189–1216. https://doi.org/10.1020/18574872234871806.

Chen, C.-F., 2008. Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: evidence from Taiwan. Transp. Res. Part A 42 (4), 709–717. https://doi.org/10.1016/j.tra.2008.01.007.

Chow, C.K.W., 2015. On-time performance, passenger expectations and satisfaction in the Chinese airline industry. J. Air Transp. Manag. 47, 39–47. https://doi.org/10.1016/j.jairtraman.2015.04.003.

Dolnicar, S., Grabler, K., Grün, B., Kulnig, A., 2011. Key drivers of airline loyalty. Tour. Manag. 32 (5), 1020–1026. https://doi.org/10.1016/j.tourman.2010.08.014.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification, second ed. Wiley-Interscience, New York.

Eboli, L., Mazzulla, G., 2009. An ordinal logistic regression model for analysing airport passenger satisfaction. EuroMed J. Bus. 4 (1), 40–57. https://doi.org/10.1108/14502190910956684.

Felbermayr, A., Nanopoulos, A., 2016. The role of emotions for the perceived usefulness in online customer reviews. J. Interact. Mark. 36, 60–76. https://doi.org/10.1016/j.intmar.2016.05.004.

Feldmann, R., Sanger, J., 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge university press, New York.

Flanagin, A.J., Metzger, M.J., 2013. Trusting expert- versus user-generated ratings online: the role of information volume , valence , and consumer characteristics. Comput. Hum. Behav. 29 (4), 1626–1634. https://doi.org/10.1016/j.chb.2013.02.001.

Forgas, S., Moliner, M.A., Sánchez, J., Palau, R., 2010. Antecedents of airline passenger loyalty: low-cost versus traditional airlines. J. Air Transp. Manag. 16 (4), 229–233. https://doi.org/10.1016/j.jairtraman.2010.01.001.

Gaskin, S.P., Griffin, A., Hauser, J.R., Katz, G.M., Klein, R.L., 1993. The Voice of the customer. Mark. Sci. 12 (1), 1–27. https://doi.org/10.1287/mksc.12.1.1.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U. S. A 101 (Suppl. l), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Grigoroudis, E., Siskos, Y., 2009. Customer Satisfaction Evaluation: Methods for Measuring and Implementing Service Quality. Springer Science & Business Media.

Guo, Y., Barnes, S.J., Jia, Q., 2017. Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. Tour. Manag. 59, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009.

Hackeling, G., 2014. Mastering Machine Learning with Scikit-Learn. Packt Publishing Ltd, Birmingham, UK.

Hashimoto, K., Kontonatsios, G., Miwa, M., Ananiadou, S., 2016. Topic detection using paragraph vectors to support active learning in systematic reviews. J. Biomed. Inform. 62, 59–65. https://doi.org/10.1016/j.jbi.2016.06.001.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer.

Hoffman, M.D., Blei, D.M., Bach, F., 2010. Online learning for latent dirichlet allocation. Adv. Neural Inf. Process. Syst. 23, 856–864.

Hu, M., Liu, B., 2004. Mining and Summarizing Customer Reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining 168–177.

Jin, J., Ji, P., Gu, R., 2016. Identifying comparative customer requirements from product online reviews for competitor analysis. Eng. Appl. Artif. Intell. 49, 61–73. https://doi.org/10.1016/j.engappai.2015.12.005.

Josephat, P., Ismail, A., 2012. A logistic regression model of customer satisfaction of airline. Int. J. Hum. Resour. Stud. 2 (4), 255. https://doi.org/10.5296/ijhrs.v2i4.2868.

Lee, T.Y., Bradlow, E.T., 2007. Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews.

Li, W., Yu, S., Pei, H., Zhao, C., Tian, B., 2017. A hybrid approach based on fuzzy AHP and 2-tuple fuzzy linguistic method for evaluation in-flight service quality. J. Air Transp. Manag. 60, 49–64. https://doi.org/10.1016/j.jairtraman.2017.01.006.

Liau, B.Y., Tan, P.P., 2014. Gaining customer knowledge in low cost airlines through text mining. Ind. Manag. Data Syst. 114 (9), 1344–1359. https://doi.org/10.1108/IMDS-07-2014-0225.

Liou, J.J.H., Tzeng, G.-H., 2007. A non-additive model for evaluating airline service quality. J. Air Transp. Manag. 13 (3), 131–138. https://doi.org/10.1016/j.jairtraman.2006.12.002.

Liou, J.J.H., Hsu, C.-C., Yeh, W.-C., Lin, R.-H., 2011. Using a modified grey relation method for improving airline service quality. Tour. Manag. 32 (6), 1381–1388. https://doi.org/10.1016/j.tourman.2011.01.013.

Lu, H., Xie, L., Kang, N., Wang, C.-J., Xie, J.-Y., 2017. Don't forget the quantifiable relationship between Words : using recurrent neural Network for short text topic discovery. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 1192–1198. AAAI-17.

Lucini, F.R., Fogliatto, F.S., da Silveira, G.J.C., Neyeloff, J., Anzanello, M.J., Kuchenbecker R, de S., Schaan, B.D., 2017. Text mining approach to predict hospital admissions using early medical records from the emergency department. Int. J. Med. Inform. 100, 1–8. https://doi.org/10.1016/j.ijmedinf.2017.01.001.

Mattila, A.S., 2004. The impact of service failures on customer loyalty. Int. J. Serv. Ind. Manag. 15 (2), 134–149. https://doi.org/10.1108/09564230410532475.

Morgan, R.M., Hunt, S.D., 1994. The commitmemt-trust theory of relationship marketing. J. Mark. 58 (3), 20–38.

Mostafa, M.M., 2013. An emotional polarity analysis of consumers' airline service tweets. Soc. Netw. Anal. Min. 3 (3), 635–649. https://doi.org/10.1007/s13278-013-0111-2.

Namukasa, J., 2013. The influence of airline service quality on passenger satisfaction and loyalty. TQM J. 25 (5), 520–532. https://doi.org/10.1108/TQM-11-2012-0092.

Park, J.-W., Robertson, R., Wu, C.-L., 2004. The effect of airline service quality on passengers' behavioural intentions: a Korean case study. J. Air Transp. Manag. 10 (6), 435–439. https://doi.org/10.1016/j.jairtraman.2004.06.001.

Park, J.-W., Robertson, R., Wu, C.-L., 2005. Investigating the effects of airline service quality on airline image and passengers' future behavioural intentions: findings from Australian international air passengers. J. Tour. Stud. 16 (1), 2–11. https://doi.org/10.1375/jhtm.13.2.161.

Python Software Foundation, 2017. The Python Language Reference. Retrieved from. https://www.python.org/.

Shivashankar, S., Srivathsan, S., Ravindran, B., Tendulkar, A.V., 2011. Multi-view methods for protein structure comparison using latent dirichlet allocation. Bioinformatics 27 (13), 61–68. https://doi.org/10.1093/bioinformatics/btr249.

Sparks, B.a., So, K.K.F., Bradley, G.L., 2016. Responding to negative online reviews: the effects of hotel responses on customer inferences of trust and concern. Tour. Manag. 53, 74–85. https://doi.org/10.1016/j.tourman.2015.09.011.

Sreenivasan, N.D., Lee, C.S., Goh, D.H., 2012. Tweeting the friendly skies: investigating information exchange among Twitter users about airlines. Program 46 (1), 21–42. https://doi.org/10.1108/00330331211204548.

Surowiecki, J., Silverman, M.P., 2007. The wisdom of crowds. Am. J. Phys. 75 (2), 190–192.

Tirunillai, S., Tellis, G.J., 2014. Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation. J. Mark. Res. 51 (4), 463–479. https://doi.org/10.1509/jmr.12.0106.

Tsafarakis, S., Kokotas, T., Pantouvakis, A., 2017. A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. J. Air Transp. Manag. 1–15. https://doi.org/10.1016/j.jairtraman.2017.09.010.

Wan, Y., Gao, Q., 2015. An ensemble sentiment classification system of twitter data for airline services analysis. In: Proceedings - 15th IEEE International Conference on Data Mining Workshop. ICDMW, pp. 1318–1325. https://doi.org/10.1109/ICDMW.2015.7.

Xu, X., Li, Y., 2016. The antecedents of customer satisfaction and dissatisfaction towardvarious types of hotels: A text mining approach. Int. J. Hosp. Manag. (55), 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003.

Xu, X., Liu, W., Gursoy, D., 2018. The impacts of service failure and recovery efforts on airline customers' emotions and satisfaction. J. Travel Res. 1–18. https://doi.org/10.1177/0047287518789285.

Yao, B., Yuan, H., Qian, Y., Li, L., 2015. On exploring airline service features from massive online review. In: 2015 12th International Conference On Service Systems And Service Management. ICSSSM. https://doi.org/10.1109/ICSSSM.2015.7170270. *2015*.

Zhang, X., Jeske, D., Li, J., W, V., 2016. A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. Comput. Stat. Data Anal. 94, 238–249. https://doi.org/10.1016/j.csda.2015.08.009.

Zhao, W., Zou, W., Chen, J.J., 2014. Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinf. 15 (Suppl. 1), S11. https://doi.org/10.1186/1471-2105-15-S11-S11.

Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W., 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinf. 16 (Suppl. 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8.