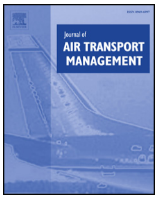
Journal of Air Transport Management 71 (2018) 28–44

Contents lists available at ScienceDirect

Journal of Air Transport Management

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



Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews

Kiljae Lee[*], Chunyan Yu

Embry-Riddle Aeronautical University, USA

A R T I C L E I N F O

Keywords:
Airport service quality
Google maps
Text mining
LDA
Sentiment analysis

A B S T R A C T

The purpose of this paper is to demonstrate that user-generated online contents can be used as an alternative data source for assessing airport service quality, which effectively complements and cross-validates the conventional service quality surveys. We apply sentiment analysis and topic modeling technique to 42,137 reviews collected from Google Maps. The results are compared to the well-publicized ASQ ratings conducted by Airport Council International. The sentiment scores computed from the textual Google reviews are very good predictors of the associated Google star ratings, with $r_s(96) = 0.89$, $p < .01$ in 2016. The correlation could be further improved ($r_s(96) = 0.90$) by customizing the sentiment lexicon leveraging the information gained from the previous year's analysis. Also, both the sentiment scores and Google star ratings are found to have a reasonably strong association with the ASQ ratings, with $r_s(78) = 0.63$, $p < .01$ and $r_s(78) = 0.64$, $p < .01$, respectively, in 2016, excluding outliers. These results indicate that the online reviews provide a good proxy for airport service quality ratings and an effective means to cross-validate the conventional industry standard survey results. Further, the study extracts 25 latent topics from the Google reviews through a topic modeling analysis. The 25 topics show good correspondence with the ASQ service attributes, suggesting that the ASQ program effectively covers all the service quality attributes of airport users. Also, further analysis indicates that the relative importance of service attributes varies depending on the size of the airports and that some ASQ service attributes may not be relevant anymore for most passengers.

1. Introduction

Airports are in the service industry; thus service quality is essential to airport operation and management. Airports strive to meet the needs of their customers including passengers, shippers, and airlines. Passenger's perception of service quality and their level of satisfaction have become important indicators of airports' performance, and are measured through surveys conducted either internally or externally. One of the most publicized passenger satisfaction surveys is Airports Council International's (ACI) Airport Service Quality program (hereinafter ASQ), which was initiated in 2006. At present, 320 airports across 80 countries[1] participate in the ASQ survey. The ACI-ASQ program conducts quarterly in-person questionnaire surveys of sample passengers at participating airports and requires a minimum of 350 onsite survey participants per quarter (1400 per year) at each airport. Following a strict plan developed by ACI, the staff at participating airports or third-party companies conduct surveys at the airports with a standardized questionnaire. Regular audits are undertaken by ACI to ensure compliance. The survey rates airport performance by 34 service attributes in eight categories including access, check-in, passport control, security, navigation, facilities, environment, and arrival. ACI recognizes the best airports by size and by region in the annual ASQ awards based on the survey results. The awards are extensively cited by the winning airports for promotional purpose. However, access to the survey results are limited to the participating airports,[2] and the general public and non-participating airports can only see the list of winners in various award categories (by size and region).

With the growing popularity of web-based opinion platforms, passengers increasingly offer voluntary reviews of airport services on various platforms, such as Skytrax (www.airlinequality.com), TripAdvisor, and Google reviews. These platforms let travelers leave a star-rating along with reviews about various aspects of airport services. These ratings and reviews are great sources of information for both travelers and service providers. However, past reviews are quickly

[*] Corresponding author. College of Business, Embry-Riddle Aeronautical University, 600 S. Clyde Morris Blvd., Daytona Beach, FL, 32114, USA.
E-mail addresses: kiljae.lee@erau.edu (K. Lee), yuc@erau.edu (C. Yu).

[1] http://www.aci.aero/Customer-Experience-ASQ/ASQ-Awards, accessed on August 15, 2017.

[2] According to private communications with managers at a participating airport, each participating airport pays a roughly $16,000 to $20,000 annual fee, and spends additional $50,000 to $100,000 to carry out the survey.

https://doi.org/10.1016/j.jairtraman.2018.05.004
Received 23 August 2017; Received in revised form 17 May 2018; Accepted 20 May 2018
0969-6997/ © 2018 Elsevier Ltd. All rights reserved.

buried in a massive amount of newer reviews, and the single consolidated rating scores depicting each airport reflects all ratings accumulated over the years without distinguishing service attributes or explicating changes over time. That is, travelers and service providers would not be able to fully utilize the rich and valuable information embedded in these ratings and reviews in their raw format.

The recent development in data mining (i.e., techniques of discovering patterns and trends from a large data set) (Larose, 2005) and text mining (i.e., a subset of data mining that aims at extracting information from texts) (Zhong et al., 2012) provides various means to analyze these rating and reviews. These analyses can offer complementary indicators to cross-validate the ASQ's survey results and to extract the key attributes of service quality perceived by passengers that can be compared to those from ASQ surveys, and to expand the coverage of analysis beyond the 320 participating airports in the ASQ program.

Although airports have been among the leaders in applying social media, mobile and digital technology in delivering services and communicating with customers, there has been limited research in the field of airport management that investigate the growing contents on the aforementioned platforms. This paper is intended as a prefatory step to fill this gap.

The objective of this paper is to demonstrate a complementary assessment approach that can be used: 1) to evaluate passengers' perception of airport service quality, 2) to cross-validate ASQ's survey results, and 3) to examine the degree to which ASQ's service attributes match the service attributes expressed in Google reviews. To achieve the objective, the study performs pairwise comparisons between Sentiment Scores extracted from Google reviews, Google star ratings, and ASQ ratings. Next, the study extracts major topics from the textual reviews and compares them with the ASQ's survey attributes. The paper shows that Google reviews provide a rich source of data to develop airport service quality indicators that complement and extend ASQ surveys.

The rest of the paper is organized as follows: Section 2 reviews literature in airport service quality and provides an overview of studies that use text mining in airport management and related fields; Section 3 describes our methodology for selecting a platform, collecting reviews, and extracting metrics to be used for comparison with ASQ's ratings and survey attributes; Section 4 analyzes the data and discusses the results; Section 5 discusses some specific issues related to using Google Maps as data source; Section 6 summarizes the main findings and offers concluding remarks.

2. Literature review

2.1. Service quality as a construct

Service quality is considered one of the most debated topics in the service marketing literature. Yet, there is one thing the researchers appear to agree on: perceptions of service quality are based on multiple dimensions, even though there is no general agreement as to the nature or content of the dimensions (Brady and Cronin, 2001). Inherently, airport service quality is a multi-dimensional construct that represents a broad range of passenger experiences. As Pantouvakis and Renzi (2016) pointed out that there are two general tracks of airport service quality literature: (1) to identify the different dimensions or attributes of airport service quality through conceptual or empirical modeling; (2) to identify quality drivers that lead to customers' satisfaction.

Rhoades et al. (2000) used factor analysis to identify four dimensions that contain twelve attributes, including passenger service (food and beverage, restrooms, retail and duty free, special services), airport access (parking, rental car, ground transportation), airline-airport interface (gate boarding areas, baggage claim, information display), and the inter-terminal transportation as a single attribute dimension. The study surveyed 150 airport directors and consultants through a mail questionnaire. Yeh and Kuo (2003) identified six “manageable” service attributes (Comfort, Processing time, Convenience, Courtesy of staff, Information visibility, and Security) through a panel of experts, then applied a fuzzy multi-attribute decision making (MADM) model to generate a service quality index to evaluate the comparative level of passenger service performance among 14 Asia Pacific airports. The study is limited to the attributes that are “manageable by the airport” rather than “all” service attributes experienced by passengers.

Emphasizing the importance of passengers' perspective, Fodness and Murray (2007) proposed that passengers' expectation of airport service quality has three key dimensions with five subdimensions including function (effectiveness, efficiency), interaction, and diversion (maintenance, productivity, decor). They conducted an empirical test of the model based on 700 responses collected through a mail survey. Using factor analysis (Bezerra and Gomes, 2015), extracted seven dimensions of airport service quality as perceived by the passengers, and then examined how each of these dimensions affects passengers' overall satisfaction. Excluding one of the extracted dimensions (i.e., Price), Bezerra and Gomes (2016) estimated a six-factor model to measure airport service quality (Check-in, Security, Mobility, Ambience, Basic facilities, and Convenience). Both studies were based on on-site survey data collected at one Brazilian airport. Their results may not be generalizable to all airports as the perceived service quality is subjective and context-dependent (Brady and Cronin, 2001). This notion of context dependency was further examined by Pantouvakis and Renzi (2016). Pantouvakis and Renzi (2016) collected 922 usable responses through in-terminal personal interviews at Rome Fiumicino Airport over a two month period in 2014. They identified three “distinct, independent and invariant” service quality dimensions, namely, “Servicescape and Image,” “Signage” and “Service.” Further, their empirical results provided some evidence that passengers' satisfaction or dissatisfaction perception of airport service quality vary according to their nationalities. In other words, the perception of airport service quality is context dependent.

There have been efforts to identify various drivers that influence passengers' perception of airport service quality. Suárez-Alemán and Jiménez (2016) investigated whether passengers' perception of airport quality is influenced by airport management schemes and characteristics that are not directly observable. Their results indicate that airport ownership, the degree of regulation, level of GDP per capita are among the drivers of airport service quality. Brida et al. (2016) examined the effects of information and communication technologies on passengers' perception of airport service quality. Based on a survey conducted by Chilean Aviation Authority at Santiago International Airport (SCL) in 2013, the study found that factors related to flights and airport information have an important impact on the passengers' perception of airport services.

As the discussions above attest, there is no established consensus on the dimensions and attributes of airport service quality. Further, as shown by the previous studies, the airport service quality is a context-dependent construct. These suggest that examining the consistency between ASQ's service quality attributes and the collective perception of service quality is an empirical question. The present study extends the service quality literature, not by proposing yet another set of service dimensions but by empirically comparing ASQ ratings against the ratings from a large number of online reviews, and comparing ASQ service attributes against the topics extracted from those reviews that reflect passengers' collective experiences.

2.2. Text mining and sentiment analysis in airport management and related fields

The ever-increasing volume of comments and reviews on the Internet offers new opportunities to capture passengers' perceptions and expectations of airport service quality on a global scale. Recent research in text mining and data science makes this possible through various computational, algorithmic alternatives. There has been a recent uprush of endeavors exploring the possibilities of using (often unstructured) information from online resources to investigate travel and transport related issues. For example, Költringer and Dickinger (2015) presented an automatic web content mining approach to compare destination branding information from 1) the official websites of destination marketing organizations, 2) local news media sites, and 3) UGC (User Generated Contents) from review pages and travelers' blogs. Using Vienna as a destination example, they concluded that UGC is the richest and most diverse source of information that reveals a variety of aspects of the city that often are not covered by official websites and news media sites. Kuflik et al. (2017) proposed a generic framework for automatizing the procedure for collecting, filtering, classifying, and semantically processing transport-related tweets. The study found that transport-related customer reviews expressed on social media tend towards negative sentiments. The paper illustrates some of the challenges that need to be addressed in developing an automatic analysis of social media data and when aggregating micro-level social media data concerning transport experiences. Twitter has been extensively used as the source of information for detecting traffic accidents (Mai and Hranac, 2013), improving predictions on traffic incidents (Grosenick, 2012; Gu et al., 2016), disclosing traffic congestion and surge (D'Andrea et al., 2015; Zhang et al., 2016), and extracting sentiment to assess transit riders' satisfaction of public transportation systems (Collins et al., 2013). One of the major challenges in using Twitter data in a systematic framework is to improve the precision of filtering out irrelevant tweets from the massive amount of tweet messages.[3]

Sentiment analysis approach on customer reviews has been particularly thriving in the field of hospitality and tourism management. Sentiment analysis is one of the popular text mining approaches that focus on identifying positive and negative opinions, emotions, and evaluations from the large textual data set (Wilson et al., 2005). For example, Xiang et al. (2015) collected 60,648 hotel customer ratings and reviews from Expedia.com. Through sentiment analysis, they demonstrated that guest experience attributes are significantly associated with hotel satisfaction. Xie et al. (2014) collected 4994 reviews of 843 hotels in five major cities in Texas and showed that overall ratings, ratings of “value for money,” variation and volume of reviews, and the number of management responses are all significantly associated with hotel performance. After reviewing 22 recent studies, Xiang et al. (2017) noted that most researchers would elect one of the three popular websites (i.e., TripAdvisor, Expedia, and Yelp) as their main data source. It is interesting to note that those three platforms are considerably different from each other, and do not represent the same population. Xiang et al. (2017) show that the three platforms had significant discrepancies in the representation of hotel industry, and in terms of review length, review topics, review sentiment, and helpfulness of the reviews using sentiment analysis and topic modeling (i.e., algorithms that automatically summarize large archives of texts by discovering hidden topics) (Blei and Lafferty, 2009). The platform bias (i.e., the bias associated with using a single platform as a source of data) is the most oft-cited methodological limitation of analyzing online contents. This calls for triangulation, comparing results using multiple data sources, as an essential strategy for improving external validity (Ruths and Pfeffer, 2014; Tufekci, 2014).

In the field of airport management, the earliest attempt to text-mine the online reviews was targeted at evaluation of airport service quality (Bogicevic et al., 2013). Bogicevic et al. (2013) conducted a content analysis on 1095 reviews posted on Skytrax (www.airlinequality.com) between 2010 and 2013. The study identified 14 airport service quality attributes based on the frequencies of words in the reviews. They conceptualized that the words that appear most frequently together with high star ratings as satisfiers, and the words that appear most frequently with low ratings as dissatisfiers. They showed that security check, signage, and dining options are dissatisfiers whereas cleanliness and shopping option are satisfiers. Wattanacharoensil et al. (2017) also collected reviews on airports from Skytrax, choosing the top five, the middle five, and the last five from the ranking of Skytrax's best 100 airports worldwide in 2014. By manually cross-tabulating 647 reviews, they found that travelers' “experience outcome” (especially in emotional and memory aspects of air travelers) is determined primarily by functional experience (i.e., basic/fundamental processes in airport) and service personnel; travelers do not appreciate any additional experience (i.e., from aesthetic and hedonic activities) when these fundamental processes and human interactions are not satisfactory.

Gitto and Mancuso (2017) conducted sentiment analysis of passengers' perception of service quality at the five largest airports in Europe based on 895 sentences selected from reviews posted on Skytrax between September 2013 to February 2014. Their results suggested that passengers' perception of non-aviation services is mostly influenced by food & beverage and the shopping area, and their perception of aviation services is mostly influenced by check-in, baggage claim, and security control procedures. Nghiêm-Phú and Suter (2018) collected and analyzed passenger reviews of Las Vegas McCarran International Airport on TripAdvisor to identify the attributes that are associated with the image of the airport. Based on a usable sample of 427 reviews, the study found that McCarran International Airport shares most of the common airport service attributes identified by previous studies, but also has several attributes unique to its host city, Las Vegas. The authors claimed that the airport had played an active role in creating visitors' first and last impressions of the city of Las Vegas.

Most of the past studies utilizing text mining and sentiment analysis rely on reviews and contents collected from a single website. In the field of airport management, the majority of such studies have focused on Skytrax and occasionally on TripAdvisor, none of these studies attempt to compare their results with other assessment measures. As pointed out by (Ruths and Pfeffer, 2014; Tufekci, 2014), the validity of the results from such big data analysis could be improved through triangulation with multi-platform analysis. The present study extends this line of research in airport management in two ways. First, we adopt the much-recommended triangulation approach by comparing the ratings and topics extracted from online reviews with those from ASQ surveys. Second, we introduce a new web-based general opinion platform (i.e., Google Maps) to the literature that has never been investigated by any research in airport management as well as in the related fields.

3. Methodology

This study uses user-generated online contents as an alternative data source for assessing airport service quality. In particular, we apply sentiment analysis and topic modeling technique on reviews collected from Google Maps and compares the results with the well-publicized ASQ ratings and their service attributes. This section explains the rationale for choosing Google reviews as the platform to collect passengers' perceptions of airport service quality and describes the overall analytical framework and the data collection process.

3.1. Target data

The largest volumes of online reviews on airports in English can be found on Twitter, TripAdvisor, Skytrax, and Google Maps. While Twitter provides a convenient Application Protocol Interface (API) to crawl its data, it is not a suitable platform for our purpose because most of the tweets that contain the keyword “airport” or the hashtag “airport” do not include the type of evaluative messages relevant to our research objective; and our chosen topic modeling algorithm (i.e., Latent Dirichlet Allocation) does not perform well with the short

[3] In this work, the authors showed a relative superiority of SVM (Support Vector Machine) to Naïve Bayes or Decision Tree algorithm for the tasks of filtering out irrelevant tweets.

messages like tweets (Mehrotra et al., 2013; Zhao et al., 2011). Reviews posted on TripAdvisor mostly focus on airlines, hotels, restaurants, and local attractions but not exclusively on airports.

While TripAdvisor and Google Maps focus on information exchange among customers, Skytrax focuses more on ranking airports and presenting World Airports Awards. Similar to the ASQ awards, Skytrax awards are often cited by the winner airports for publicity. Skytrax started to accumulate online reviews as early as July 2002, but the growth in the number of the airport reviews has been slow.[4] The numbers of reviews per year per airport are not sufficiently large enough to meet our needs for many airports. For example, only 11 reviews for Hartsfield-Jackson Atlanta International Airport (ATL), 9 reviews for Beijing Capital International Airport (PEK) between January 1 to August 1, 2015. The total number of all airport reviews posted in 2014 was only 2,568, and only 1698 between January 1 to August 1, 2015. More importantly, the site has relatively weak exposure to the general public, even though it is well known to aviation industry professionals. We presume that there is substantially lesser chance for casual visitors than aviation professionals to leave reviews in Skytrax, potentially leading to a strong self-selection bias.

Google reviews, on the other hand, are posted mostly by casual visitors who happen to search the airport in Google or use Google Maps before, during, or after visiting the place. Furthermore, unlike Twitter, Google Maps solicits a quantitative star-rating along with a textual review. This enables us to test the consistency between the sentiments reflected in the textual reviews and their associated star rating scores. Therefore, we chose Google reviews as the primary data source for this study.

3.2. Analytical Framework

The study is carried out in four steps as shown in Fig. 1:

Step 1 examines the consistency between the contents of the textual reviews and the associated star ratings within Google Maps platform. In Google Maps, people write a textual review and give a star rating at the same time. We conduct a sentiment analysis (Liu, 2012) to quantify the reviewers' emotional valence expressed in the textual reviews toward the airport services, and then compare the sentiment scores with the associated star ratings. Since those who give positive reviews generally assign higher star ratings and vice versa, we expect the sentiment scores extracted from the textual reviews to have a highly positive correlation with the associated star ratings. This result will also provide a rationale for performing **Step 4,** which assumes that the textual reviews represent the experienced service attributes that determine the reviewers'ratings.

Step 2 compares the Google star ratings with ASQ ratings. The average annual Google star rating scores of the selected airports are regressed on ASQ ratings to find a general linear trend between these two metrics and investigate outliers.

Step 3 compares the sentiment scores with ASQ ratings. The degree of correlation between the sentiment scores and ASQ ratings is tested. If significant, the results further strengthen the proxy value of the review texts for airport service quality.

Step 4 extracts the dominant topics from the textual reviews using the Latent Dirichlet Allocation (LDA) model proposed by Blei and coauthors (Blei, 2012; Blei et al., 2003). The resulting topics are compared to the 34 ASQ service attributes. The results provide an empirical basis to evaluate the extent to which the topics discussed by passengers in Google reviews are consistent with the service attributes of the ASQ questionnaire. Lastly, the relative weights of the extracted topics among the airports of different sizes are computed to examine whether those topics are equally distributed across all airports of different sizes. Since the perception of airport service quality is context dependent (Brady and Cronin, 2001), we expect the relative weight on the topics to be different depending on the size of the airport.

These four sets of comparisons and analysis allow us to address the following research questions:

Q1. Is the sentiment score a good indicator of Google star ratings?

Q2. Do sentiment scores match ASQ ratings?

Q3. Do Google star ratings match ASQ ratings?

Q4. Do topics extracted from reviews match ASQ's service attributes?

Q5. Do the extracted topics equally distributed across all airports of different sizes?

Q6. How to improve the performance of sentiment analysis over time?

3.3. Data collection

To ensure that the Google review data can be directly compared with the ASQ results, our sample includes the top 100 airports in passenger traffic volume that participated in the ASQ survey at least once between 2013 and 2016.

Using a Python[5] script, 123,067 reviews posted between November 20, 2007, and November 10, 2016, were collected. Since Google discloses the timeline of each review in a blurred format (e.g., two days ago, three weeks ago, six months ago, two years ago), the dates of specific reviews were established by counting backward from the day the data were collected (e.g., two days ago from November 10, 2016, is translated into November 8, 2016).

Python enchant library is used to classify the data into English and non-English reviews. An “English” tag is attached when 50% of the words in a review passes the English spell check test. Among the 123,067 reviews, 42,137 are deemed as “English” reviews (see Table 1). Of the 42,137 reviews, 55% were posted in 2016, 27% in 2015, 7% in 2014, 6% in 2013, and 5% between 2007 and 2012. Table 2 presents the descriptive statistics of the Google star ratings for the 2008–2016 period. We can easily see the explosive growth of the reviews beginning in 2014–2015 when Google started to reveal the airport review section in its search result pages for airports.

4. Data analysis

To test the feasibility of using Google textual reviews to assess airport service quality attributes, we conduct a sentiment analysis (Liu, 2012) in R programing language environment using AFINN sentiment lexicon (Nielsen, 2011). The sentiment analysis is performed as an inner join between the tokenized list of the reviews and the sentiment lexicon that contains a list of emotionally laden keywords with a positive or a negative score tag. This procedure generates a sentiment score for each review so that we can compare them with Google star ratings as well as ASQ ratings.

Next, topics from the textual reviews are extracted using the Latent Dirichlet Allocation (LDA) algorithm. LDA is a probabilistic model of text and is designed to identify latent topics from a large set of documents without requiring any human annotation. Blei (2012) provides a graphical illustration of the model as shown in Fig. 2 as well as the mathematical equation (Equation (1), Blei, 2012).

K - Total number of topics; βk - Topics distribution over K; D - Total

[4] It is worth noting that the main Skytrax methodology (that results in the awards) is not the online reviews but an independent annual survey of the airports. Skytrax states in their website that “the Skytrax World Airport Awards are voted for by air travelers in the largest, annual global airport customer satisfaction survey” which is “operated as an independent study with no entry fees or charges to any airport.” http://www.worldairportawards.com/accessed on May 17, 2018

[5] Python is an interpreted high-level language for general-purpose programming. Since the APIs that Google Maps provides has a strong restriction on the number of reviews to fetch per query, and they use AJAX (Asynchronous JavaScript + XML) interface, we had to write a private Python script to scrape the reviews rather than using Google Maps' standard APIs or a (commercial) web scraping software.

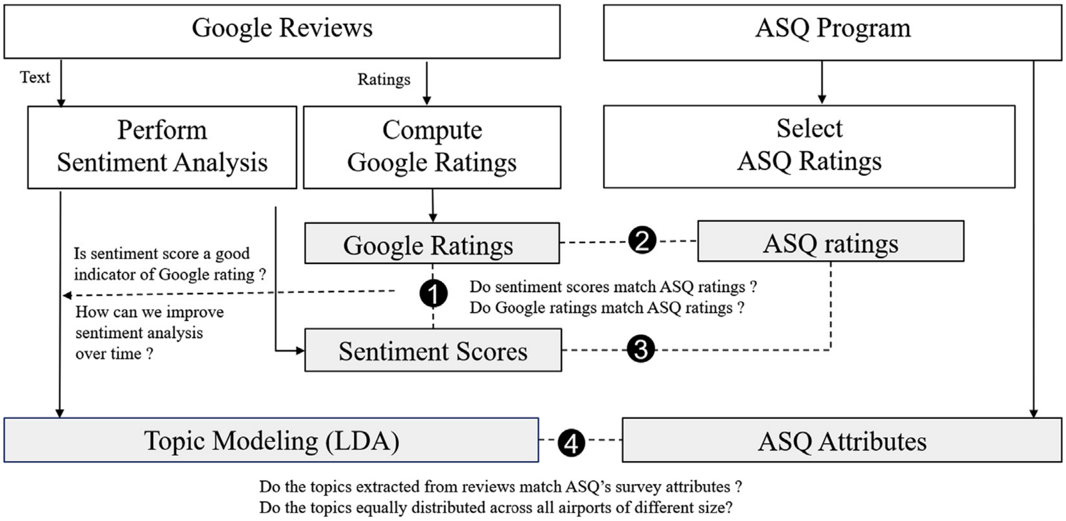

Google Reviews
ASQ Program
Text
Ratings
Perform Sentiment Analysis
Compute Google Ratings
Select ASQ Ratings
Google Ratings
2
ASQ ratings
Is sentiment score a good indicator of Google rating ?
How can we improve sentiment analysis over time ?
1
Do sentiment scores match ASQ ratings ?
Do Google ratings match ASQ ratings ?
Sentiment Scores
3
Topic Modeling (LDA)
4
ASQ Attributes
Do the topics extracted from reviews match ASQ's survey attributes ?
Do the topics equally distributed across all airports of different size?

Fig. 1. Analytical framework.

Table 1
Distribution of reviews: English versus other languages.

Total Records Collected	Without Review		With Reviews			
	(Rating only)		In Other Language		In English	
123,067	58,593	48%	22,337	18%	42,137	34%

Table 2
Descriptive statistics of reviews in English and average Google star ratings.

	2008	2009	2010	2011	2012	2013	2014	2015	2016
No of Reviews	10	43	50	664	1188	2482	3143	11,388	23,168
No of Airports	8	14	20	72	85	94	96	99	98
Mean	3.88	3.04	3.48	3.53	3.8	3.73	3.71	3.75	3.74
Median	5	3.25	4.18	3.65	3.97	3.78	3.76	3.88	3.78
Minimum	1	1	1	1	1.5	1.5	2	2	2.47
Maximum	5	5	5	5	5	4.86	4.83	4.78	4.69
Standard Deviation	1.81	1.79	1.72	0.82	0.69	0.62	0.61	0.54	0.47

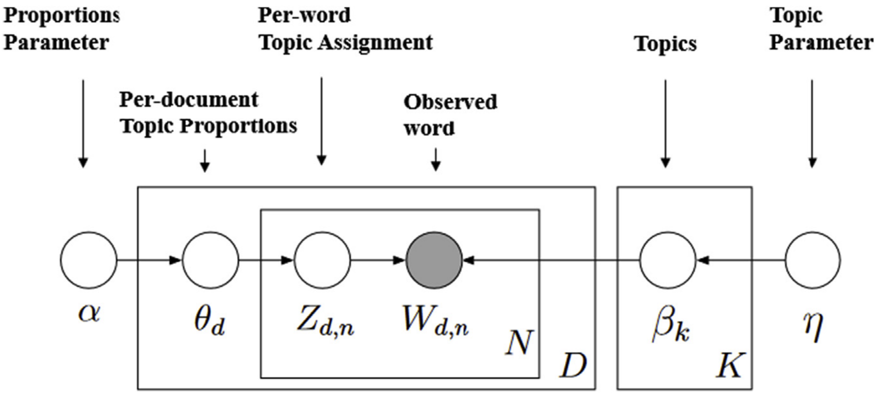


Fig. 2. LDA as a graphical model (Blei, 2012).

$$p(\beta_{1:K}, \theta_{1:D}, \mathrm{z}_{1:D}, \mathrm{w}_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(W_{d,n}|\beta_{1:K,} Z_{d,n}) \right) \tag{1}$$

number of documents.

θd - Per-document topic proportions; N - Number of words in a document.

Zd,n - Per-word topic assignment; Wd.n - Observed word; α,η - Dirichlet Parameters.

This probabilistic algorithm is gaining increasing popularity to make sense out of large amounts of textual contents. The algorithm assumes that each document is composed of multiple latent topics and each topic is expressed as a collection of words. By maximizing inter-class variance, LDA estimates the probabilities of these topics and words in every document at the same time (Blei et al., 2003; Silge and Robinson, 2016). The extracted topics are then used to examine the degree to which passengers' interests ("topics") being discussed in their reviews agree with the ASQ survey questionnaires.

4.1. Computation of sentiments scores from Google review texts

Three general sentiment lexicons are publicly available for the sentiment analysis: NRC Emotion (Mohammad and Turney, 2013), Bing Liu (Hu and Liu, 2004), and AFINN(Nielsen, 2011). These lexicons assign positive or negative sentiment scores to thousands of common English words. The AFINN lexicon assigns scores between −5 and 5 to each of its 2476 entries; Bing assigns positive or negative to each word; NRC categorizes 14,182 words in positive or negative sentiment along with 8 emotional dimensions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) (Mohammad and Turney, 2013). All three lexicons are combined as a data frame in “tidytext” package (Silge and Robinson, 2016) in R.

We chose AFINN as it showed the best pretest performance in our dataset. The “tidytext' framework provides a handy procedure to compute per review sentiment scores by executing an inner-join operation between the “unnested” records (i.e., one-term per row) of reviews and sentiment lexicon data frame. After the inner-join, we compute the net sentiment (positive-negative) per each review. For instance, in the sample reviews shown below, the net sentiment score for the first review is −4. The net sentiment score for the second review is 2, computed by (3 + 1+3 + 1)/4.

“Holy long ass (−4) way to the terminal after security.”

“I'm quite impressed (+3) that such a huge (+1) airport is so organized! There are a lot of shops and cafes after security check in case you have some time to spare. KLM is king here:) Access to the city center is super (+3) easy (+1) by taking the train directly from the terminal.”

Table 3 Presents the descriptive statistics of the sentiment scores for the 2008–2016 period. The sentiment scores for 2015 and 2016 are provided in Appendix. It should be noted that our analysis is focused on 2015 and 2016 Google reviews, as the sample sizes for earlier years are too small.

4.2. Sentiments scores versus Google Star Ratings

As shown in Fig. 1, the first step of our analysis is to conduct a within-Google comparison, testing the consistency between the sentiment scores and Google star ratings. As expected, per review sentiment scores from the review texts appear to be strongly correlated with the associated Google star ratings (See Fig. 3). While the overall relationship is clear, we see many data points showing prediction errors. Given that we use unigram-based analysis that does not consider any complex issues such as negation (e.g., “Not fantastic”) or sarcasm (“Nice perfume, you marinate in it?”), these errors are expected. However, once the data is summarized at airport level and filtered by year, these noises cancel each other out and show a much stronger correlation as the number of data per year increases (See Fig. 4).

Shapiro-Wilk normality tests are conducted for both Google star ratings and sentiment scores. The p-value for Google star ratings is 0.002 in 2015 and 0.362 in 2016; and the p-value for the sentiment scores is 0.335 in 2015 and 0.760 in 2016. That is, the 2015 Google star ratings are not from a normally distributed population. Therefore, the

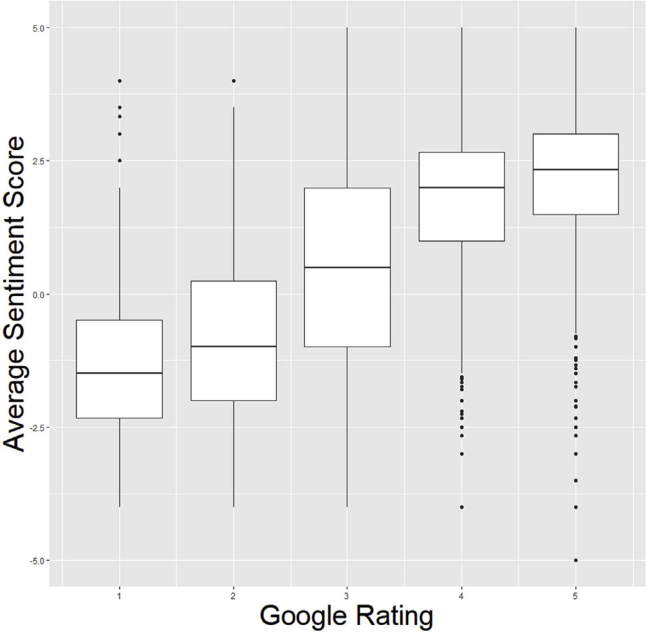


Fig. 3. Sentiment scores versus Google star ratings: Per review.

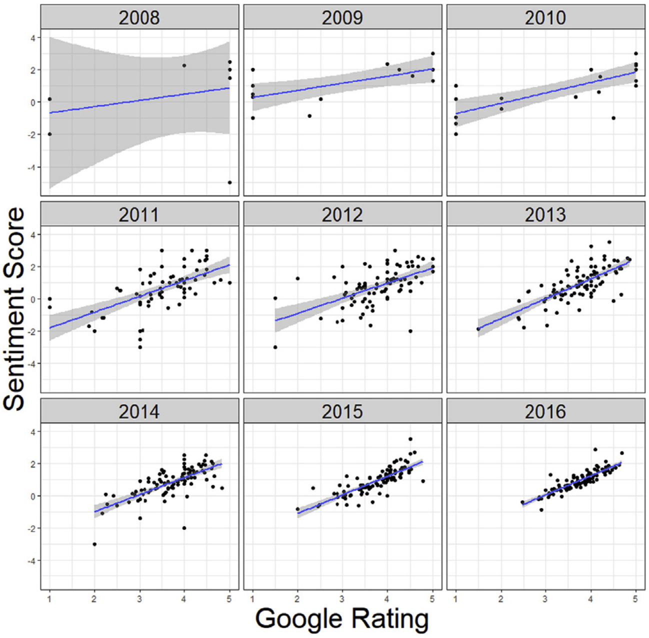


Fig. 4. Sentiment scores versus Google star ratings: Yearly at airport level.

Table 3
Descriptive statistics of sentiment scores.

	2008	2009	2010	2011	2012	2013	2014	2015	2016
No of Reviews	10	43	50	664	1188	2482	3143	11,388	23,168
No of Airports	8	14	20	72	85	94	96	99	98
Mean	0.43	1.18	0.87	0.68	0.8	0.91	0.82	0.93	0.94
Median	1.75	1.48	1	0.88	0.91	0.83	0.9	1.06	0.9
Minimum	−5	−1	−2	−3	−3	−1.88	−3	−0.83	−0.87
Maximum	2.5	3	3	3	3	3.5	2.54	3.5	2.89
Standard Deviation	2.65	1.2	1.48	1.3	1.16	1.04	0.92	0.78	0.63

Table 4
Spearman's rank correlations between sentiment scores and Google star ratings.

Year	# of Reviews Written in English	r_s	p
2009	43	.59	< .05
2010	50	.79	< .01
2011	664	.69	< .01
2012	1188	.57	< .01
2013	2482	.70	< .01
2014	3143	.71	< .01
2015	11,388	.84	< .01
2016	23,168	.89	< .01
Total	42,137[a]		

[a] The records from 2007 and 2008 are excluded from the table due to the small quantity (most of them were ratings-only reviews, and only 11 reviews were written in English).

Table 5
Descriptive statistics of ASQ overall passenger satisfaction ratings.[a]

	2013	2014	2015	2016
No of Airports	94	96	99	98
Mean	4.07	4.12	4.19	4.23
Median	4.03	4.05	4.12	4.16
Minimum	3.1	3.04	3.21	3.39
Maximum	4.97	4.97	4.98	4.99
Standard Deviation	0.4	0.38	0.4	0.39

[a] The ASQ sample consists of the 100 target airports that participated in ASQ survey at least once between 2013 and 2016. However, airports that did not participate all 4 quarterly surveys in a year are excluded from our analysis.

non-parametric Spearman's rank correlation is used to compare Google star ratings and sentiment scores. The correlations between the sentiment scores and Google star ratings at airport level are strong and have been increasing over time: $r_s(96) = 0.89$, $p < 0.01$ in 2016 (See Table 4).

4.3. Google reviews versus ASQ ratings

Descriptive statistics for ASQ overall passenger satisfaction scores (ASQ ratings) are presented in Table 5. The p-value from Shapiro-Wilk tests for normality is 0.004 in 2015 and 0.001 in 2016, indicating lack of normality with the ASQ scores. As such, Spearman's rank correlation is adopted.

Step 2 of the study compares Google star ratings with ASQ ratings. As shown in Fig. 5, there are significant correlations between the two sets of ratings in both 2015 and 2016: $r_s(87) = 0.35$, $p < .01$ in 2015, and $r_s(82) = 0.51$, $p < .01$ in 2016.

Step 3 compares sentiment scores with ASQ ratings as shown in Fig. 6. Again there are significant correlations between the two sets of ratings: $r_s(87) = 0.40$, $p < .01$ in 2015, $r_s(82) = 0.52$, $p < .01$ in 2016.

Both comparisons show that the correlations between Google reviews and ASQ ratings improve over time as the number of reviews increase. However, both Figs. 5 and 6 reveal four noticeable outliers as outlined by the red boxes. These outliers (PEK, CAN, HGH, and PVG[6]) have very high ASQ ratings, but very low Google star ratings and sentiment scores. Specifically, these four outliers exhibit Cook's Distance values (0.11, 0.09, 0.12, and 0.04 respectively) greater than the cutoff point of .038 – the data points that show Cook's D greater than three times the mean Cook's D value of all data points are considered potential outliers (Neter et al., 1996). Excluding these four outliers increases the correlations between Google reviews and ASQ ratings considerably: Google star rating versus ASQ ratings at ($r_s(78) = 0.64$, $p < .01$), and Sentiment Score versus ASQ ratings at ($r_s(78) = 0.63$, $p < .01$) in 2016 (See Table 6).

There are two possible explanations for the discrepancy between the high ASQ ratings but low Google reviews for these outliers. First, the discrepancy may reflect the different perceptions of service quality between English speaking versus non-English speaking passengers[7] (c.f., Pantouvakis and Renzi, 2016). The analysis is based on 42,137 records that contain only reviews written in English, whereas ASQ surveys are conducted at airports, randomly selecting both English speaking and Non-English speaking passengers. Second, the discrepancy may be partially due to a procedural noise that tainted ASQ survey results at these airports. This warrants a further investigation that includes non-English reviews and in-depth analysis of the outlier airports. Nevertheless, the existence of these obvious outliers suggests that there is a need for cross-validation between the conventional and the alternative approach of assessing airport service qualities.

4.4. Topic modeling and comparison with ASQ service attributes

Since 2006, ASQ has been using the standard questionnaires with 34 performance metrics reflecting passengers' perception of airport service attributes. In Step 4, we use LDA to extract the topics that are discussed in the Google reviews and compares them with the ASQ's 34 service attributes. The results illustrate the degree to which the ASQ service attributes are consistent with the service quality attributes that passengers, as a crowd, voluntarily describe.

Because LDA relies on word co-occurrence patterns at the document level, using “too short” texts often leads to a bad outcome (Phan et al., 2008; Yan et al., 2013). Therefore, the reviews with less than 50 characters are filtered out. This leaves 24,554 reviews for modeling with the standard preprocessing of the textual data. Using SMART lexicon of tm package (Feinerer, 2017) in R, we further remove 1149 stop-words (e.g., “about”, “above”, “after”, “all”), add 147 custom stopwords (e.g., city names, numbers), and replace some common bigrams with a single word (e.g., “wi fi”, to wifi, “duty free” to “dutyfree”). A Document-Term Matrix, a contingency table in which each row represents a review (24,536 rows), and each column represents a word (26,518 columns), is created to feed into the LDA model.

While LDA uses Bayesian inference to generatively estimate the posterior model distribution based only on the words shown in the texts, it requires one parameter (k: number of latent topics to identify) to begin with its iteration process (see Equation (1)). Researchers have recommended various approaches to establish the optimal k (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths and Steyvers, 2004; Zhao et al., 2015). These approaches provide a good range of possible k values that are mathematically plausible. We use the R package, ldatuning (Nikita, 2014), which simultaneously run three different approaches: topic-density minimization method of Cao et al. (2009), KL-divergence minimization method of Arun et al. (2010), and expectation maximization method of Griffiths and Steyvers (2004).

As shown in Fig. 7, the minimization methods (top chart) and the maximum likelihood method (bottom chart) agree that the ideal number of topics for our sample dataset is either 24 or 25. Consequently, two alternative LDA models are estimated with the k value at 24 and 25, respectively. After reviewing the two sets of results, we determine that the 25 latent topics are semantically more plausible with only nuanced differences. Table 7 shows the 25 extracted latent topics.

Since the extracted topics, expressed as collections of words, are inherently latent, they often contain multi-dimensional semantics. To

[6] Beijing Capital International Airport (PEK), Guangzhou Baiyun International Airport (CAN), Shanghai Pudong International Airport (PVG), Hangzhou Xiaoshan International Airport (HGH).

[7] The four outlier airports are all in China. Google is not accessible in China. Therefore, in addition to the difference between English speaking versus non-English speaking, there could also be a difference between Chinese local residents versus travelers from outside China.

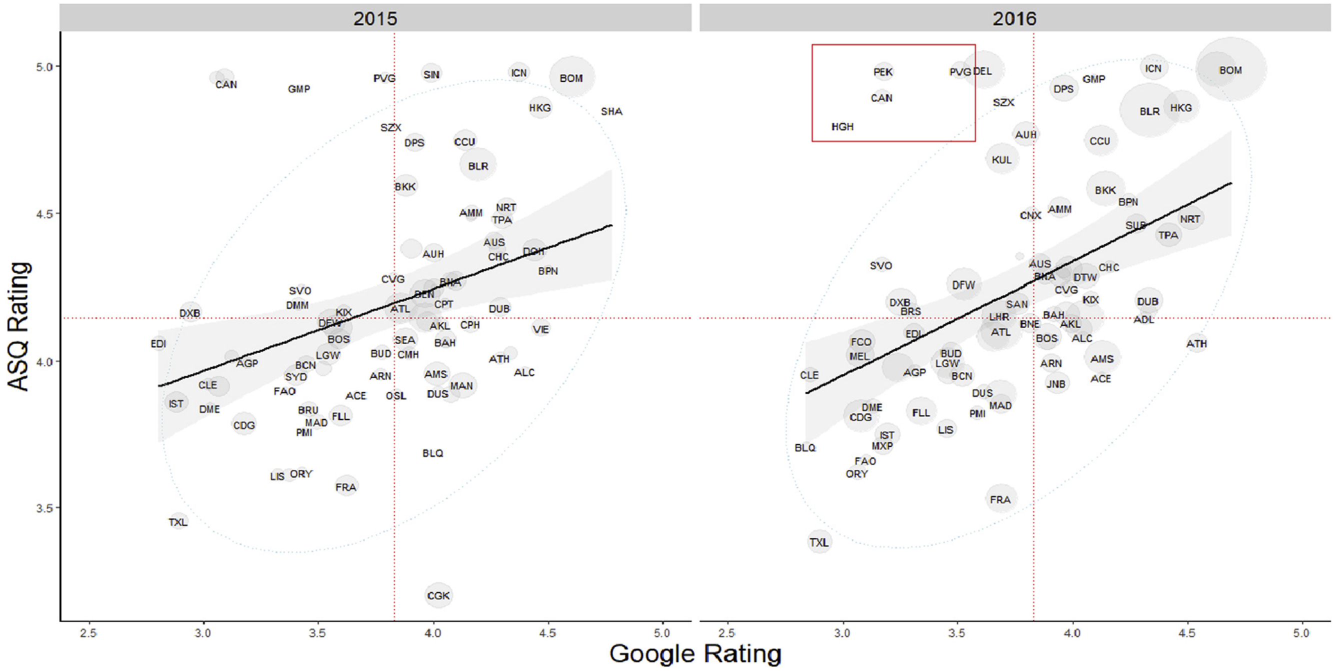


Fig. 5. Google star ratings versus ASQ ratings: 2015 and 2016.

Table 6
ASQ Ratings versus Google Reviews with or without the four outliers in 2016.

	Google Star rating vs. ASQ rating	Sentiment Score vs. ASQ rating
With Outliers	$r_s(82) = .51, p < .01$	$r_s(82) = .51, p < .01$
Without Outliers	$r_s(78) = .64, p < .01$	$r_s(78) = .63, p < .01$

provide a better understanding of the latent topics, Fig. 8 presents some examples of the topic-specific words probabilities (β). For instance, the word “security” has a 28% probability of being generated from Topic 1, whereas “tsa” has 6% probability of being generated from Topic 1.

We map the 25 latent topics onto the 34 ASQ service attributes for the last ten years (See Table 8). This procedure requires researchers' involvement to clarify the multiple semantics associated with a few topics. For example, the word “lounge” in Topic 18 is used to refer to not only business lounges but more often gate areas. The word “options' in Topic 16 is used to refer to choices for either shops or restaurants.

Overall, the extracted 25 topics match the service attributes of ASQ survey reasonably well, suggesting that ASQ program has been successfully covering most of the service quality attributes that passengers collectively express in their reviews. For example, Topic 1 corresponds well with security check experiences specified in category 11 to 14 of

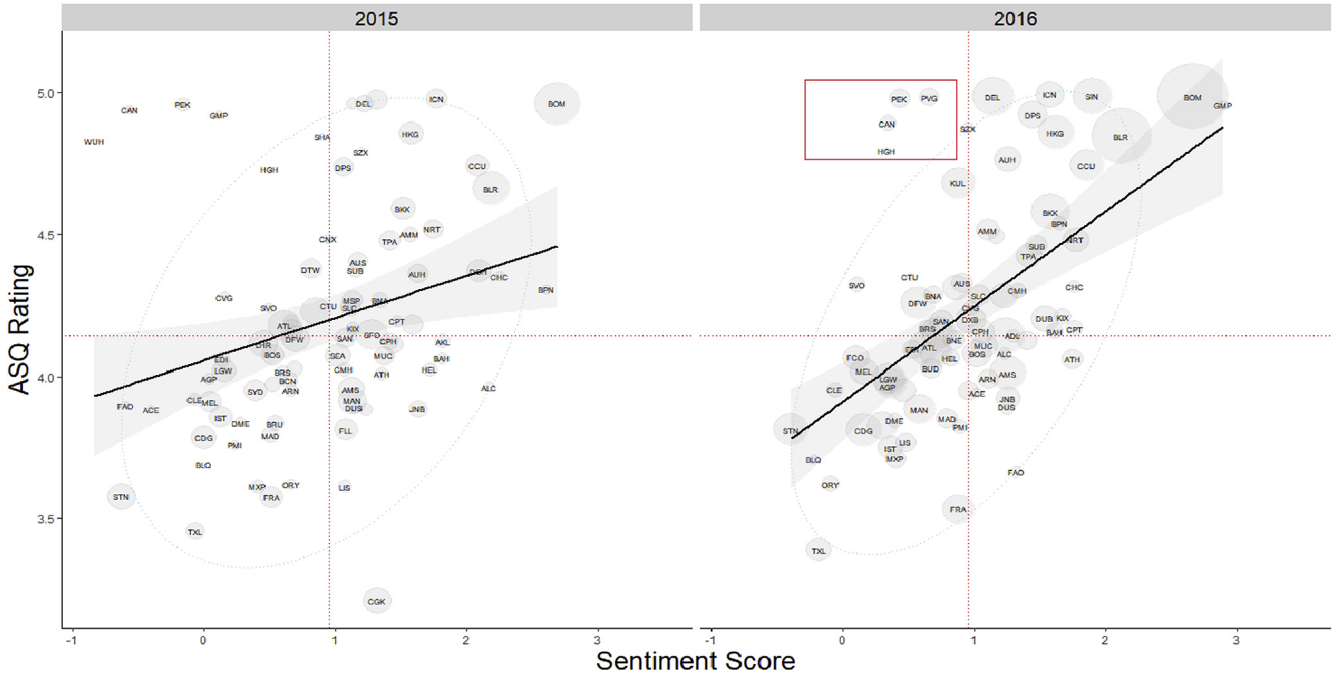


Fig. 6. Sentiment scores versus ASQ ratings: 2015 and 2016.

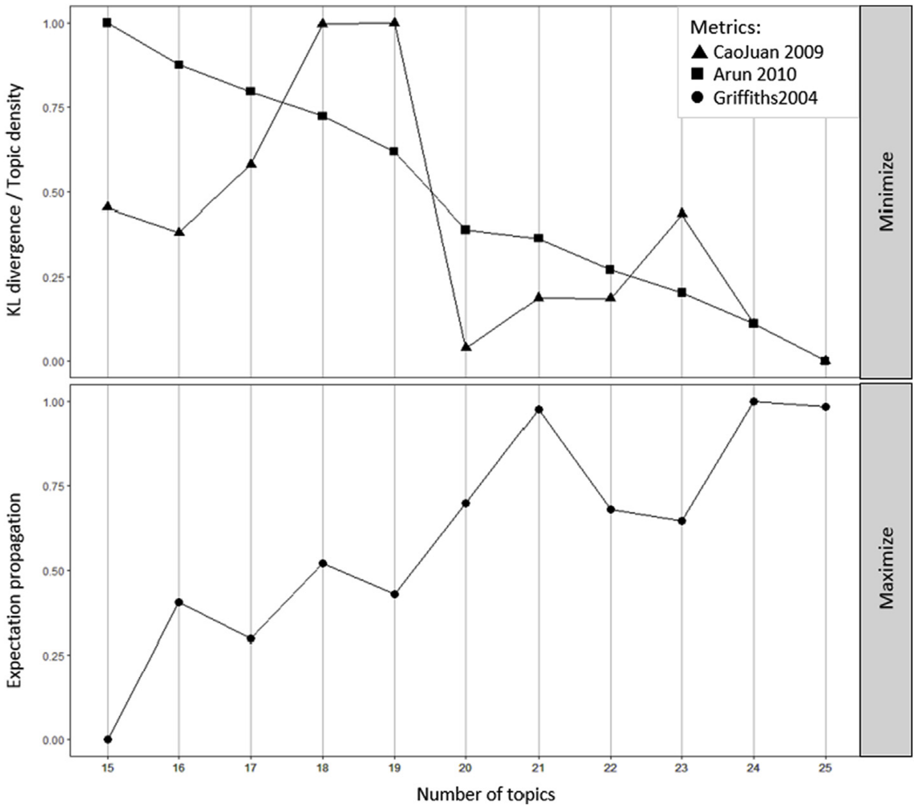


Fig. 7. Determining the number of latent topics (k).

Table 7
Extracted Latent Topics with keywords.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
1	security	minutes	food	dutyfree	terminal	gates	passport	Easy	service	worst	rude	parking	city
2	Fast	waiting	lot	money	takes	crowded	control	Free	bad	times	customs	car	expensive
3	Lines	found	shopping	poor	shuttle	avoid	Slow	Wifi	taxi	baggage	told	rental	inside
4	Tsa	walking	bus	toilets	confusing	extremely	Queue	Pretty	services	signs	pass	pick	traffic
5	Feel	stop	selection	water	delta	quickly	Bags	Navigate	information	lost	customer	charge	public
6	terrible	left	connection	shop	delays	seating	Bag	Quick	convenient	run	english	park	expect
7	Sit	family	main	building	united	taking	Queues	awesome	system	signage	employees	phone	designed
8	Fine	finally	spacious	buy	moving	cool	Worse	access	trip	claim	job	card	seats
9	checks	closed	choices	spend	tram	enjoy	arrivals	super	cost	arriving	person	return	transportation
10	country	assistance	choice	dirty	ground	issues	departures	favorite	reach	flew	speak	call	transport

	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
1	Time	hours	options	international	train	beautiful	Check	nice	shops	flight	staff	love
2	Walk	wait	lots	world	lounge	day	experience	clean	restaurants	flights	people	fat
3	departure	immigration	facilities	fly	short	home	Gate	friendly	efficient	hour	line	helpful
4	Pay	plane	plenty	huge	station	visit	Airline	makes	organized	transit	hard	busy
5	Drop	luggage	eat	domestic	stay	design	Travel	modern	modern	connecting	desk	amazing
6	arrival	passengers	decent	compared	hotel	visited	Flying	layover	stores	due	care	excellent
7	process	leave	comfortable	major	business	restaurant	horrible	maintained	europe	delayed	change	class
8	transfer	arrive	space	limited	prices	art	Ticket	layout	internet	coming	hope	local
9	counter	half	outlets	night	lounges	south	Close	pleasant	organised	late	stuff	top
10	Mins	checked	coffee	busiest	recommend	loved	construction	rest	smoking	arrived	front	stars

ASQ. Topic 12 corresponds with airport parking experiences in category 3 and 4 of ASQ. While Topic 12, 13, and 18 simultaneously match to category 4 of ASQ, ground transportation to/from the airport, T12 is about the rental car experience in association with parking facilities, T13 is about airport public transportation in general, and T18 is specifically about trains to and from the city.

It should be noted that, however, two ASQ service attributes, shaded in black in Table 8, could not be mapped onto any of the 25 extracted topics. That is, these two service attributes (availability of baggage carts, availability of bank/ATM facilities and money changers) are not mentioned often enough to be extracted as a distinctive topic. While “baggage” or “luggage” was mentioned by 1192 reviews, the cart was mentioned only 28 times and trolley 3 times. In fact, when “carts” were occasionally commented, it is more often about the fee attached to it (e.g., “… the only downsides are paid baggage cart”) rather than availability. Similarly, Bank and ATM were mentioned only in 24, and 156 reviews, respectively, which may indicate their availability is not an issue for most passengers anymore.

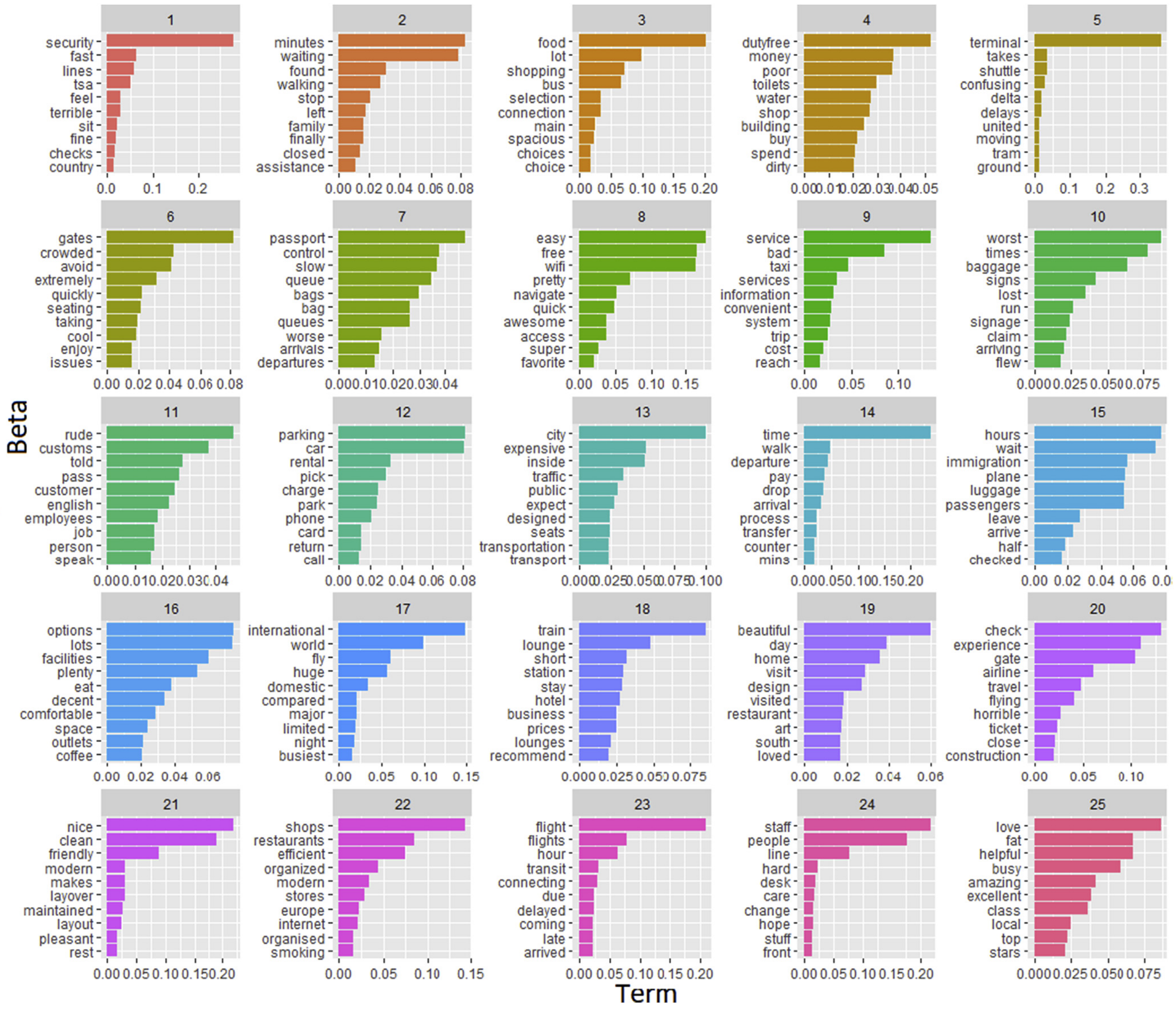


Fig. 8. Topic-specific word probabilities.

4.5. *Relative frequency of topics*

While the good correspondence between the extracted topics and ASQ survey attributes suggests that ASQ covers most of the service attributes that matter most to reviewers as a collective, but it does not tell whether the relative importance of these service attributes is equal across different airports. As the perception of service quality is context dependent (Bezerra and Gomes, 2015; Brady and Cronin, 2001; Pantouvakis and Renzi, 2016), we expect that the relative weights of the extracted topics would be different between airports of different sizes. To examine this, we compute the relative weights of topics among four airport groups of different sizes and compare them through the following procedure.

While LDA models each topic as a combination of words (with probabilities of *β),* it also estimates each document as a combination of topics (with probabilities of γ*)*. Take ***Review 51*** as an example, which states:

“It is one of the most efficient and friendliest security check I know in Europe, The restaurants are good, and there is something for every taste! The electronics shop has a rather interesting selection. Once you move to the gates, there is not much else to do so stay at the main hall as long as you can to kill some time!” – ***Review 51***.

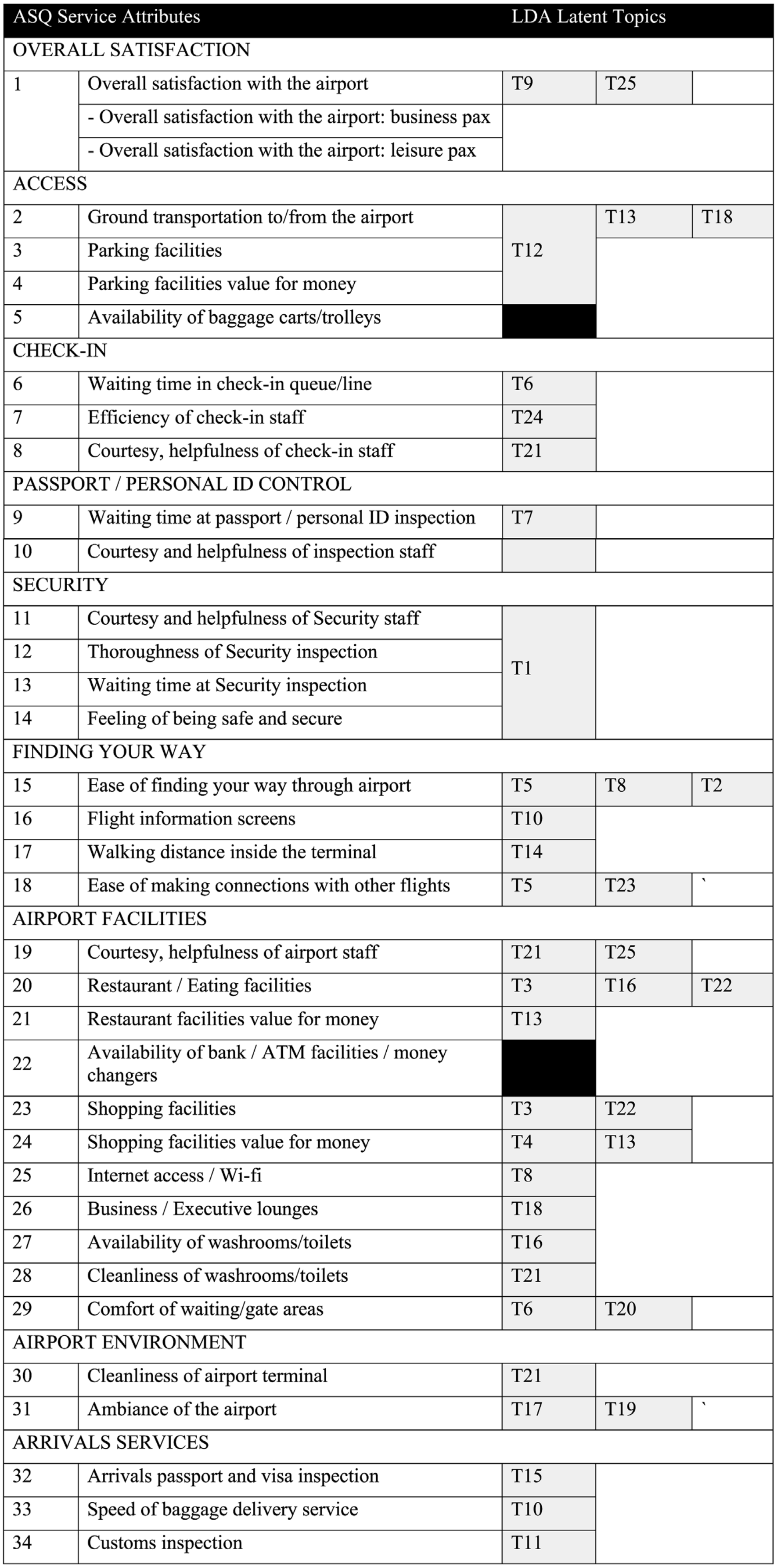
The LDA model assigns the highest probability to Topic 1 (γ = 0.1364), Topic 22 (γ = 0.1034), then Topic 6 (γ = 0.0745), and so forth. That is, there is a 13.64% chance that the words in ***Review 51*** were generated from Topic 1 (i.e., security check), 10.34% chance from Topic 22 (i.e., shops and restaurant facilities), and 7.45% chance from Topic 6 (i.e., waiting at the gate).

To identify the frequency distribution of the topics in our corpora, we introduce a heuristic assumption that each review should be categorized into one, and only one, topic group. That is, each review is categorized into one topic group that shows the highest γ value. With this assumption, the relative frequencies of the topics are shown in Table 9. The cells shaded in black in the rightmost column represent the top 3 topics that are the most frequently discussed by passengers: Topic 8 (Free Wi-Fi and easy navigation), Topic 16 (Variety in facilities), and Topic 21 (Nice, clean, and modern). And, the shaded cells in each column represent the top 3 topics in each airport size group.

4.6. *Extracted topics and airport size*

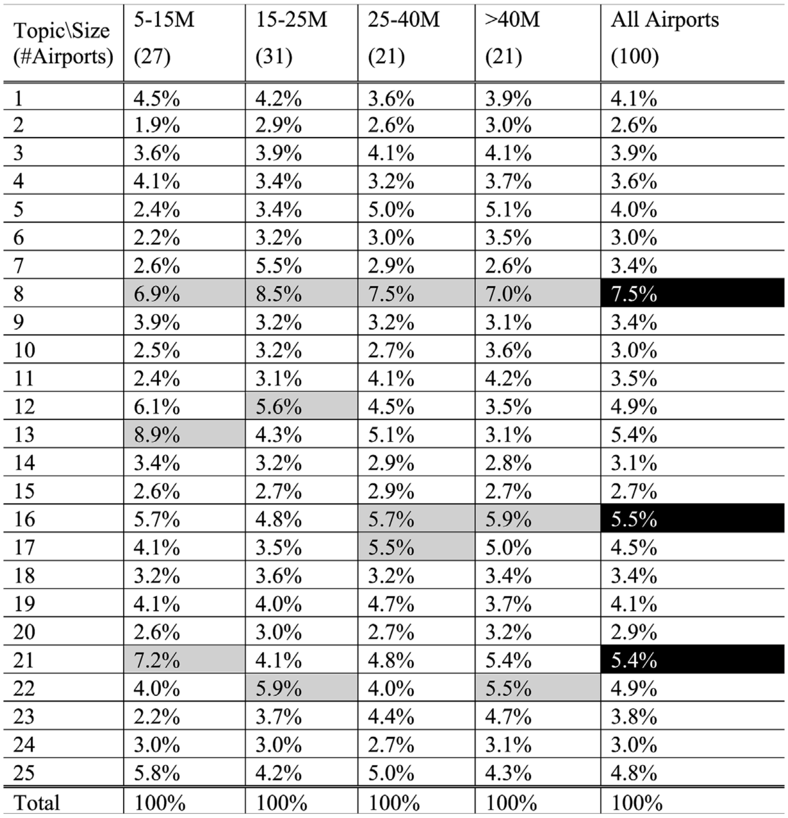
Are the occurrences of topics distributed equally to all airports of different size? Our data, summarized in Table 8, suggest that the are not equal. Chi-square test of independence shows a significant interaction

Table 8
Mapping of ASQ service attributes to latent topics extracted from reviews.



ASQ Service Attributes		LDA Latent Topics		
OVERALL SATISFACTION				
1	Overall satisfaction with the airport	T9	T25	
	- Overall satisfaction with the airport: business pax			
	- Overall satisfaction with the airport: leisure pax			
ACCESS				
2	Ground transportation to/from the airport	T12	T13	T18
3	Parking facilities			
4	Parking facilities value for money			
5	Availability of baggage carts/trolleys	■		
CHECK-IN				
6	Waiting time in check-in queue/line	T6		
7	Efficiency of check-in staff	T24		
8	Courtesy, helpfulness of check-in staff	T21		
PASSPORT / PERSONAL ID CONTROL				
9	Waiting time at passport / personal ID inspection	T7		
10	Courtesy and helpfulness of inspection staff			
SECURITY				
11	Courtesy and helpfulness of Security staff	T1		
12	Thoroughness of Security inspection			
13	Waiting time at Security inspection			
14	Feeling of being safe and secure			
FINDING YOUR WAY				
15	Ease of finding your way through airport	T5	T8	T2
16	Flight information screens	T10		
17	Walking distance inside the terminal	T14		
18	Ease of making connections with other flights	T5	T23	`
AIRPORT FACILITIES				
19	Courtesy, helpfulness of airport staff	T21	T25	
20	Restaurant / Eating facilities	T3	T16	T22
21	Restaurant facilities value for money	T13		
22	Availability of bank / ATM facilities / money changers	■		
23	Shopping facilities	T3	T22	
24	Shopping facilities value for money	T4	T13	
25	Internet access / Wi-fi	T8		
26	Business / Executive lounges	T18		
27	Availability of washrooms/toilets	T16		
28	Cleanliness of washrooms/toilets	T21		
29	Comfort of waiting/gate areas	T6	T20	
AIRPORT ENVIRONMENT				
30	Cleanliness of airport terminal	T21		
31	Ambiance of the airport	T17	T19	`
ARRIVALS SERVICES				
32	Arrivals passport and visa inspection	T15		
33	Speed of baggage delivery service	T10		
34	Customs inspection	T11		

Table 9
Relative frequencies of extracted topics by airport size.



Topic\Size (#Airports)	5-15M (27)	15-25M (31)	25-40M (21)	>40M (21)	All Airports (100)
1	4.5%	4.2%	3.6%	3.9%	4.1%
2	1.9%	2.9%	2.6%	3.0%	2.6%
3	3.6%	3.9%	4.1%	4.1%	3.9%
4	4.1%	3.4%	3.2%	3.7%	3.6%
5	2.4%	3.4%	5.0%	5.1%	4.0%
6	2.2%	3.2%	3.0%	3.5%	3.0%
7	2.6%	5.5%	2.9%	2.6%	3.4%
8	6.9%	8.5%	7.5%	7.0%	7.5%
9	3.9%	3.2%	3.2%	3.1%	3.4%
10	2.5%	3.2%	2.7%	3.6%	3.0%
11	2.4%	3.1%	4.1%	4.2%	3.5%
12	6.1%	5.6%	4.5%	3.5%	4.9%
13	8.9%	4.3%	5.1%	3.1%	5.4%
14	3.4%	3.2%	2.9%	2.8%	3.1%
15	2.6%	2.7%	2.9%	2.7%	2.7%
16	5.7%	4.8%	5.7%	5.9%	5.5%
17	4.1%	3.5%	5.5%	5.0%	4.5%
18	3.2%	3.6%	3.2%	3.4%	3.4%
19	4.1%	4.0%	4.7%	3.7%	4.1%
20	2.6%	3.0%	2.7%	3.2%	2.9%
21	7.2%	4.1%	4.8%	5.4%	5.4%
22	4.0%	5.9%	4.0%	5.5%	4.9%
23	2.2%	3.7%	4.4%	4.7%	3.8%
24	3.0%	3.0%	2.7%	3.1%	3.0%
25	5.8%	4.2%	5.0%	4.3%	4.8%
Total	100%	100%	100%	100%	100%

($\chi^2(96) = 240.64$, $p < .01$) between the relative weight of topics and airport size. Table 8 shows the proportion of topics grouped by the size of airport passenger traffic. Passengers for all airports share Topic 8 (Free WiFi and easy navigation) as one of their top 3 subjects of interests. However, passengers at small airports (< 5–15 M) care most about Topic 13 (Transportation from/to the city) and Topic 21(cleanness and kindness), passengers at mid-sized airports(15–25 M) pay more attention to Topic 22(Foods and shops) than others, and passengers at larger airports (25–40 M; > 40 M) pay more attention to Topic 16 (Variety of facilities) and Topic 17 (ambiance) than others. These suggest that the relative importance of service quality attributes varies depending on the size of the airport. Overall, transportation to/from the city (Topic 13) and cleanness and kindness of airport staff (Topic 21) are more prominent issues for the smaller airports, whereas customs inspection (Topic 11) and nice ambiance (Topic17) are more prominent at the larger airports than others.

5. Special considerations in using Google reviews as data source

As this is the first study that assesses airport service quality based on Google reviews, in this section, we discuss some additional findings associated with the exploration.

5.1. Term frequency analysis

One of the previous research on airport service quality (Bogicevic et al., 2013) used term frequency analysis of 1095 reviews randomly selected from Skytrax. By comparing the frequency of the terms and average ratings of the reviews that contain the terms, they argued that there are some service categories that can be considered either satisfiers (e.g., cleanness, shopping) or dissatisfiers (e.g., security, dining). We conduct a term frequency analysis to examine if the reviews from Google Maps produce a consistent result.

The results (in Log Scale) are plotted in Fig. 9. The terms that were shown in more than 50 reviews covering at least 5 airports are selected to remove particularities. Fig. 9 identifies the words most frequently associated with ratings: positive, neutral, or negative. For instance, keywords such as “arrivals,” “waiting,” and “walking” are considered valence neutral; they tend to appear in both positive and negative reviews with about equal frequencies. While “clean,” “navigate,” and “architecture” are associated more often with positive valence, words such as “attitude,” “joke,” and “customer” are more often associated with negative valence. This serves as a term level snapshot of the reviews, showing the most prominent terms with associated ratings.

This provides a quick test of Bogicevic's (2013) notion of satisfiers and dissatisfiers. The test-results are mixed. Consistent with Bogicevic's results (words marked with green arrows), our data show that “shop” and “clean” are associated with high ratings (i.e., satisfiers) and “sign” is associated with low ratings (i.e., dissatisfier). However, other terms (terms marked with red arrows) that they categorized as dissatisfiers are associated with either a high rating (e.g., “food”) or a neutral rating (e.g., “security”). Instead, in our data set, “bag(s)” and “luggage” (those marked in grey) show a dissatisfier-like tendency. This may indicate a platform-specific sampling difference (Ruths and Pfeffer, 2014) between Skytrax and Google Maps. However, a more general explanation would be that whether a word is considered satisfier or dissatisfier is determined by the aggregated service conditions of an airport group from which the data is collected.

5.2. Customizing sentiment lexicon to airport review contents

As discussed in Section 4, the general purpose sentiment lexicon (AFINN) performs exceptionally well for evaluating textual reviews to infer ratings in Google Maps. However, we may improve the performance by customizing the sentiment lexicon specifically for airport reviews in Google. Because each entry in Google review has both a star

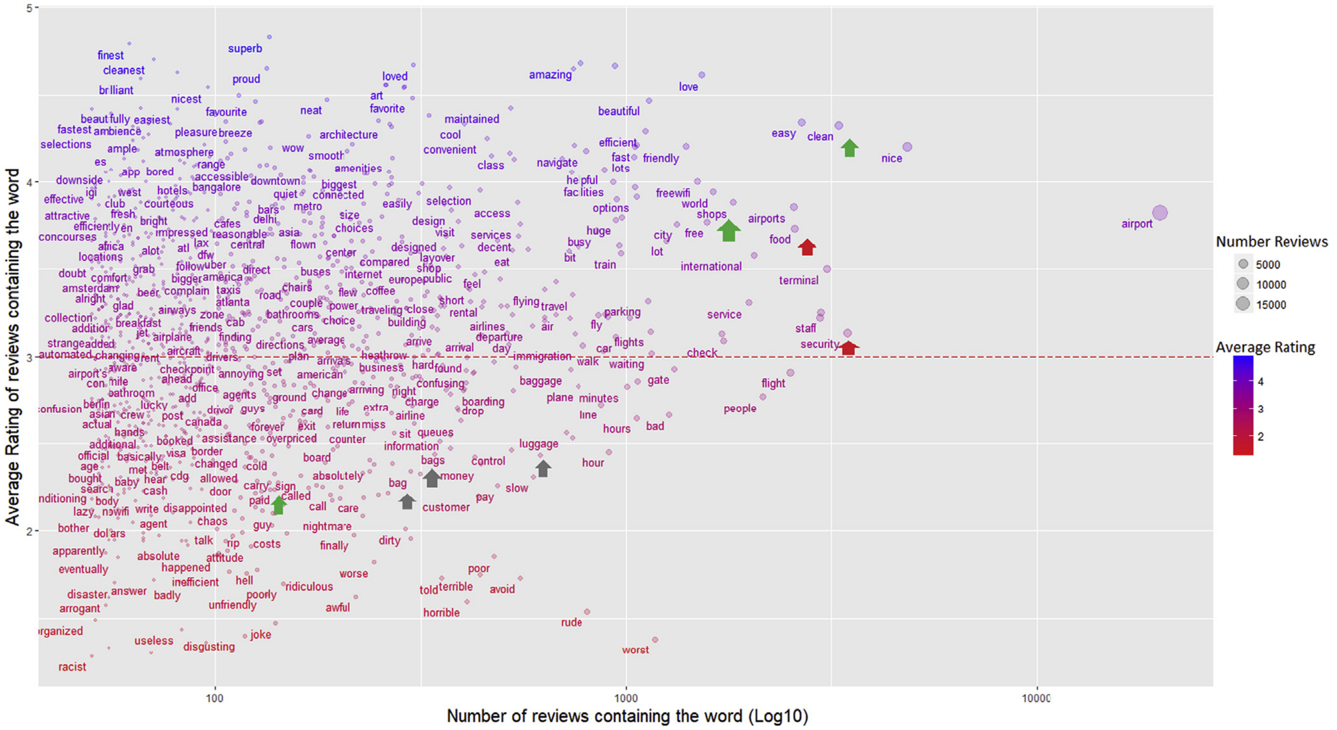


Fig. 9. Average ratings versus word frequencies.

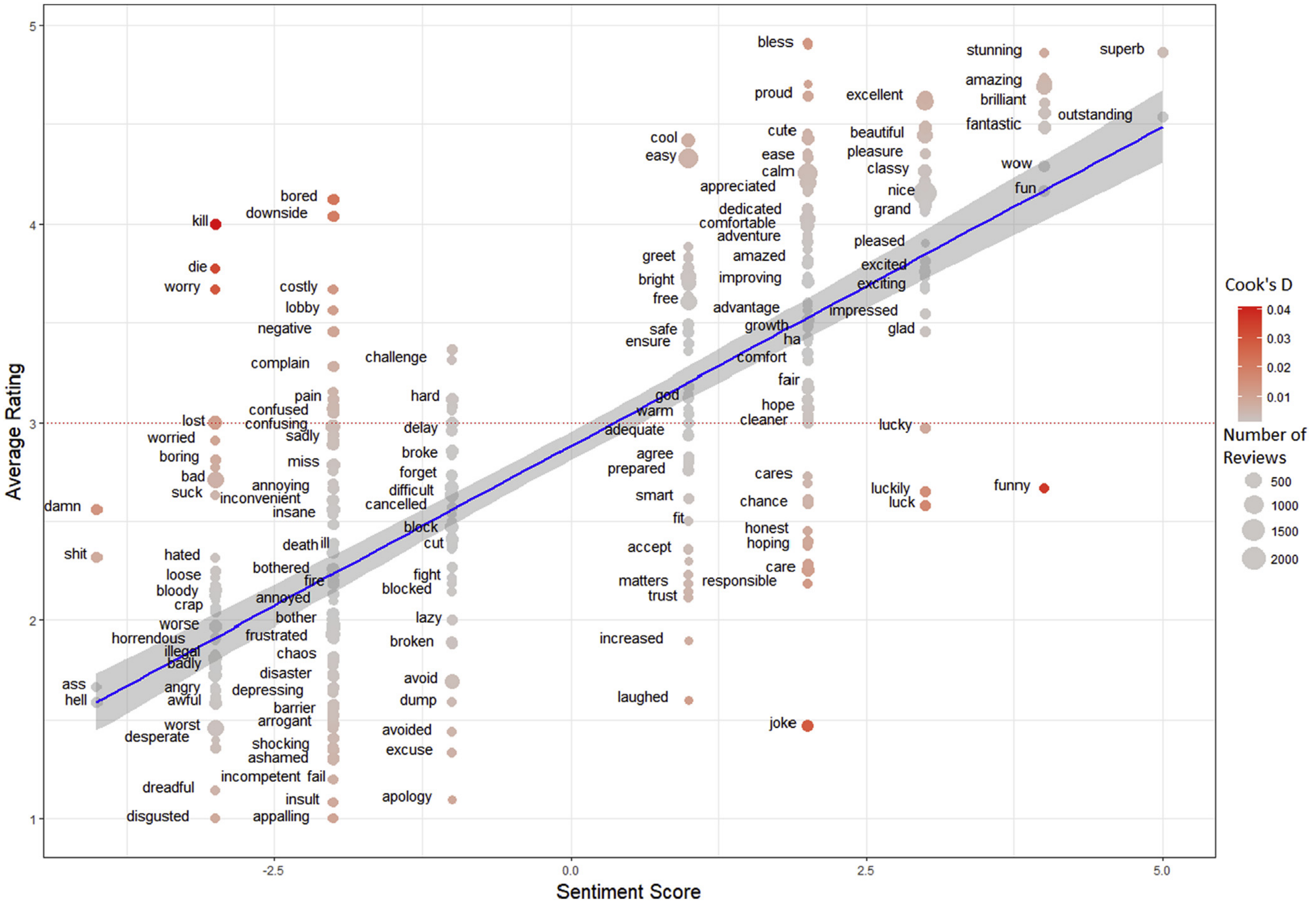


Fig. 10. The performance of AFINN lexicon in Google Rating.

rating score and textual review, we can identify the words in the AFINN lexicon that are most or least successful in predicting the valence of these reviews. We conduct a test to see whether this performance information from previous years' data could be used to improve the performance of the sentiment analysis for the following years.

Fig. 10 demonstrates the performance of high-frequency AFINN words applied to evaluate the reviews posted until the end of 2015 (dot size represents frequency, dot color represents Cook's Distance). Again, observations with Cook's D greater than three times the mean of Cook's Ds are identified as outliers (Neter et al., 1985). The closer the dot color is to red, the higher its Cook's D. For example, the word "kill" has an AFINN score of −3 (highly negative) while the average star rating of the reviews that used the word "kill" in the airport reviews was +4 (highly positive), such as "lots of shops to kill your waiting time!" or "there is an interesting little museum to kill time". The word "joke" is an example in the opposite: positive AFINN but negative ratings, such as "free Wifi is a bad joke," "meet and greet at the airport was a joke." Words like "bored" or "downside" are frequently used in Google reviews with a various form of negation, such as "You will never get bored here!," "They also have a lot of shopping should you get bored," or "Good airport, the only downside is the internet."

From the AFINN lexicon, we exclude the 22 words that show higher Cook's D value than the outlier cutoff point (0.009) in the reviews posted between 2007 and 2015. We apply the new filtered lexicon to the 2016 reviews (which has been set aside as a test set), the correlation between sentiment scores and Google star ratings improved from $r_s(96) = 0.89$ to $r_s(96) = 0.90$. This suggests that the word-usage pattern in this particular context persists. Therefore using the pair structure between rating score and textual review in Google Maps, we can improve the performance of the sentiment analysis over time by customizing the sentiment lexicon to the context of airport assessment.

6. Summary and conclusions

6.1. Summary of findings

While researchers are continuously seeking consensus on the genuine dimensions and attributes of airport service quality, ASQ has established a dominant industry standard for assessing airport service quality in terms of 34 service attributes.

In this study, we explore a potential complementary approach to assessing airport service quality based on Google reviews. For the airports that do not participate in ASQ program, this can be used as an alternative measure to compare their services against competitors. This approach can also be used to cross-validate the ASQ survey results. The study further demonstrates that the validity of the service quality attributes defined by conventional survey could be cross-checked by the topics extracted from the massive textual review data. Considering the lack of consensus on the dimensions and attributes, and given the context specificity of the service quality construct, this empirical approach to capture the attributes of airport service quality from users collective opinion appears to be a valuable avenue that warrants further investigation.

The study finds that the sentiment scores from Google reviews for the top 100 airports that participate in ASQ surveys are highly correlated with the associated Google star ratings. The correlation improves over time as the number of reviews increases, reaching $r_s(96) = 0.89$, $p < .01$ in 2016. This provides strong evidence that the lexicon based sentiment analysis is a reliable method to estimate the reviewers' quantitative ratings, and that the textual reviews, in fact, contain good information to infer passengers' perception of satisfaction. This is especially true when the number of reviews per year is sufficiently large. The study indicates that the correlation could be further improved ($r_s(96) = 0.90$, $p < .01$ in 2016) by customizing the general purpose sentiment lexicon into an airport-review specific sentiment lexicon. This improvement process is showcased by filtering out the words with the low predictive performance from the sentiment lexicon and applying the refined lexicon to perform the following year's sentiment analysis.

Based on the 2015 and 2016 data, our results indicate that both sentiment scores and Google ratings are reasonably good predictors of ASQ ratings. This is especially true when the four "outliers" are removed from the analysis, resulting in $r_s(78) = 0.63$, $p < .01$ and $r_s(78) = 0.64$, $p < .01$ in 2016, respectively. The four "outliers" (PEK, PVG, HGH, and CAN) exhibit exceptionally higher ASQ ratings than their sentiment scores and Google star ratings. These outliers may either suggest that there is a significant difference in service quality perceptions between English speaking passengers and local non-English speaking passengers or indicate that there might be a procedural noise that tainted ASQ survey results at these airports. This warrants a further investigation that includes analysis of the reviews by non-English speaking passengers. Nevertheless, the existence of such noticeable outliers itself is a good indication that there is a need for cross-validation between the conventional and the proposed alternative approaches.

The study extracts 25 latent topics from 42,137 reviews using LDA (Latent Dirichlet Allocation) algorithm. These 25 topics match the standard ASQ service attributes fairly well. This suggests that ASQ program has been successfully covering most of the service quality attributes that passengers care about and pay attention to. There are, however, a few ASQ service attributes (Availability of baggage carts/trolleys; Availability of bank/ATM facilities/money changers) that seem to be not as relevant as in the past, given the dynamic nature of the air transport industry. We conjecture that those attributes are now well facilitated by most of the airports so that passengers in most cases simply take them for granted. While ASQ has been using the identical standard survey attributes over the last 10 years, passengers' service perceptions are likely to have evolved with time and technologies.

Further analysis of extracted topics indicates that not all service attributes are equally important for airports of different sizes. For the small airports, good transportation to/from the city and cleanness and kindness of airport staff are more prominent issues, whereas, for the larger airports, customs inspection and nice ambiance appear to be more important. This suggests that there is a different order of priority for each airport for improving *its* passengers' perception of service quality. This echoes the views that service quality perception is context dependent (Brady and Cronin, 2001) and airport service qualities are culturally subjective (Pantouvakis and Renzi, 2016).

6.2. Contributions

This study contributes to the literature in three ways. First, it contributes to the airport service quality literature by empirically comparing the service attributes of ASQ against the topics extracted from an alternative data source that reflect passengers' collective experiences. Though it is a purely empirical question, examining the degree to which the ASQ attributes match the passengers' collective perceptions is an important undertaking given that there is no consensus on the nature of dimensions and attributes of airport service quality in the literature. Second, this study showcases how to adopt the much-recommended triangulation approach by comparing the ratings and topics extracted from online reviews with those from ASQ surveys to improve the validity of results. Third, we introduce Google Maps as a new source of airport-related opinion reviews that has never been investigated by any study in the airport management and the related fields.

For the decision makers and managers of commercial airports, this study illustrates an alternative approach to assess their passengers' perception of service quality and compare its fluctuation to those of their competitors. For researchers, the methods presented here can be used to facilitate more studies to investigate the relationships between the service quality and other aspects of airport performance, such as profitability, efficiency, and productivity. For instance, a recent study conducted on 30 international airports points out that there is no significant correlation between perceived service quality and profitability (r = 0.22, p = .209) (Merkert and Assaf, 2015). Airports, therefore, can find themselves in a situation where service level provision and profitability become conflicting objectives in their strategic management decisions (Merkert and Assaf, 2015). Perceived service quality is an important variable that cannot be merged with profitability as was often the case in the literature (Merkert and Assaf, 2015). By taking our approach, researchers may dramatically expand the scope of their target airports including those never participated in any survey program. This is an important contribution to the literature in airport management because many studies have cited lack of viable service quality indicators as a reason not to consider airport service quality in their studies.

While the Google review based approach showcased here does not supplant the existing assessment methods, it does complement them by cross-validating the results of the conventional assessment. Moreover, our finding of disparities in the relative importance of various service attributes between airports of different sizes will help guide future in-depth analysis of specific group(s) of airports to develop customized strategies to improve their service quality.

6.3. Limitations and future study

There are a number of limitations with this study. First, our analysis uses reviews written in English only. Out of the 123,067 records crawled from Google Maps, 42,137 records were written in English. That is, over 65% of the records are excluded either because of no reviews or because of non-English reviews. As discussed in section 4, the exclusion of the non-English reviews might have been one of the explanations for the outliers (all in China). At the time of finalizing this paper (March 2018), we notice that Google Maps has recently launched a new feature that displays an automatic English translation of every non-English review. Future research may be able to use these translated reviews to resolve this limitation.

Second, our sentiment analysis uses an unigram method that does not consider more complex bigram or trigram techniques such as negating qualifiers (e.g., "*not* interesting" or "not excited"). While we found the presented approach performs exceptionally well in analyzing the airport reviews (and even better after customizing the lexicon), a more sophisticated n-gram technique or a machine learning approach may help improve the results.

Third, as stated in Section 3, Google reviews are date-stamped relative to "today's date." That is, the date of a specific review was established by counting backward from the day the reviews were collected (November 10, 2016). This means our reviews for 2015 are actually from November 11, 2014 to November 10, 2015, not from Calendar year 2015. This shift of date stamping goes all the way back to 2007. Future researchers should be aware of this limitation. The ideal date for collecting reviews would be December 31 to match calendar years.

Our future study will use Google reviews to identify the key drivers of the perceived service quality, to examine the relative importance of the service attributes among different groups (e.g., passengers flying on network carriers vs. low-cost carriers), and to classify the service attributes (e.g., attributes that are more vs. less manageable by airports). Since Google Maps covers virtually all commercial airports throughout the world, future studies will exploit the enhanced flexibility in selecting the airports to investigate.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jairtraman.2018.05.004.

Appendix. Sentiment Scores and Google Ratings (2015, 2016)

Code	Airport Name	2015		2016	
		Sentiment Score	Google Rating	Sentiment Score	Google Rating
ACE	Lanzarote Airport	−0.400	3.667	1.024	4.129
ADL	Adelaide International Airport	0.999	4.083	1.292	4.316
AGP	Malaga-Costa del Sol Airport	0.035	3.189	0.341	3.315
AKL	Auckland International Airport	1.816	4.032	1.402	3.991
ALC	Alicante Airport	2.174	4.400	1.234	4.045
AMM	Queen Alia International Airport	1.567	4.167	1.108	3.946
AMS	Amsterdam Airport Schiphol	1.117	4.014	1.259	4.131
ARN	Stockholm-Arlanda Airport	0.664	3.773	1.104	3.906
ATH	Athens International Airport	1.352	4.292	1.752	4.545
ATL	Hartsfield-Jackson Atlanta International Airport	0.615	3.857	0.658	3.689
AUH	Abu Dhabi International Airport	1.627	4.000	1.257	3.798
AUS	Austin Bergstrom Airport	1.170	4.266	0.909	3.859
BAH	Bahrain International Airport	1.811	4.053	1.615	3.919
BCN	Barcelona El Prat Airport	0.644	3.448	0.461	3.522
BDL	Bradley International Airport	0.794	3.750	1.284	4.034
BKK	Suvarnabhumi Airport	1.512	3.879	1.582	4.145
BLQ	Bologna Airport	0.000	4.000	−0.222	2.840
BLR	Kempegowda International Airport	2.189	4.196	2.119	4.336
BNA	Nashville International Airport	1.340	4.076	0.687	3.882
BNE	Brisbane Airport	0.520	3.519	0.850	3.815
BOM	Chhatrapati Shivaji International Airport	2.683	4.604	2.661	4.690
BOS	Boston Logan International Airport	0.525	3.588	1.020	3.889
BPN	Sultan Aji Muhammad Sulaiman Int'l Airport	2.600	4.500	1.649	4.246
BRS	Bristol Airport	0.601	3.123	0.644	3.298

BRU	Brussels Airport	0.543	3.459	0.594	3.425
BUD	Budapest Ferenc Liszt International Airport	0.686	3.778	0.677	3.474
BWI	Baltimore Washington International Airport	1.081	3.954	0.788	3.733
CAN	Bai Yun Airport	−0.563	3.100	0.340	3.169
CCU	Netaji Subhas Chandra Bose Int'l Airport	2.078	4.141	1.851	4.121
CDG	Paris Charles de Gaulle Airport	−0.003	3.179	0.160	3.080
CGK	Jakarta Soekarno-Hatta International Airport	1.315	4.023	0.892	4.051
CHC	Christchurch International Airport	2.246	4.286	1.764	4.163
CKG	Chongqing Jiangbei International Airport	3.500	4.500	*# review insufficient*	
CLE	Cleveland-Hopkins International Airport	−0.070	3.024	−0.064	2.859
CMH	Port Columbus International Airport	1.065	3.893	1.330	3.960
CNX	Chiang Mai International Airport	0.944	4.160	1.163	3.819
CPH	Copenhagen Airport Kastrup	1.410	4.161	1.052	3.973
CPT	Cape Town International Airport	1.471	4.050	1.767	4.329
CTU	Chengdu Shuangliu International Airport	0.955	4.000	0.513	3.769
CUN	Cancun International Airport	−0.495	2.882	−0.053	3.360
CVG	Cincinnati/Northern Kentucky International Airport	0.156	3.828	0.977	3.974
DEL	Bandaranaike International Airport	1.219	3.093	1.140	3.612
DEN	Indira Gandhi International Airport	0.846	3.962	0.868	3.914
DFW	Denver International Airport	0.695	3.555	0.579	3.528
DME	Dallas Forth Worth International Airport	0.283	3.032	0.393	3.132
DMM	King Fahd International Airport	0.658	3.411	1.124	3.687
DOH	Doha International Airport	2.088	4.439	1.891	4.393
DPS	Ngurah Rai International Airport	1.058	3.917	1.446	3.965
DTW	Detroit Metropolitan Wayne County Airport	0.816	3.908	1.225	4.057
DUB	Dublin Airport,	1.591	4.291	1.540	4.332
DUS	Düsseldorf International Airport	1.145	4.021	1.250	3.613
DXB	Dubai International Airport	0.668	2.944	0.973	3.254
EDI	Edinburgh Airport	0.139	2.806	0.528	3.310
FAO	Faro Airport	−0.597	3.357	1.315	3.103
FCO	Rome Leonardo Da Vinci/Fiumicino Airport	−0.560	2.472	0.099	3.083
FLL	Fort Lauderdale Hollywood International Airport	1.084	3.598	0.303	3.342
FRA	Frankfurt Airport	0.515	3.623	0.879	3.689
GMP	Seoul Gimpo International Airport	0.111	3.417	2.889	4.091
HEL	Helsinki Vantaa Airport	1.717	4.333	0.820	3.890
HGH	Hangzhou Xiaoshan International Airport	0.500	2.333	0.333	3.000
HKG	Hong Kong International Airport	1.574	4.466	1.618	4.474
ICN	Incheon International Airport	1.774	4.376	1.573	4.352
IST	Istanbul Atatürk Airport	0.129	2.882	0.365	3.195
JNB	O. R. Tambo International Airport	1.627	4.074	1.260	3.934
KIX	Kansai International Airport	1.138	3.610	1.674	4.081
KUL	Kuala Lumpur International Airport	1.273	3.955	0.880	3.693
LAS	Las Vegas McCarran International Airport	1.338	3.963	0.821	3.820
LGW	London Gatwick International Airport	0.154	3.543	0.349	3.454
LHR	London Heathrow Airport	0.454	3.586	0.724	3.683
LIS	Lisbon Portela Airport	1.075	3.326	0.476	3.451
MAD	Madrid Barajas Airport	0.505	3.494	0.795	3.690
MAN	Manchester Airport	1.125	4.127	0.586	3.683
MEL	Melbourne Airport	0.048	3.064	0.159	3.075
MSP	Minneapolis/St. Paul International Airport	1.124	4.095	0.863	3.983
MUC	Munich Airport	1.371	4.051	1.075	4.027
MXP	Milan Malpensa Airport	0.409	3.375	0.400	3.175
NRT	Tokyo Narita International Airport	1.743	4.319	1.771	4.514
ORY	Paris Orly Airport	0.667	3.429	−0.092	3.064
OSL	Oslo Airport Gardermoen	1.232	3.841	0.954	3.455
PEK	Beijing Capital International Airport	−0.163	3.056	0.431	3.179
PMI	Palma de Mallorca Airport	0.233	3.440	0.891	3.585
PVG	Shanghai Pudong International Airport	1.126	3.789	0.657	3.514
SAN	San Diego International Airport	1.073	3.977	0.746	3.757
SEA	Seattle-Tacoma International Airport	1.023	3.879	0.712	3.653
SFO	San Francisco International Airport	1.277	3.965	1.245	4.052
SHA	Shanghai Hongqiao International Airport	0.900	4.778	0.504	3.556
SIN	Singapore Changi International Airport	1.313	3.993	1.898	4.630
SLC	Salt Lake City International Airport	1.109	4.000	1.037	3.941
STN	London Stansted Airport	−0.631	2.436	−0.390	2.468
SUB	Juanda International Airport	1.148	4.279	1.480	4.277

SVO	Sheremetyevo International Airport	0.493	3.423	0.110	3.169
SYD	Sydney Airport	0.394	3.402	0.367	3.238
SZX	Shenzhen Bao'an International Airport	1.194	3.818	0.950	3.704
TPA	Tampa International Airport	1.415	4.301	1.415	4.416
TXL	Berlin Tegel Airport	−0.067	2.890	−0.180	2.901
VIE	Vienna International Airport	1.442	4.467	1.558	4.377
WUH	Wuhan Tianhe International Airport	−0.833	2.000	−0.867	2.900
YYZ	Toronto Lester B. Pearson International Airport	0.631	3.660	0.477	3.429
ZRH	Zurich Airport	1.653	4.519	1.608	4.289

References

Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N., 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. Springer, Berlin, Heidelberg, pp. 391–402.

Bezerra, G.C.L., Gomes, C.F., 2015. The effects of service quality dimensions and passenger characteristics on passenger's overall satisfaction with an airport. J. Air Transport. Manag. 44, 77–81.

Bezerra, G.C.L., Gomes, C.F., 2016. Measuring airport service quality: a multidimensional approach. J. Air Transport. Manag. 53, 85–93.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.

Blei, D.M., Lafferty, J.D., 2009. “Topic Models”, Text Mining: Classification, Clustering, and Applications, vol. 10. CRC Press, pp. 34 No. 71.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022.

Bogicevic, V., Yang, W., Bilgihan, A., Bujisic, M., 2013. Airport service quality drivers of passenger satisfaction. Tourism Rev. 68 (4), 3–18 Emerald Group Publishing Limited.

Brady, M.K., Cronin, J.J.J., 2001. Some new thoughts on conceptualizing perceived service quality: a hierarchical approach. J. Market. 65 (3), 34–49.

Brida, J.G., Moreno-Izquierdo, L., Zapata-Aguirre, S., 2016. Customer perception of service quality: the role of Information and Communication Technologies (ICTs) at airport functional areas. Tourism Management Perspectives 20, 209–216.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S., 2009. “A density-based method for adaptive LDA model selection”. Neurocomputing 72 (7–9), 1775–1781.

Collins, C., Hasan, S., Ukkusuri, S.V., 2013. A novel transit rider satisfaction metric: rider sentiments measured from online social media data. Journal of Public Transportation 16 (2), 21–45.

D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F., 2015. Real-time detection of traffic from twitter stream analysis. IEEE Transactions on Intelligent Transportation Systems, IEEE 16 (4), 2269–2283.

Deveaud, R., SanJuan, E., Bellot, P., 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numérique 17 (1), 61–84. Available at: http://www.cairn.info/revue-document-numerique-2014-1-page-61.html, Accessed date: 14 July 2017.

Feinerer, I., 2017. Introduction to the Tm Package Text Mining in R. Gvailable at: https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf, Accessed date: 14 July 2017.

Fodness, D., Murray, B., 2007. Passengers' expectations of airport service quality. J. Serv. Market. 21 (7), 492–506.

Gitto, S., Mancuso, P., 2017. Improving airport services using sentiment analysis of the websites. Tourism Management Perspectives 22, 132–136 Elsevier.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 101. National Academy of Sciences, pp. 5228–5235 Suppl 1.

Grosenick, S., 2012. Real-time Traffic Prediction Improvement through Semantic Mining of Social Networks. University of Washington.

Gu, Y., Qian, Z.S., Chen, F., 2016. From Twitter to detector: real-time traffic incident detection using social media data. Transport. Res. C Emerg. Technol. 67, 321–342 Elsevier.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 168–177.

Költringer, C., Dickinger, A., 2015. Analyzing destination branding and image from online sources: a web content mining approach. J. Bus. Res. 68, 1836–1843 Elsevier.

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., Shoor, I., 2017. Automating a framework to extract and analyse transport related social media content: the potential and the challenges. Transport. Res. C Emerg. Technol. 77, 275–291 Elsevier.

Larose, D.T., 2005. Introduction to Data Mining. Wiley Online Library.

Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5 (1), 1–167.

Mai, E., Hranac, R., 2013. Twitter interactions as a data source for transportation incidents. In: Proc. Transportation Research Board 92nd Ann. Meeting.

Mehrotra, R., Sanner, S., Buntine, W., Xie, L., 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 889–892.

Merkert, R., Assaf, A.G., 2015. Using DEA models to jointly estimate service quality perception and profitability–Evidence from international airports. Transport. Res. Pol. Pract. 75, 42–50 Elsevier.

Mohammad, S.M., Turney, P.D., 2013. NRC Emotion Lexicon. Technical Report. National Research Council Canada, Ottawa, Canada.

Neter, J., Wasserman, W., Kutner, M.H., 1985. Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Richard D. Irwin, Homewood, IL.

Neter, J., Wasserman, W., Kutner, M.H., Li, W., 1996. Applied Linear Statistical Models. Irwin.

Nghiêm-Phú, B., Suter, J.R., 2018. Airport image: an exploratory study of McCarran international airport. J. Air Transport. Manag. 67, 72–84 Elsevier.

Nielsen, F., 2011. “Afinn”, Informatics and Mathematical Modeling. Technical University of Denmark.

Nikita, M., 2014. Tuning of the Latent Dirichlet Allocation Models Parameters [R Package Ldatuning Version 0.2.0]. Comprehensive R Archive Network (CRAN) Available at: https://cran.r-project.org/web/packages/ldatuning/index.html, Accessed date: 14 July 2017.

Pantouvakis, A., Renzi, M.F., 2016. Exploring different nationality perceptions of airport service quality. J. Air Transport. Manag. 52, 90–98.

Phan, X., Nguyen, L., Horiguchi, S., 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th, Available at: http://dl.acm.org/citation.cfm?id=1367510, Accessed date: 14 July 2017.

Rhoades, D.L., Waguespack, J.B., Young, S., 2000. Developing a quality index for US airports. Manag. Serv. Qual.: Int. J. 10 (4), 257–262.

Ruths, D., Pfeffer, J., 2014. Social media for large studies of behavior. Science 346 (6213), 1063–1064 American Association for the Advancement of Science.

Silge, J., Robinson, D., 2016. Tidytext: Text Mining and Analysis Using Tidy Data Principles in R. Available at:https://doi.org/10.21105/joss.00037.

Suárez-Alemán, A., Jiménez, J.L., 2016. Quality assessment of airport performance from the passengers' perspective. Research in Transportation Business and Management 20, 13–19.

Tufekci, Z., 2014. Big questions for social media big data: representativeness, validity and other methodological Pitfalls. ICWSM 14, 505–514.

Wattanacharoensil, W., Schuckert, M., Graham, A., Dean, A., 2017. An analysis of the airport experience from an air traveler perspective. J. Hospit. Tourism Manag. 32, 124–135.

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 347–354.

Xiang, Z., Du, Q., Ma, Y., Fan, W., 2017. A comparative analysis of major online review platforms: implications for social media analytics in hospitality and tourism. Tourism Manag. 58, 51–65 Elsevier.

Xiang, Z., Schwartz, Z., Gerdes Jr., J.H., Uysal, M., 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? Int. J. Hospit. Manag. 44, 120–130 Elsevier.

Xie, K.L., Zhang, Z., Zhang, Z., 2014. The business value of online consumer reviews and management response to hotel performance. Int. J. Hospit. Manag. 43, 1–12 Elsevier.

Yan, X., Guo, J., Lan, Y., Cheng, X., 2013. “A Biterm Topic Model for Short Texts”, Proceedings of the 22nd International Conference on World Wide Web - WWW ’13. ACM Press, New York, New York, USA, pp. 1445–1456.

Yeh, C.-H., Kuo, Y.-L., 2003. Evaluating passenger services of Asia-Pacific international airports. Transport. Res. E Logist. Transport. Rev. 39, 35–48 Elsevier.

Zhang, Z., Ni, M., He, Q., Gao, J., Gou, J., Li, X., 2016. An exploratory study on the correlation between twitter concentration and traffic surge 2. Transport. Res. Rec.: Journal of the Transportation Research Board 2553, 90–98.

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 (13), S8. Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S13-S8, Accessed date: 14 July 2017.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X., 2011. Comparing twitter and traditional media using topic models. In: European Conference on Information Retrieval. Springer, pp. 338–349. Available at: http://link.springer.com/chapter/10.1007/978-3-642-20161-5_34, Accessed date: 12 July 2017.

Zhong, N., Li, Y., Wu, S.-T., 2012. Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, IEEE 24 (1), 30–44.