Monitoring Airport Service Quality: A Complementary Approach to Measure Perceived Service Quality using Online Reviews

Completed Research Full Paper

Kiljae Lee
Embry-Riddle Aeronautical University
kiljae.lee@erau.edu

Chunyan Yu
Embry-Riddle Aeronautical University
yuc@erau.edu

Abstract

Based on 42,063 airport reviews collected from Google Maps, we conducted a sentiment analysis and a topic modeling. We showed that the sentiment scores computed from textual reviews are good estimates of their paired star-ratings (r=0.63, p<0.01). Next, using the LDA (Latent Dirichlet Allocation), we extracted latent topics from the textual reviews and compared them with the standard categories utilized in the Airport Service Quality survey (ASQ). The topics extracted from reviews correspond well with the categories used in ASQ. We, in turn, compared the online ratings with the ratings annually updated by ASQ. While online reviews discuss almost identical topics with those of ASQ, the correlation between the ratings from two was weak (r=0.2). We suggest that the text mining approach using online reviews not only provides an inexpensive, dynamic, and locally customizable means of monitoring airport quality but also complements the standard survey by offering an alternative metric.

Keywords

Airport management, Airport Service Quality, Text Mining, Online reviews, LDA, Sensitivity Analysis

Introduction

The extant literature on airport service quality, as well as the major commercial survey conducted for airports, rely mostly on offline data collected through on-site questionnaires. Given the big data trend in which a massive amount of reviews on airport services are being generated every minute through various digital and social media channels, it is imperative for the airline managers to develop methods to leverage such a trend to harvest insights directly from passengers in real time. Unlike in other disciplines, however, in the airport management studies, little research has attempted to leverage online review contents using computational approach.

Background

The increasing competition on service among airports has triggered the need for a more effective and comprehensive measure of airport quality (Rhoades et al. 2000) over the last two decades. Airport service quality is a multi-dimensional construct that represents a broad range of passenger experiences from physical facilities, interactions, and services (Brady and Cronin Jr 2001). In many instances, the perceived quality of an airport is highly subjective and context dependent (Bezerra and Gomes 2015a). Therefore, it is critical to ensure that collected data should appropriately represent the first-hand experience of passengers to generate useful managerial insights. No metrics offered by the extant literature, yet, established a dominant position and many of them are criticized for being “unable to capture the quality perceptions from the perspective of passengers"(George et al. 2013). Most of the extant research on airport service quality rely on offline data collected through on-site or mailed out questionnaires(e.g., Bezerra and Gomes 2015b; El-deen et al. 2016; Jeon and Kim 2012). These questionnaires are designed on the opinions of domain experts (e.g., Rhoades et al. 2000) or built on focused group interviews (e.g., Fodness and Murray 2007). If the metric scales misrepresent the general perception of the passengers of a particular airport or fail to capture the change of passenger expectations, those metrics may lead to a "misguided efforts" to improve the competitive service quality (Fodness and Murray 2007).

In the commercial domain, there is a dominant research standard established over the last ten years by Airports Council International (ACI). ACI has initiated an extensive annual survey program on airport service quality (ASQ) in 2006. Since then, the standardized annual survey has been consistently conducted around the world. In 2016, more than 250 airports participated in this survey program (ASQ, 2016). The airport staffs or third-party survey companies gather the survey data following the strict plan developed by ACI which also regularly audits participating airports to ensure compliance with the strict standard. The study defines 34 service areas under eight categories which include access, check-in, passport control, security, navigation, facilities, environment, and arrival (See Table 1).

OVERALL SATISFACTION	
1	Overall satisfaction with the airport
2	Overall satisfaction with the airport: business pax
3	Overall satisfaction with the airport: leisure pax
ACCESS	
4	Ground transportation to/from the airport
5	Parking facilities
6	Parking facilities value for money
7	Availability of baggage carts/trolleys
CHECK-IN (AT THIS AIRPORT)	
8	Waiting time in check-in queue/line
9	Efficiency of check-in staff
10	Courtesy, helpfulness of check-in staff
PASSPORT / PERSONAL ID CONTROL	
11	Waiting time at passport / personal ID inspection
12	Courtesy and helpfulness of inspection staff
SECURITY	
13	Courtesy and helpfulness of Security staff
14	Thoroughness of Security inspection
15	Waiting time at Security inspection
16	Feeling of being safe and secure
FINDING YOUR WAY	
17	Ease of finding your way through airport
18	Flight information screens
19	Walking distance inside the terminal

20	Ease of making connections with other flights
AIRPORT FACILITIES	
21	Courtesy, helpfulness of airport staff
22	Restaurant / Eating facilities
23	Restaurant facilities value for money
24	Availability of bank / ATM facilities/money changers
25	Shopping facilities
26	Shopping facilities value for money
27	Internet access / Wi-fi
28	Business / Executive lounges
29	Availability of washrooms/toilets
30	Cleanliness of washrooms/toilets
31	Comfort of waiting/gate areas
AIRPORT ENVIRONMENT	
32	Cleanliness of airport terminal
33	Ambiance of the airport
ARRIVALS SERVICES	
34	Arrivals passport and visa inspection
35	Speed of baggage delivery service
36	Customs inspection

Table 1. Airport Service Quality Metrics (ACI 2006 ~ 2016)

They collect data from at minimum, 350 passengers per quarter (1,400 passengers per year) per airport. The results are provided back to the participating airports, and ACI recognizes and rewards the best airports every year. The result, in turn, is extensively cited by the high ranking airports for promotion. Moreover, the data is used as a valuable reference to improve their services. Since every participating airport uses same questionnaires every year, the survey has become a de facto industry benchmark (ibid).

While the 36 questions exhaustively encompass all aspects of airport services, these questions may not be equally relevant for all types of airports all the time. It is likely that using the same metric consistently for over a decade has contributed to establishing a strong industry-wide standard. The consistency, however, might not allow sufficient flexibility to capture passenger expectations that presumably co-evolve with time, technology, and culture. Also, the exhaustiveness of questions might blur the importance of a few dominant aspects of a particular airport that passengers place high weights in their expectation and, thus, in rating their satisfaction.

Along with the standardized service benchmark, airport managers need complemental methods to monitor whether and how passengers' expectation evolves over time and whether and how the relative weights on different aspects of airport services vary depending on the size and location of the airport.

One plausible approach can be found in text mining. To make sense of the ever-increasing volume of textual data on the web, research in text mining offers various computational alternatives. Many academic fields seemingly far from computer science begin to take text mining approach to disrupt their mainstream research traditions (e.g., Jockers 2013). In the airport management studies, however, there are only a few initial attempts to takes such approaches to analyze a massive amount of online review contents on airport service quality(e.g., Bezerra and Gomes 2015b; Bilgihan, Vanja Bogicevic Wan Yang Anil and Bujisic 2015).

Methodology

In the present study, we take two text mining techniques (i.e., probabilistic topic modeling and opinion mining) to extract the key features from a large number of textual reviews and quantify the emotional valence expressed in them to complement the mainstream survey methods built on onsite questionnaires. Specifically, among many probabilistic topic modeling algorithms available to annotate large archives of documents with topical information, we take the Latent Dirichlet Allocation (LDA) model proposed by Blei, Ng, and Jordan (Blei et al. 2003; Blei 2012). Through LDA, we extract the dominant topics from the reviews grouped by size and year. The resulting topics are compared against the 36 standard questions that ACI uses in all airports. We also use the opinion mining (a.k.a., sentiment analysis) technique to computationally calculate sentiments toward the airport services from review taking a natural language processing (Liu 2012). The resulting sentiment score, as a predictor, will be regressed on the overall satisfaction score for each airport. The sentiment scores will also be regressed on the overall rating from the major commercial survey conducted during the same periods as the reviews were posted (e.g., 2014, 2015, 2016).

Data Collection

The largest number of reviews on airports in English can be found on Twitter, SKYTRAX (airlinequiality.com), and Google maps. While the Twitter provides a convenient Application Protocol Interface(API) to crawl data, we exclude Twitter from our project because it is well documented that topic modeling technique like LDA does not work well with the short messages like tweets (Mehrotra et al. 2013). Moreover, we found from our preliminary test suggests that most of the tweets that contain 'airport' keyword or airport hashtag do not include the type of evaluative message relevant to our research objective.

SKYTRAX website (Airportquality.com) has a section that exclusively holds reviews on the airport service quality. However, the number of review data SKYTRAX is relatively small for our analysis. The entire number of airport reviews between Jan 1, 2015, to Aug 1, 2015, was only 1,698. While the website is well known to professionals in the aviation industry, it has a relatively weak exposure to the general public. We presume that there is a weaker chance for casual visitors than airport professionals to leave reviews in SKYTRAX potentially leading to a result containing a stronger self-selection bias.

For this analysis, therefore, we collect Google map reviews on top 100 international airports from the ASQ metric of service quality. As of Oct 30, 2016, Google map contains 123,068 reviews on the top100 airports since 2007. The reviews on this site are mostly written by the general public who, presumably, happen to search the airport before, during, or after visiting the place. Further, compared to other online texts, such as Twitter, Google review solicits review along with a quantitative rating. This allows us to test the consistency between the valence reflected on the textual reviews and their paired quantitative ratings. We used Python to crawl the reviews systematically from Google maps.

Preliminary Analyses

In order to test the feasibility of using textual reviews as a source of quantified predictor of airport service quality, we conduct a sentiment analysis (Liu 2012)using AFINN (Nielsen 2011) sentiment lexicon in R. Sentiment analysis was performed as an inner join between tokenized list of the reviews and the sentiment lexicon which contains a list of emotionally laden keywords with a positive or a negative tag. We examine how good the computationally calculated emotional polarity scores from the airport review data are to predict the actual overall satisfactions scores.

Next, to extract topics that customers refer to when evaluating an airport, we take the Latent Dirichlet Allocation (LDA) (Blei et al. 2003; Blei 2012). LDA is an unsupervised learning algorithm designed to identify latent topics from a large set of documents without making any prior annotation of the documents. This probabilistic modeling technique is gaining an increasing popularity to make sense out of large amounts of textual contents. The algorithm assumes that each document is composed of multiple topics and each latent topic is expressed only as a collection of words. By maximizing inter-class variance, LDA estimates the probabilities of these topics and words at the same time.

Preliminary Results of Sentiment Analysis

From the 123,068 Google reviews for 100 airports, we take 42,063 records for the analysis excluding 81,005 non-English reviews. Three sentiment lexicons are available as a dataset in tidytext package in R. The three lexicons are NRC Emotion (Mohammad and Turney 2013) and Bing Liu (Hu and Liu 2004), AFINN (Nielsen 2011). We use AFINN to compute per review sentiment and then compute per airport sentiment. As expected, the average of the per review sentiment scores from the Google review texts was strongly correlated with Google star ratings, r=.63, p<0.01 (See Figure 1).

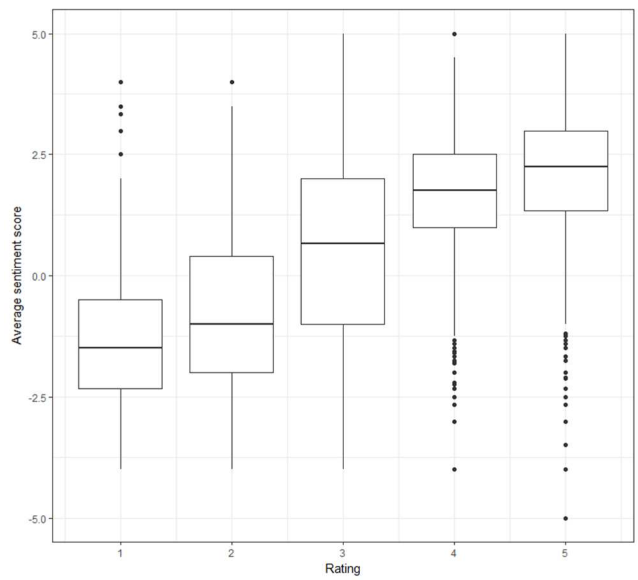

Figure 1. Average Sentiment Score vs. Average Star Ratings

The correlation between per airport sentiment scores from Google reviews and per airport Google star ratings is significantly high, r(97)=.88, p<0.01. This suggests that the emotional valence reflected in customers' review texts is a good estimator of their overall evaluation of the airport that they marked as a Google star-rating score. The keywords that frequently associated with either positive or negative valence are identified as shown in Figure 2. For instance, keywords such as "security," "waiting," "baggage" are identified as valence neutral. While "clean," "navigate," and "facilities" are associated more with a positive valence, words such as "attitude," "joke," and "customer" are more often associated with negative valence.

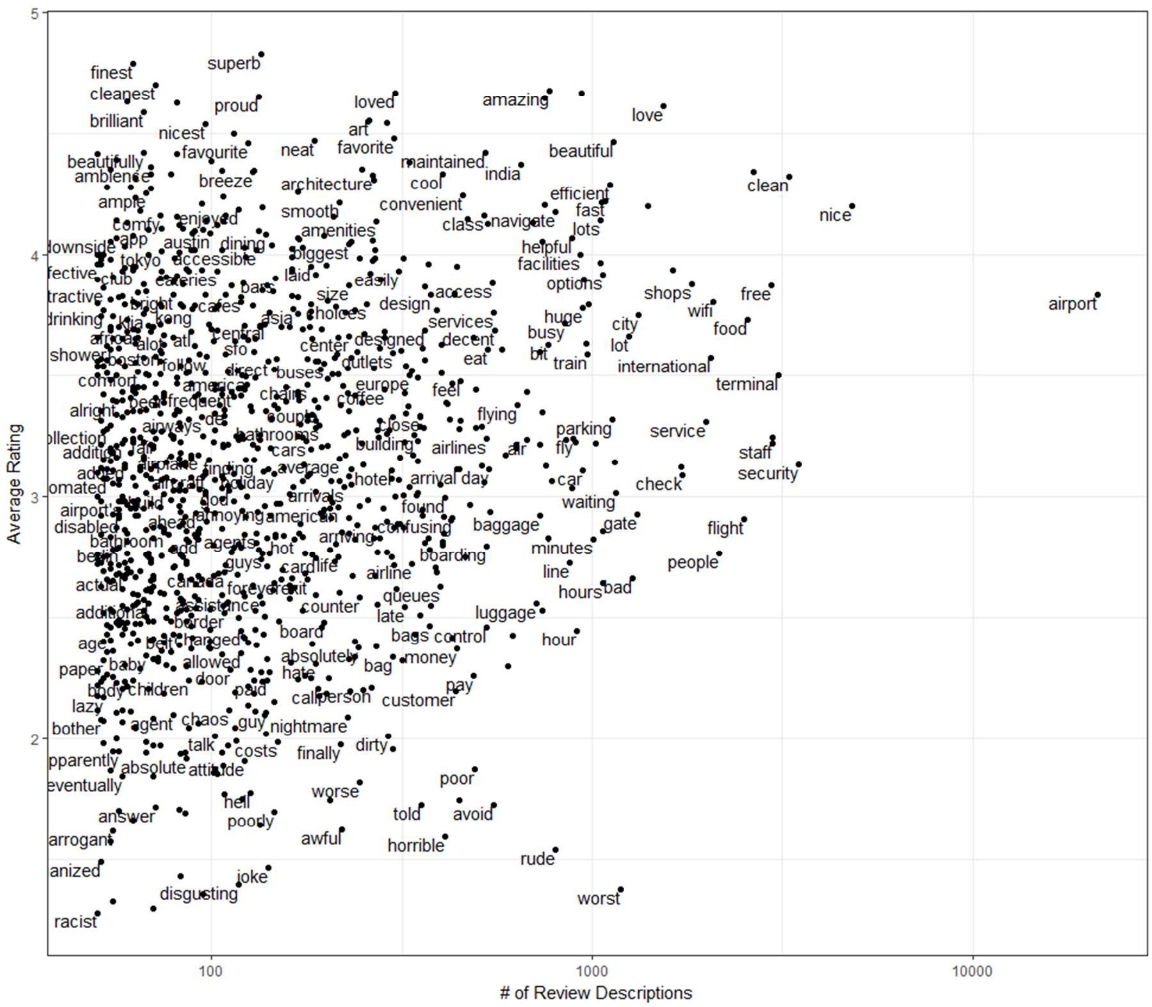

Figure 2. Average Ratings vs. Word Frequencies

However, the correlation between Google star-ratings and ASQ ratings[1] is at a suboptimal level, explaining very little variability for each other, r (95) =.21, p=.04, R^2 = 0.03. This, tentatively, suggests that Google reviewers (English speaking group in this case) and ASQ survey participants (the majority of which are local) do not necessarily agree with each other on the quality of service of an airport.

Preliminary Results of Topic Modeling

Fitting the LDA model to 42,063 review records (written in English) results in 20 topics as listed in Table 2.

[1] Since we have access to only Year 2013 and 2014 data from ASQ, this comparison was made between the Google reviews in 2014 and ASQ survey on 2014. More detailed comparison analysis using ASQ 2015, 2016 will be followed in the completed work.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	parking	world	international	time	staff	shops	security	experience	food	customer
2	car	excellent	fast	travel	helpful	restaurants	luggage	layover	options	told
3	signs	taxi	terminals	boarding	baggage	dutyfree	bags	walking	gates	found
4	avoid	class	quick	takes	rude	lounge	flying	money	organized	information
5	rental	country	super	day	home	departure	leave	shop	crowded	service
6	pay	traffic	ago	shuttle	plane	arrival	bag	trip	transit	ticket
7	drop	designed	convenient	pass	english	coming	arrive	employees	poor	left
8	pick	building	top	fine	hard	seating	stay	traveling	toilets	phone
9	short	passenger	major	stop	compared	main	hotel	extra	stores	person
10	charge	visited	months	stars	claim	space	planes	price	water	lost
	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
1	wifi	love	people	clean	terminal	worst	nice	gate	huge	friendly
2	free	beautiful	check	easy	flight	feel	service	wait	bit	bad
3	city	pretty	minutes	modern	flights	horrible	lot	train	tsa	lines
4	lots	shopping	line	fly	waiting	terrible	times	walk	plenty	slow
5	efficient	amazing	hour	navigate	hours	expect	decent	passengers	expensive	extremely
6	awesome	inside	immigration	bus	airlines	outlets	services	station	access	run
7	facilities	maintained	passport	cool	connecting	coffee	visit	close	eat	transfer
8	busy	loved	customs	selection	connection	dirty	layout	night	confusing	process
9	fat	wonderful	control	spacious	domestic	floor	delays	plane	public	europe
10	comfortable	live	queue	favorite	airline	absolutely	simple	spend	design	job

Table 2. 20 topics with 10 top keywords

Since each topic, expressed as a collection of words, is inherently latent, not all topics can be briefly verbalized without losing its underlying conceptual structure. To facilitate conceptualizing the topics, we may refer to the per-topic-per-word probabilities (i.e., *beta*) as shown in Figure 3.

We map these 20 topics with the 36 service areas that ASQ has used as their base categories of airport service for the last ten years (See Table 3). Each author performed this task individually, and then the results are combined with discussion. Overall, the extracted topics nicely correspond to the categories of ASQ survey. For instance, Topic 1 illustrates the experience of airport parking which corresponds to category 4 and 5 of ASQ. Both Topic 7 and Topic 19 involves security check process represented which correspond to category 13,14, 15, and 16 of ASQ. While the reviews having a high score on Topic 7 are focusing on the experience of security check process, the reviews high on Topic 19 discuss the overall interactive experience dimension which includes the interaction with TSA.

Of note, three survey categories of ASQ do not directly correspond to one of the 20 extracted topics. These three categories (7. availability of baggage carts, 24. Availability of bank/ATM facilities and Money changers; 28. Business/Executive lounges) appear to be either negligible (7) or specific to a smaller group of passengers (24, 28).

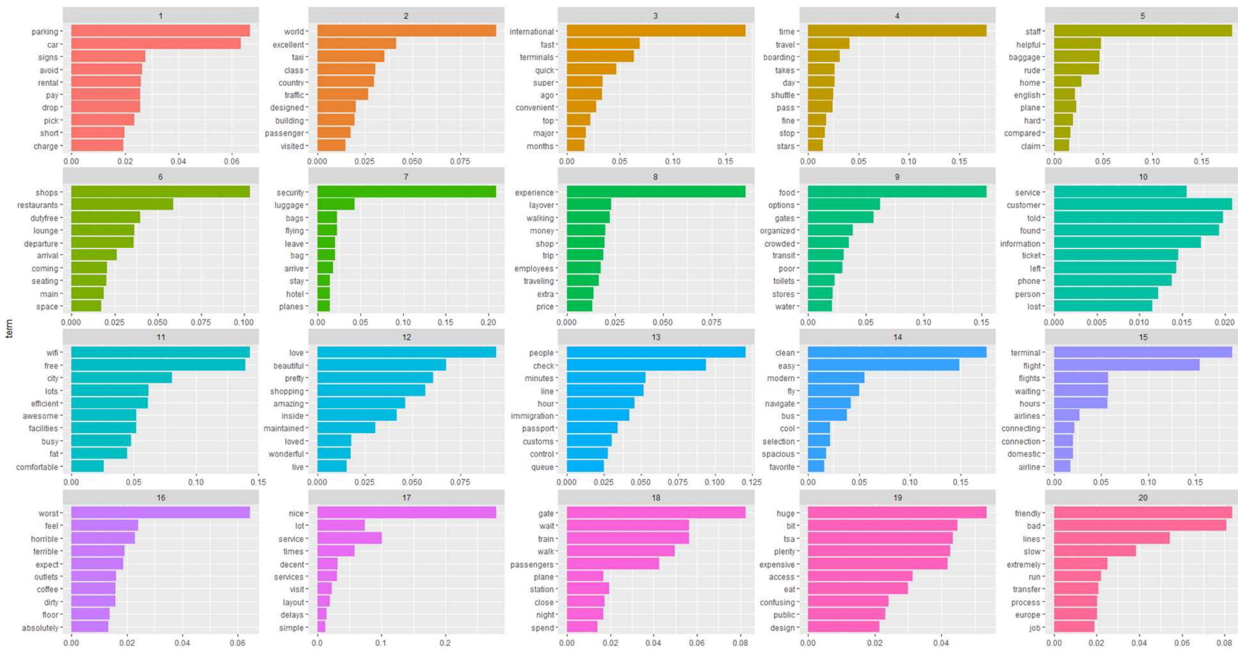

Figure 3. 20 topics 5 top terms by beta

Planned Analyses

In the next stage, a relative weight of each topic from the entire passengers' review corpus will be obtained. A trend analysis will be performed to test whether there have been any thematic changes that should be considered in the structure of the commercial survey. Further, the sub-group topic analyses will be conducted to examine whether there is a significant difference in topics as well as weights depending on the size of the airports.

Conclusion

Most research on airport service qualities relies on offline data collected through on-site questionnaires. In commercial domain, the mainstream surveys on airports are not sufficiently dynamic to capture the passenger expectations that rapidly co-evolve with time and technology. Also, these are not necessarily relevant for aff the airports of different size and location to isolate the key areas to improve the passengers' perception of a particular airport. Through this study, we seek to demonstrate that the aggregated voice of passengers is a good predictor of customized insights to complement the existing commercial survey approaches.

The present study collects online reviews from Google maps. First, from the crawled corpus, we compute the aggregated sentiment scores per review per airport. These scores were highly correlated with the reviewers' quantitative ratings suggesting that the emotional valence expressed in passengers' review can be used a reliable estimate of their quantitative ratings. Next, we extracted latent topics from the review dataset using LDA (Latent Dirichlet Allocation) model. As expected, the algorithmically identified topics from the airport reviews match well with the conventional categories used in the mainstream commercial survey (i.e., Airport Service Quality by ACI). However, the correlation between the ratings from the online reviews and the ratings from ASQ that uses the traditional survey are not impressive. These results suggest that the text analysis of airport reviews does provide an inexpensive and dynamic alternative of monitoring airport service qualities. It may provide a benchmark to critically evaluate the results of the mainstream commercial survey and offer locally customizable insights that may not be readily available from the globally standardized approach only. However, given the low correlations between the two ratings (Google review vs. ASQ survey), in practice, the results from each approach should complement, rather than replace, each other.

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

Table 3. 6 ASQ categories vs. 20 topics extracted from airport reviews

REFERENCES

Bezerra, G. C., and C. F. Gomes. 2015a, "The Effects of Service Quality Dimensions and Passenger Characteristics on Passenger's overall Satisfaction with an Airport," *Journal of Air Transport Management* (44:pp. 77-81.

Bilgihan, Vanja Bogicevic Wan Yang Anil, and M. Bujisic. 2015, "Airport Service Quality Drivers of Passenger Satisfaction," .

Blei, D. M. 2012, "Probabilistic Topic Models," *Communications of the ACM* (55:4), pp. 77-84.

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

Brady, M. K., and J. J. Cronin Jr. 2001, "Some New Thoughts on Conceptualizing Perceived Service Quality: A Hierarchical Approach," *Journal of Marketing* (65:3), pp. 34-49.

El-deen, R. M. B., S. B. Hasan, and N. M. Fawzy. 2016, "The Effect of Airport and in-Flight Service Quality on Customer Satisfaction," *Journal of Faculty of Tourism and Hotels, Fayoum University* (10:1/2), .

Fodness, D., and B. Murray. 2007, "Passengers' Expectations of Airport Service Quality," *Journal of Services Marketing* (21:7), pp. 492-506.

George, B. P., T. L. Henthorne, and T. R. Panko. 2013, "ASQual: Measuring Tourist Perceived Service Quality in an Airport Setting," *International Journal of Business Excellence* (6:5), pp. 526-536.

Hu, M., and B. Liu. 2004. "Mining and Summarizing Customer Reviews," pp. 168-177.

Jeon, S., and M. Kim. 2012, "The Effect of the Servicescape on Customers' Behavioral Intentions in an International Airport Service Environment," *Service Business* (6:3), pp. 279-295.

Jockers, M. L. 2013. *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press.

Liu, B. 2012, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies* (5:1), pp. 1-167.

Mehrotra, R., S. Sanner, W. Buntine, and L. Xie. 2013. "Improving Lda Topic Models for Microblogs Via Tweet Pooling and Automatic Labeling," pp. 889-892.

Mohammad, S. M., and P. D. Turney. 2013, "No Title," *Nrc Emotion Lexicon*.

Nielsen, F. 2011, "Afinn," *Richard Petersens Plads, Building* (321:.

Rhoades, D. L., B. Waguespack Jr, and S. Young. 2000, "Developing a Quality Index for US Airports," *Managing Service Quality: An International Journal* (10:4), pp. 257-262.