Use of Electronic Word of Mouth as Quality Metrics: A Comparison of Airline Reviews on Twitter and Skytrax

Lin Lu
Fairfield University
llu@fairfield.edu

Amit Mitra
Auburn University
mitraam@auburn.edu

Yen-Yao Wang
Auburn University
yzw0008@auburn.edu

Yu Wang
Auburn University
yzw0049@auburn.edu

Pei Xu
Auburn University
pzx0002@auburn.edu

Abstract

User-generated content (UGC) at online platforms serves as a critical data source in the service industry as it can be accessed in real-time and reflect customers' changing focus on service aspects. Drawing upon the importance-performance analysis framework, we propose a methodology to derive service quality metrics by utilizing the heterogeneous sources of UGC with customized text mining techniques and examining the effectiveness of these quality metrics. UGC data related to major U.S. airlines were collected from non-social media (Skytrax) and social media platforms (Twitter) from 2014 to June 2019. The results suggest that the topic distributions and the UGC-derived weighted service quality (WSQ, which represents the weighted sentiment based on service aspects) significantly vary between the non-social media and social media platforms. In addition, the WSQ scores derived from two platforms are significant indicators of the objective service quality measurement (i.e., airline quality rating) with stronger predictive power from the social media derived WSQ score.

1. Introduction

Service quality is essential for retaining customer patronage ([1], [2]) and market share [3]. The provision of supervisor service quality is essential for the airline industry as multiple services encounter stages that can affect passengers' satisfaction towards partial or overall aspects of service [4].

Subjective perceptual surveys and objective criteria are two common approaches to measure service quality. Scholars have either leveraged the existing instruments (e.g., SERVQUAL [5]); SERVPERF [6]) or developed new instruments to measure service quality in the airline industry (e.g., AIRQUAL [7]). However, the survey approach is prone to many kinds of drawbacks and biases such as the time consumed to collect complete datasets [8], sample size limits [9], the limited research expandability [10], and social desirability bias [11]. Notably, in the fast-paced airline industry, the timeliness of survey-based results becomes a challenging issue [12].

The Airline Quality Rating (AQR) combining multiple operational performance such as on-time performance, overbooking, mishandled baggage, and customer complaints serves as an alternative and objective method for assessing airline service quality. Although this approach provides a periodic, objective, and comparable basis, scholars still argue that customer perception rather than operational performance drives customers' attitudes about service quality (e.g., [13]). Given the potential biases and the lack of timeliness that the survey approach may involve in measuring service quality, several researchers have called for more novel approaches to better measure service quality in the airline industry (e.g., [4], [14]).

Nowadays, as review websites and social media have become potent channels for consumers to post service-related issues and/or rate their satisfaction, the proliferation of user-generated content (UGC) or online word of mouth (WOM) across these channels offers an unprecedented opportunity to collect and monitor customer feedbacks ([15], [16]) and predict service quality [14]. The online communications between individuals concerning their perceptions of goods and services [17] come in many forms such as consumer reviews, microblogging, or expert blogs [18]. Popular web platforms such as Skytrax, TripAdvisor, Google Reviews, etc., let travelers leave star ratings and reviews about various aspects of airline services. In addition, a majority of airlines have maintained social media accounts (e.g., Twitter) to interact with their customers.

Compared to surveys and operational measures, UGC is publicly accessible, can be collected in real-time, and can reflect customers' dynamic focus on service. It also eliminates the response style biases and sampling issues due to the limited coverage of customers [14]. More importantly, since effective measurement of service quality must be based on customers' experiences [2], this makes UGC particularly suitable.

URI: https://hdl.handle.net/10125/79497
978-0-9981331-5-7

The purpose of this paper is to propose a methodology to derive service quality metrics by utilizing the heterogeneous sources of UGC with customized text mining techniques and statistical methods to examine if UGC derived metrics can serve as a valid measure of airline service quality. Specifically, we aim to answer the following research questions: *(a) What are the service aspects discussed in UGC over the past years? (b) Are WSQ derived from UGC on the social media different from that of the non-social media platforms? (c) Is there a significant predictive relationship between WSQ and AQR? and (d) Are social media metrics relatively stronger indicators of AQR compared with non-social media metrics?*

The remainder of this paper is organized as follows. The following section provides background and related works. Section 3 presents the research framework and hypotheses development. Section 4 details the methods we use in data acquisition, data preprocessing, sentiment analysis, topic modeling, and hypotheses testing. We discuss our experimental results in evaluating the utility of online reviews on service quality in Section 5. The last two sections discuss the findings and the directions for future research.

2. Background and Related Works

Although UGC may offer new opportunities for measuring service quality, it imposes certain challenges. For example, with the high velocity of UGC, previous reviews are soon buried in the large wave of new reviews [17]. In addition, star ratings of reviews have several limitations. First, there may be biases in star ratings based on their published sources [19] or star ratings do not match the review sentiment [20], which lowers its reliability. Further, star ratings may not be applied to a specific part of a document and are typically missing in certain forms of reviews such as tweets [18]. If using these raw reviews and ratings, travelers and service providers would not efficiently exploit the rich information [21].

The above limitations highlight the capacity of text mining. Text mining belongs to data mining that aims at extracting information from texts [22] through which numerous measures can be collected or derived from UGC, including textual features such as the length of a review, semantic features such as words and topics, sentiment feature that assess consumer emotional polarity towards a specific topic [23], and other features such as ratings (e.g., review websites provide) and reviewer identity [24]. Topic modeling and sentiment analysis are two main methods that help derive the above semantic and sentiment features, respectively. These analyses provide approaches to analyze online reviews.

Existing service quality literatures using UGC typically employ a sample of UGC to extract features or measures that allow for detecting, describing or predicting meaningful patterns. [25], [26], and [27] utilized tweets to identify polarity directions based on classified service attributes. [28], [29], and [23] collected Skytrax reviews to seek for relationships between service aspects performance and customer satisfaction. Other studies ([30], [31], and [4]) used data from the review website such as TripAdvisor to measure service quality using unstructured data and converting them into managerial insights. Only few studies (e.g., [14], [21]) combined data from multiple sources to compare online review derived measures with ratings and/or operational metrics.

While the above line of research has generated novel insights, our understanding regarding the effectiveness of UGC on airline service quality has yet to receive systematic scrutiny for the following reasons. First, prior research tends to ignore the heterogeneous nature of UGC in favor of a single data source of UGC. Given the heterogeneous nature of UGC across online platforms [32] this approach will miss unique contributions from heterogeneous data sources to form more meaningful and comparable quality metrics. Second, the size, time-lapse, and sampling of UGC varies greatly, which significantly limits the data quality and their generalizability and contribution to knowledge. Third, service aspects and satisfaction are more likely to be discussed as two issues. Thus, there is a lack of an integrated measure that considers both the importance and performance of service aspect. Finally, implementation of text mining methods such as topic modeling and sentiment analysis may need careful customization based on the differences of textual features and research objectives. However, there is a lack of explorations on how customizing the text mining approaches will make a difference in results.

In this paper, we leverage Importance-Performance Analysis (IPA) ([33], [34]) as our theoretical framework to address the research objectives. These analyses can offer complementary indicators to extract the key attributes of service quality perceived by passengers that can be compared between platforms, to cross-validate with AQR results, and to expand the coverage of analysis beyond participating airlines in the AQR program.

3. Hypotheses Development

To derive the valid UGC metrics on measuring service quality, we leverage IPA ([33], [34]) as our theoretical framework. The IPA framework is used to evaluate the attributes of a product or service based on measures of their importance and performance from the perceptual viewpoint of the customer ([34]). It is a commonly used framework for understanding customer satisfaction and guiding strategic planning schemes. The IPA framework considers importance as a weight of a service aspect's performance to identify areas that need managerial attention. Figure 1 shows the IPA grid with four quadrants divided by the two dimensions.

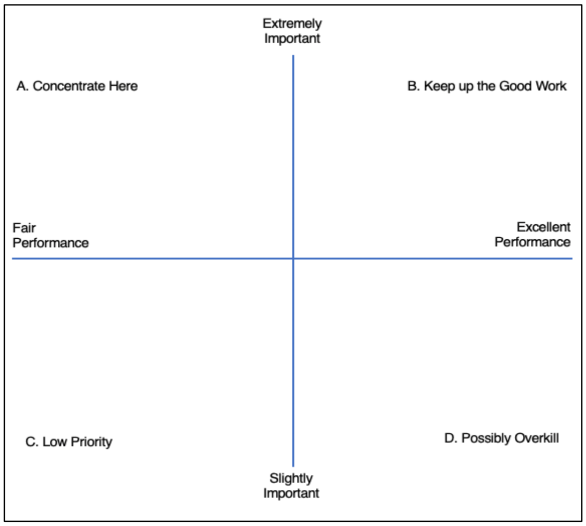

Figure 1. Importance-Performance analysis grid [34]

For IPA, determination of the importance of a service aspect is a crucial task [35]. When sharing opinions about service-related issues at online platforms, customers may discuss various topics from customer service, booking experience, connection experience, boarding experience, to baggage experience. We posit that the importance of these service aspects varies across non-social media and social media platforms, namely, the distribution of service topics is different across platforms.

For the non-social media platforms that dedicate to review, customers typically review multi-aspect of service as the text entry can reach a maximum of 3500 characters and the platform also asks for ratings in aspects of ground service, seat comfort and cabin staff service, etc. For example, Skytrax, one of the largest airline review sites, is an international air transport rating organization that provides the world's only airline quality ranking business through which passengers not only could find the relevant airline or airport information but also share their personal experience evaluation for their service [29]. On the other hand, social media platforms such as Twitter are considered a microblogging platform, which offers a fast-paced way to discover new content and see what is trending [32]. The unique nature of Twitter is also attractive to customers due to its fast response to receive service [36]. In addition, Twitter is characterized by a simplistic design of 140 characters limit of a post and has been one of the world's most popular social media channels that appeals to both airlines and customers in sharing information and interacting in real-time to address issues before, during, and after service [37]. Given the distinct characteristics of non-social media and social media platforms, we hypothesize:

H1: The distributions of service topics are different between social media and non-social media online platforms.

In addition to the importance of a service aspect, determination of the performance of the important service aspect is another critical task for IPA. We use the average sentiment of a topic to measure the performance of a service aspect. When customers discuss a service-related topic at online platforms, the sentiment of this topic reflects the customers' perceptions of the service concerning the related attributes [35]. For example, if a customer has a good experience with a service, this customer may share this experience positively. Otherwise, this customer may discuss this experience negatively if he/she decides to share it on online platforms. Therefore, the sentiment of a service topic could be considered as the impression that a customer has of the company during the journey of receiving or requesting a service. It can be regarded as the actual performance of the service concerning the related attributes [35].

An important service aspect and its performance (i.e., the average sentiment of a topic) just indicate a single aspect of overall service consumption. Service production and consumption often unfold over a series of consumption episodes [38] and require customer to engage in multiple service encounters in an extended period of time [39]. Thus, at the aggregate level of service quality metrics, we need to consider different attribute weights to better understand the overall service quality perceptions [39]. We propose that the WSQ scores summation of importance times performance for all topics allow us to better capture different service attribute weights. We further posit that the WSO scores vary across non-social media and social media platforms. As discussed, the fast-paced nature of Twitter triggers both airlines and customers to interact in real-time to address issues before, during, and after service. Thus, promptness in addressing issues in a timely manner will be key when it comes to the service space in the Twitter setting [36]. Although users of review sites (i.e., non-social media platforms) also expect a managerial response, promptness is not a key to write reviews [40]. Therefore, it is plausible to hypothesize:

H2: The WSQ scores for an airline are different between the two types of online platforms.

Prior research has indicated the predictive power of UGC in various settings such as customer engagement (e.g., [41]), customer acquisition (e.g., [42]), firm equity value (e.g., [43], offline sales (e.g., [44]), and quality management (e.g., [45]). For example, [45] demonstrates how to quantify UGC and extract important features to discover and analyze product defects. Drawing upon prior research, we also expect that the WSQ scores derived from UGC can have good predictive power on the objective measure of airline quality (i.e., AQR) because these provide helpful information for firms to realize which service attributes are essential, their performance, and any service attributes that firms need to make an improvement. More importantly, we argue that the WSQ score derived from social media platforms is a better indicator of AQR than non-social media platforms. Social media metrics tend to be more socially contagious than non-social media platforms [43]. In addition, the fast-paced nature of Twitter may trigger firms to respond more quickly to service attributes that need immediate attention. We hypothesize:

H3: The WSQ scores from social media platforms is a better indicator of AQR than non-social media platforms.

4. Methods

We propose a data-driven approach that consists of four main phases as shown in Figure 2 to answer the research questions. In phase I, the data is collected from three web sources, which include user-generated text reviews from Skytrax (also known as airlinequality.com) and Twitter, and numerical values summarized from Airline Quality Rating (AQR). The text reviews will then enter preprocessing steps to prepare for: sentiment analysis in phase II and topic modeling in phase III. Quality metrics derived from the online reviews are finally constructed and compared by platforms in Phase IV. The following subsections details the procedures and methods used in each phase.

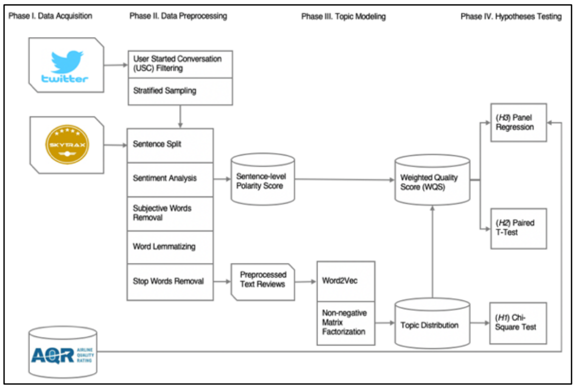
Phase I. Data Acquisition
Phase II. Data Preprocessing
Phase III. Topic Modeling
Phase IV. Hypotheses Testing
twitter
SKYTRAX
User Started Conversation (USC) Filtering
Stratified Sampling
Sentence Split
Sentiment Analysis
Subjective Words Removal
Word Lemmatizing
Stop Words Removal
Sentence-level Polarity Score
Preprocessed Text Reviews
Word2Vec
Non-negative Matrix Factorization
Weighted Quality Score (WQS)
Topic Distribution
(H3) Panel Regression
(H2) Paired T-Test
(H1) Chi-Square Test
AQR AIRLINE QUALITY RATING

Figure 2. Analyses flow

4.1. Data Acquisition

Our dataset consists of customer reviews during January 1, 2014 – June 30, 2019 from Skytrax and Twitter for 10 U.S. airlines. The exact review start date for each airline company varies based on when the first comment occurred in each web platform. Tweets that @ the official airline accounts of interests (@AlaskaAir, @Allegiant, @AmericanAir, @Delta, @FlyFrontier, @HawaiianAir, @JetBlue, @SouthwestAir, @SpiritAirlines, @united) are crawled using a Python module called Twitterscraper [45]. We scraped the tweets excluding retweets. The AQR data is extracted from annual reports at airlinequalityrating.com as an objective service quality measure.

4.2. Data Preprocessing

User-generated text reviews typically contains a lot of noises and can be expressed in a variety of ways such as using different tense and words yet for a similar meaning. Preprocessing is preparing the raw data to be technically correct, consistent, and informative for the analyses. While the two online platforms, Skytrax and Twitter, exhibit different characters in their review volumes and contextual features. For instance, a Skytrax review can be much longer than a tweet while the count of its reviews is often far less than tweets within a time period.

To equally represent online reviews from both platforms and achieve comparable results, both customized and common preprocessing treatments are applied on reviews collected from the two types of platforms. Firstly, all Skytrax reviews are preserved to capture customer voices from the non-social media platform while only tweets originated from customers are selected and we call these tweets as user-started conversations (USC). Postings initiated by the official twitter accounts are screened out to eliminate reviews that not directly relate with customer perceived service. Stratified random down sampling by airline, year, and month is conducted for tweets to achieve a similar order of magnitude with Skytrax in terms of the number of reviews and to allow for efficient data explorations. For both platforms, each review is then split into sentences, which later become the unit of analyses. The sentence-level data provides a consistent form of text observations and avoid the potential bias on sentiment analysis and topic modeling caused by the review length difference [18] between the two platforms.

A polarity score is then calculated for each sentence. The subjectivity score and subjective words within each sentence are also identified. We propose the removal of subjective words to reduce the chance of topic clusters forming by similar polarity. Specifically, positive words may form a cluster themselves, and so do the negative words. To examine the effectiveness of this subjective words' removal on topic modeling, the topic clusters derived by both non-removal and removal of subjective words are compared.

There are a number of tools available that can implement sentiment analysis. The approach used in this paper is through the Python TextBlob package for sentence split, polarity scoring and subjectivity words detection [47]. TextBlob has been extensively validated by previous research and exhibit significantly higher accuracy for short reviews over medium and long reviews [18]. Finally, word lemmatization by tag is conducted utilizing Wordnet Lemmatizer. Stop words and symbols are removed to obtain the informative preprocessed text. In reflecting these changes, both original and cleaned reviews are saved in the resulting dataset.

4.3. Topic Modeling

We use the non-negative matrix factorization (NMF) to realize topic modeling based on features selected from term frequency inverse document frequency (TF-IDF). The optimal number of clusters is determined by the topic coherence calculated through our trained word2vec model [48]. The word2vec model organizes words from the preprocessed reviews in a 500-dimensional space and that semantically similar words are close to each other. The topic coherence is a quantitative measure to evaluate if the topics are meaningful by calculating the average similarity between all pairs of the top-n words describing the topic. Two authors, with a background in business, reviewed the featured words and sentences from each topic cluster. If reaching a consensus, the subjective labeling is given; else, the third author will join the discussion until a consistent labelling is reached.

4.4. Hypotheses Testing

Based on sentiment analysis and topic modeling, each review sentence derives a polarity score and is clustered into one of the k topics. We propose that for each topic, its mean polarity indicates the performance, and its weight or ratio presents the importance of that service aspect. Then the WSQ score is calculated using the summation of each topic's mean polarity multiplied by its weight. Note that the sentiment and the weight of a topic as well as the WSQ score are all aggregated monthly for each airline to cancel out potential short-term impacts. Descriptions and symbols of these UGC derived service quality metrics are detailed in Table 1, and they serve as the numerical basis for the following hypotheses testing.

Table 1: Service quality metrics

$Note: s_{ij}\ denotes\ the\ sentiment\ of\ sentence\ j\ in\ topic\ i$

Metrics	Descriptions	Symbols
Service aspect	Topic cluster	$i = 1, 2, \dots, k$
Importance of a service aspect	Percentage of sentences in topic i among all sentences in k topics	$p_i = \frac{count\ of\ sentences\ in\ topic\ i}{total\ count\ of\ sentences\ in\ k\ topics}$
Performance of a service aspect	mean sentiment of topic i	$s_i = \frac{\sum_{sentence\ j\ in\ topic\ i}(s_{ij})}{total\ count\ of\ sentences\ in\ topic\ i}$
WSQ	summation of importance * performance of topic i	$\sum_{i=1}^{k}(p_i * s_i)$

To test if the online review topics distributions significantly vary between the two platforms (hypothesis H1), Chi-square test is used based on the crosstab of the count of sentences. The paired T-test checks for whether the non-social media and social media platforms significantly differs in WSQ scores (hypothesis H2). Finally, we examine the relationships between online review derived quality metrics and AQR to check if WSQ score from social media platforms is a better indicator of AQR than non-social media platform (hypothesis H3). This is achieved via panel regression models to identify the effects, significance, and variances explained from WSQ.

5. Results

5.1. Data Description

The count of preprocessed sentences of each airline by platform is shown in Table 2. All Skytrax reviews enter the preprocessing steps while tweets are stratified random sampled by year and month based on each airline's total number of USC. Specifically, Allegiant, FlyFrontier and SpiritAirlines are sampled at 10% and all other airlines use a sampling rate at 1%. This sampling mechanism is to maintain the information of tweet volume by time and bring a similar number of sentences from the two types of platforms for each airline.

Table 2: Count of preprocessed sentences from online reviews

Airline	Skytrax (all)	Twitter (r.d. sampled)
AlaskaAir	3,870	4,744
Allegiant	10,192	9,121
AmericanAir	29,787	26,983
Delta	13,867	27,489
FlyFrontier	14,367	24,771
HawaiianAir	1,892	791
JetBlue	4,426	9,037
SouthwestAir	6,641	18,831
SpiritAirlines	26,889	26,646
United	27,436	25,330

5.2. Topics Distributions by Platforms

We run topic modeling with a number of clusters ranging from three to seven. The NMF model with five topics is selected as the optimal. Its mean topic coherence reaches the highest as shown in Figure 3, which indicates the formation of five clusters provides a more meaningful topic identification compared with other number of clusters. Also, the top featured words and reviews describing each topic provide clear clues for manually assigning topic labels.

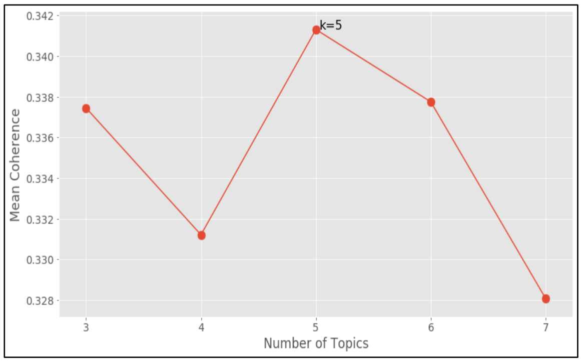

Figure 3. The optimal number of clusters

From Figure 4, customer service (topic 1) has been a hot topic among online reviews over the years. This is consistent with the report that 43% of airlines made social media customer service a top priority in 2018. The second dominant topic is boarding and baggage (topic 3). The rest of the reviews revolve around flight booking and connection (topic 0), general flight experience (topic 2), and flights' on-time performance (topic 4). We also find that the topic distribution stayed relatively constant over the five-and-half years as only slight ratio changes occurred.

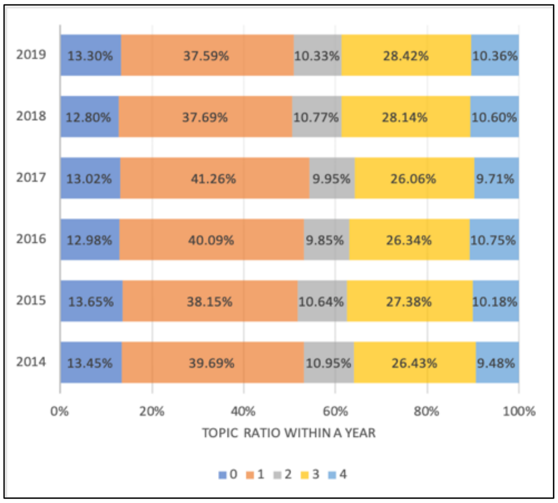

Figure 4. Topic distributions of all reviews

Both types of platforms have a higher percentage of reviews relating to customer service (topic 1) and boarding or baggage issues (topic 3), as shown in Table 3. For individual difference between platforms, the Chi-square test statistic is 10525.04 with degrees of freedom = 4 (p-value < 0.01). This indicates that topic distributions between Skytrax and Twitter are significantly different as the count of sentences for several topics differs greatly by the platform. Twitter has dominant reviews in customer service while Skytrax reviews mention boarding and baggage the most. This finding is consistent with the characters of the social media platform, which is more efficient in communicating service-related issues and get airline responses. This result supports H1.

Table 3: Topic distributions by platforms
(Unit: # - count of preprocessed sentences)

Cluster	Top featured words	Topic	Skytrax #	Twitter #
0	flight, cancel, book, attendant, day, crew, miss, make, connect, pm	Booking & connection	20,780	20,401
1	service, customer, experience, call, care, help, phone, staff, lack, line	Customer service	40,805	81,707

2	fly, time, experience, year, travel, every, use, arrive, today, make	General experience	18,088	14,483
3	get, seat, plane, go, bag, pay, check, tell, help, gate	Boarding & baggage	43,947	40,931
4	delay, hour, wait, plane, airport, sit, minute, two, another, gate	On-time performance	15,747	16,221

5.3. WSQ Scores

We obtain 660 observations (66 months * 10 airline companies) with the derived WSQ. The value of WSQ scores ranges from -1 to 1 as the percentage of a topic and sentiment score is from 0 to 1, and -1 to 1, respectively. The paired T-test result (t statistics as -3.49 and p-value < 0.01) shows a significant difference of WSQ scores between Skytrax and Twitter, in which the non-social media platform has an average of 0.015 lower WSQ score than the social media platform. This result supports H2.

5.4. Predictive Power of WSQ for AQR

The third hypothesis targets on whether the online review derived metrics, WSQ, can be a good indicator for the objective airline service quality measure, AQR. To examine the predictive power of WSQs on AQR, we use panel regressions and set the WSQ scores derived from Skytrax and Twitter as explanatory variables to predict AQR, respectively. The time index is by month, and the observation index is by airline company. To form a balanced panel data, observations from Allegiant and SpiritAirlines are excluded since they do not have complete AQR scores during 2014 – 2019. This brings 528 observations (66 months * 8 airline companies) with the records of WSQs and AQR. Table 4 provides the panel data preview.

Table 4: WSQ by companies and platforms

(Use 2014.01 as example, has a total of 528 rows)

Year	Month	Company	WSQ_ Skytrax	WSQ_ Twitter	AQR
2014	1	AlaskaAir	0.26	0.19	-0.87
		AmericanAir	0.07	0.07	-1.66
		Delta	0.12	0.06	-1.46
		Flyfrontier	0.03	-0.02	-1.74
		HawaiianAir	0.06	0.25	-0.59
		JetBlue	0.06	0.07	-1.44
		SouthwestAir	0.07	0.05	-2.13
		United	-0.01	0.07	-2.8

WSQ scores derived from Skytrax and Twitter are used as the explanatory variable in predicting AQR, respectively. The Lagrange Multiplier (LM) tests indicate that for both sets of models, fixed effects and random effects models are more appropriate than pooled ordinary least squares (OLS). And the Hausman tests suggest fixed effects models are better off random effects models. Results from fixed effects models indicate that WSQ scores from both platforms has positive effects on AQR. And WSQ score in Twitter can be a better indicator of AQR than WSQ score derived from Skytrax based on its significance. This result supports H3. The R-Squares show the between estimator can explain 73.6% or 61.6% of the between variation, and the fixed effects estimators can explain 0.6% and 0.9% of within variation. These results are shown in Table 5.

Table 5: Estimation results for panel regression models

(***, **, * indicates p-value < 0.001, <0.01, <0.05, respectively)

Model	WSQ_ Skytrax	Intercept	R2		WSQ_ Twitter	Intercept	R2
Pooled OLS	2.845***	-1.139***	0.081		3.890***	-1.271***	0.132
Between	12.779**	-1.725***	0.736		8.717*	-1.643***	0.616
Fixed Effects	0.643		0.006		1.002*		0.009
Random Effects	0.813*	-1.019***	0.009		1.119**	-1.063***	0.013

6. Conclusion

This study compares the differences of UGC derived metrics between the non-social media and social media platforms and examines whether UGC derived metrics from these two types of platforms can be indicators for operational metrics.

Extracted reviews for 10 US airlines from Skytrax and Twitter present an overall stable topic distribution from January 2014 to June 2019. While there is a significant difference between the two platforms, Skytrax dominants in reviews of boarding and baggage and Twitter tweets mainly focus on customer service. The study extends previous findings on Skytrax reviews in [23] that most passenger opinions concern two critical services: the check-in and the baggage claim. On the other hand, Twitter, as a social media platform, allows for interactive exchanges and could serve as a direct communication means between companies and customers [27].

Moreover, the two platforms show a significant difference in WSQ scores, and Skytrax has a slightly lower WSQ score in average comparing with Twitter. Also, both WSQ scores in Skytrax and Twitter positively relates with AQR, and the WSQ score of Twitter is a significant predictor for AQR. This finding complements the results from [14], in which the average sentiment score of tweets for US airlines was found to be significantly positively related with AQR.

7. Discussion

The paper makes several important contributions to the literature of employing online reviews to assess service quality. First, instead of using an overall sentiment score as a measure of service quality, we argue that the importance of topics matters and therefore proposed the weighted sentiment under the framework of importance-performance analysis. Second, we find that the reviews on social media platforms differ from the reviews on non-social media platforms, which could relate with the variation of functionalities and textual features in platforms. Finally, we cross validate the online review derived metrics with the results of industry standard AQR and found that metrics from social media could serve as a better indicator with significance. Researchers should be cautious about the review channels when using reviews as a quality measure.

8. Reference

[1] P.-T. Chen and H.-H. 'Sunny' Hu, "The mediating role of relational benefit between service quality and customer loyalty in airline industry," *Total Quality Management & Business Excellence*, vol. 24, no. 9–10, pp. 1084–1095, 2013.

[2] P. L. Ostrowski, T. V O'Brien, and G. L. Gordon, "Service quality and customer loyalty in the commercial airline industry," *Journal of travel research*, vol. 32, no. 2, pp. 16–24, 1993.

[3] S. Aksoy, E. Atilgan, and S. Akinci, "Airline services marketing by domestic and foreign firms: differences from the customers' viewpoint," *Journal of Air Transport Management*, vol. 9, no. 6, pp. 343–351, 2003.

[4] A. Brochado, P. Rita, C. Oliveira, and F. Oliveira, "Airline passengers' perceptions of service quality: Themes in online reviews," *International Journal of Contemporary Hospitality Management*, 2019.

[5] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "A conceptual model of service quality and its implications for future research," *Journal of marketing*, vol. 49, no. 4, pp. 41–50, 1985.

[6] A. Parasuraman, V. A. Zeithaml, and L. Berry, "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality," *1988*, vol. 64, no. 1, pp. 12–40, 1988.

[7] H. Nadiri, K. Hussain, E. H. Ekiz, and Ş. Erdoğan, "An investigation on the factors influencing passengers' loyalty in the North Cyprus national airline," *The TQM Journal*, 2008.

[8] C. R. Kothari, *Research methodology: Methods and techniques*. New Age International, 2004.

[9] J. Kotrlik and C. Higgins, "Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research," *Information technology, learning, and performance journal*, vol. 19, no. 1, p. 43, 2001.

[10] T. Y. Lee and E. T. Bradlow, "Automated marketing research using online customer reviews," *Journal of Marketing Research*, vol. 48, no. 5, pp. 881–894, 2011.

[11] N. Schwarz, "Self-reports: how the questions shape the answers.," *American psychologist*, vol. 54, no. 2, p. 93, 1999.

[12] "Airline Quality Rating 2020," 2020.

[13] D. L. Rhoades, "Airline service quality: Exploratory analysis of consumer perceptions and operational performance in the US and EU".

[14] X. Tian, W. He, C. Tang, L. Li, H. Xu, and D. Selover, "A new approach of social media analytics to predict service quality: evidence from the airline industry," *Journal of Enterprise Information Management*, 2019.

[15] M. Siering, A. V Deokar, and C. Janze, "Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews," *Decision Support Systems*, vol. 107, pp. 52–63, 2018.

[16] P. Xu, L. Chen, and R. Santhanam, "Will video be the next generation of e-commerce product reviews? Presentation format and the role of product type," *Decision Support Systems*, vol. 73, pp. 85–96, 2015.

[17] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert systems with applications*, vol. 36, no. 3, pp. 6527–6535, 2009.

[18] S. Al-Natour and O. Turetken, "A comparative assessment of sentiment analysis and star ratings for consumer reviews," *International Journal of Information Management*, vol. 54, p. 102132, 2020.

[19] N. Kordzadeh, "Investigating bias in the online physician reviews published on healthcare organizations' websites," *Decision Support Systems*, vol. 118, pp. 70–82, 2019.

[20] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7674–7682, 2011.

[21] K. Lee and C. Yu, "Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google

reviews," *Journal of Air Transport Management*, vol. 71, pp. 28–44, 2018.

[22] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 30–44, 2010.

[23] S. Gitto and P. Mancuso, "Improving airport services using sentiment analysis of the websites," *Tourism management perspectives*, vol. 22, pp. 132–136, 2017.

[24] Z. Xiang, Q. Du, Y. Ma, and W. Fan, "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism," *Tourism Management*, vol. 58, pp. 51–65, 2017.

[25] H. Barakat, R. Yeniterzi, and L. Martín-Domingo, "Applying deep learning models to twitter data to detect airport service quality," *Journal of Air Transport Management*, vol. 91, p. 102003, 2021.

[26] L. Martin-Domingo, J. C. Martín, and G. Mandsberg, "Social media as a resource for sentiment analysis of Airport Service Quality (ASQ)," *Journal of Air Transport Management*, vol. 78, pp. 106–115, 2019.

[27] S. Guercini, F. Misopoulos, M. Mitic, A. Kapoulas, and C. Karapiperis, "Uncovering customer service experiences with Twitter: the case of airline industry," *Management Decision*, 2014.

[28] A. Shadiyar, H.-J. Ban, and H.-S. Kim, "Extracting Key Drivers of Air Passenger's Experience and Satisfaction through Online Review Analysis," *Sustainability*, vol. 12, no. 21, p. 9188, 2020.

[29] C. Song, J. Guo, and J. Zhuang, "Analyzing passengers' emotions following flight delays-a 2011–2019 case study on SKYTRAX comments," *Journal of Air Transport Management*, vol. 89, p. 101903, 2020.

[30] N. Korfiatis, P. Stamolampros, P. Kourouthanassis, and V. Sagiadinos, "Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews," *Expert Systems with Applications*, vol. 116, pp. 472–486, 2019.

[31] E. Sezgen, K. J. Mason, and R. Mayer, "Voice of airline passenger: A text mining approach to understand customer satisfaction," *Journal of Air Transport Management*, vol. 77, pp. 65–74, 2019.

[32] Y.-Y. Wang, C. Guo, A. Susarla, and V. Sambamurthy, "Online to offline: the impact of social media on offline sales in the automobile industry," *Information Systems Research*, 2021.

[33] I. Sever, "Importance-performance analysis: A valid management tool?," *Tourism management*, vol. 48, pp. 43–53, 2015.

[34] J. A. Martilla and J. C. James, "Importance-performance analysis," *Journal of marketing*, vol. 41, no. 1, pp. 77–79, 1977.

[35] J.-W. Bi, Y. Liu, Z.-P. Fan, and J. Zhang, "Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews," *Tourism Management*, vol. 70, pp. 460–478, 2019.

[36] R. Mousavi, M. Johar, and V. S. Mookerjee, "The Voice of the Customer: Managing Customer Care in Twitter," *Information Systems Research*, vol. 31, no. 2, pp. 340–360, 2020.

[37] "Airlines and Twitter: The Good, The Bad and the Future | TravelPulse." https://www.travelpulse.com/news/airlines/airlines-and-twitter-the-good-the-bad-and-the-future.html (accessed Jun. 15, 2021).

[38] R. N. Bolton and K. N. Lemon, "A dynamic model of customers' usage of services: Usage as an antecedent and consequence of satisfaction," *Journal of marketing research*, vol. 36, no. 2, pp. 171–186, 1999.

[39] T. S. Dagger and J. C. Sweeney, "Service quality attribute weights: how do novice and longer-term customers construct service quality perceptions?," *Journal of Service Research*, vol. 10, no. 1, pp. 22–42, 2007.

[40] J. A. Chevalier, Y. Dover, and D. Mayzlin, "Channels of Impact: User reviews when quality is dynamic and managers respond," *Marketing Science*, vol. 37, no. 5, pp. 688–709, 2018.

[41] P. Xu and D. Liu, "Product engagement and identity signaling: The role of likes in social commerce for fashion products," *Information & Management*, vol. 56, no. 2, pp. 143–154, 2019.

[42] L. De Vries, S. Gensler, and P. S. H. Leeflang, "Effects of traditional advertising and social messages on brand-building metrics and customer acquisition," *Journal of Marketing*, vol. 81, no. 5, pp. 1–15, 2017.

[43] X. Luo, J. Zhang, and W. Duan, "Social media and firm equity value," *Information Systems Research*, vol. 24, no. 1, pp. 146–163, 2013.

[44] Y.-Y. Wang, T. Wang, and R. Calantone, "The effect of competitive actions and social media perceptions on offline car sales after automobile recalls," *International Journal of Information Management*, vol. 56, p. 102257, 2021.

[45] A. S. Abrahams, W. Fan, G. A. Wang, Z. Zhang, and J. Jiao, "An integrated text analytic framework for product defect discovery," *Production and Operations Management*, vol. 24, no. 6, pp. 975–990, 2015.

[46] "GitHub - taspinar/twitterscraper: Scrape Twitter for Tweets." https://github.com/taspinar/twitterscraper (accessed Jan. 25, 2020).

[47] S. Loria, "textblob Documentation," *Release 0.15*, vol. 2, 2018.

[48] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, pp. 3111–3119, 2013.