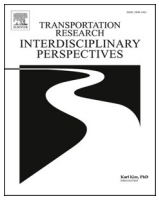
Transportation Research Interdisciplinary Perspectives 24 (2024) 101046

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives



Predicting passengers' feedback rate for airport service quality

Mohammed Saad M. Alanazi *, Karl Jenkins, Jun Li

Centre for Computational Engineering Sciences, School of Aerospace, Transport and Manufacturing (SATM), Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK

ARTICLE INFO

Keywords:
Airport service quality
Passenger review rating
Deep learning

ABSTRACT

Airport service quality evaluation is commonly found on social media sites, including Google Maps. The reviews by users of Google Maps are longer in terms of the number of words than those found on Twitter. They also include a rating, whereas those on Twitter need to be labelled. However, they are less well known than those on Twitter amongst researchers who focus on sentimental analysis. This study attempts to fill the gap in the current literature and develops architecture that is based on Long-Short Term Memory Neural Networks and Convolution Neural Networks. The combined model developed receives meta-data, such as the number of words in the review and the number of likes the review receives in addition to the key review words. The two models, the first of which predicts polarity and the second reviews ratings, were tested under several variations of parameters and showed consistency in results. The dataset was collected from Google Maps and focused on two crowded airports in the Arabic Peninsula (Doha and Dubai). They were found to be unbalanced, with positive reviews being more abundant than negative reviews.

1. Introduction

Airports are the main gateway to nations. Hence, airport service quality is crucial to imparting and sustaining an excellent image of the country in the eyes of passengers arriving via this gateway. Delivering quality services that satisfy passengers gives airports a critical competitive advantage that bolsters economic development (Usman et al., 2022). The bibliometric analysis by Bakır et al. (2022) revealed a tendency to investigate data from social media while studying airport service quality. Social media is a valuable resource when it comes to inspecting people's opinions about service quality (Brochado et al., 2019; Jain et al., 2021; Priyadarshini and Cotton, 2021; Tian et al., 2020; Zhao et al., 2018) and airport service quality in particular (Martin-Domingo et al., 2019). Zhang et al. (2022) highlight the competitive advantage associated with employing artificial intelligence with social media in business domains. Accordingly, there is a tendency to use machine learning and deep learning with sentimental analysis (Jain et al., 2021; Sadiq et al., 2021; Stojanovski et al., 2015; Zhao et al., 2018) when using people's online feedbacks as it is more effective, cheaper and more convenient than surveys (Lee et al., 2021; Lee and Yu, 2018; Li et al., 2022; Sezgen et al., 2019).

This work considers the definition of sentimental analysis by Martin-Domingo et al. (2019), who state that sentimental analysis is a data mining technique that measures a person's emotional polarity (positive and negative) or attitude towards a specific topic. The same definition can be found in Kiliç et al. (2021). This work first reviews the current literature related to airport service quality, particularly studies employing machine learning and deep learning methods, and report any gaps in the literature. The gaps identified are mostly related to datasets and techniques employed to measure the sentimental status of passengers from online reviews. Hence, this work focuses on developing a deep learning model to predict review ratings that measure passenger's sentimental status. A gap in terms of predicting passenger's rating of airport service quality is also identified. Accordingly, this study employs deep learning architectures, particularly related to Natural Language Processing (NLP), in predicting the value (1–5) assigned by passengers to their experience with airport services. This is critical due to the fact that passengers measure their experience quantitatively to deliver their view regarding their satisfaction with airport management.

2. Literature review

In comparison to other types of business, such as online product selling, studies related to airport service quality are limited, as revealed in the work of Bakır et al. (2022). This section reviews only publications from within a five-year timeframe, where there is a higher probability of

* Corresponding author at: 115 Alexandra Road, Birmingham, B5 7NN, UK.
E-mail addresses: m.s.alanazi@cranfield.ac.uk (M.S.M. Alanazi), k.w.jenkins@cranfield.ac.uk (K. Jenkins), jun.li@cranfield.ac.uk (J. Li).

https://doi.org/10.1016/j.trip.2024.101046
Received 9 April 2023; Received in revised form 4 July 2023; Accepted 14 February 2024
Available online 24 February 2024

finding studies investigating recent advancements in Artificial Intelligence (AI), i.e., deep learning. In order to obtain a comprehensive view of the current status of research related to airport service quality, a systematic literature review of relevant studies is appropriate. A recent systematic review revealed a tendency among researchers to employ machine learning and deep learning in the evaluation of airport service quality (Da Rocha et al., 2022). However, different approaches have been found. In order to validate their method, Lee and Yu (2018) compared their deep-learning based model with the ASQ ratings of Airport Council International, and they found many similarities. A similar approach was adopted by Tian et al. (2020), who used the dimensions of SERVQUAL, a well-known tool, to measure people's satisfaction with service quality. They used the number of dimensions of SERVQUAL to predict the possible number of topics that the passengers giving their opinions online would refer to as well as latent Dirichlet allocation (LDA) or Latent Semantic Analysis (LSA) and clustering.

On the other hand, it can be seen that studies that used surveys were the reference for studies that used topic modelling; the list of factors (topics) in both types of studies is quite similar. For instance, Allen et al. (2020) empirically tested a list of factors that were either related to access (signboard, security staff) or control (waiting-time to check-in, staff). Similar topics were investigated by Da Rocha et al. (2022) and Saut and Song (2022). Table 1 summarizes some studies that used sentimental analysis and topic modelling.

In terms of datasets, Lee and Yu (2018) and Li et al. (2022) selected Google Maps (GMap) as the source of passengers' viewpoints on airport service quality. Lee and Yu (2018) defend this decision by stating that GMap allows passengers to rate airport service quality in addition to giving feedback. They argue that this is more reliable because passengers themselves rate their sentimental analysis (on a scale of 1–5) instead of other people classifying reviews into positive or negative without specific numerical values. Additionally, GMap provides more space for feedback in contrast to Twitter, which imposes a very limited word count. Finally, Google Map was seen to be trustworthy (Li et al., 2022). On the other hand, Bunchongchit and Wattanacharoensil (2021) used Skytrax data (passenger reviews of 1st-100th airports) in order to build passenger segmentation. Similarly, Gitto and Mancuso (2017) used data collected from Skytrax. Finally, Stojanovski et al. (2015) used data collected from TripAdvisor.

Regarding sentimental analysis, Lee et al. (2021) extracted tweets from Twitter, whereas Barakat et al. (2021) used tweets related to airports that were already-labelled via automated labelling using a crowdsourcing tool (English and Arabic). It was noticed, as Almuqren et al. (2021) report, that there are many pre-processing steps to clean datasets extracted from Twitter, including the removal of emoticons. Almuqren et al.,'s (2021,p. 11) justification for removing emoticons was "we noticed that the classifier misunderstanding between the parentheses in the quote and in the emoticon". However, this fails to take into account (1) that tweets have a limited size and thus people may use emoticons rather than words to show their emotions due to the word restriction, and (2) emoticons are the most explicit symbols of sentimental status and may help any classifier to catch patterns inside tweets. Moreover, pre-processing could overcome the issue by replacing these emoticons with specific words.

Table 1
The methods and techniques to develop sentimental analysis model.

Author(s)	Method (sentimental/topic modelling)	Technique
(Gitto and Mancuso, 2017)	Semantria for sentimental	Sentimental analysis
(Lee and Yu, 2018)	AFINN sentiment lexicon/LDA	Sentimental analysis/topic modelling
(Tian et al., 2020)	(Number of POS words -number of NEG words) ÷ number of words in the tweet/ clustering, word frequency	Sentimental analysis/topic modelling
(Lee et al., 2021)	LDA	Topic modelling
(Bunchongchit and Wattanacharoensil, 2021)	Google sentimental analysis API	Sentimental analysis

Regarding algorithms and methods used, a simple look at Table 2 reveals that the Latent Dirichlet Allocation (LDA) is commonly used with people's online reviews for service quality.

While inspecting research that utilized deep learning and machine learning, this study found many examples of work that considered both ground airport services and airline services. Accordingly, the focus of this study is on reviewing papers that focus solely on ground airport services. This concise approach should assist airport management in optimizing their services.

In terms of comparing the results retrieved from social media and the ones extracted from a survey, social media is a prime source of people's opinions/satisfaction/sentiments, and surveys can only compete by using structured information. The major works reviewed in this study (related to sentimental analysis and airport service quality) show that social media has become a very good resource for the gathering of passengers' opinions on services provided, including airport services. However, social media data needs cleaning and in many cases labelling. Cleaning is the first step taken by almost all researchers. Labelling is required with data collected from Twitter and other sources, but this is not the case with data collected from GMap.

Lee and Yu (2018) provide a good argument regarding the preference for GMap rather than other sources, as follows. First, though tweets are often used in major research, their limited length reduces the quality of topic modelling techniques, such as LDA. Moreover, Zhu et al. (2021) found comment length, among many features, can potentially predict the critical nature of future comments. In addition, GMap compels the use of sentence-level sentimental analysis (Martin-Domingo et al., 2019). Second, TripAdvisor, an alternative source of feedback, is more related to airlines, hotels, restaurants and local attractions than to airport service quality. Lee and Yu's (2018) claim is supported by the results of Sezgen et al. (2019), who showed that reviews extracted from TripAdvisor, using 'airport' as one of the search keywords, mostly described airlines and the hospitality in airplanes. In fact, the word 'airport' is barely to be found in any research related to airport service quality that uses TripAdvisor. Third, SkyTrax, another source of feedback, has limited feedback regarding the majority of airports globally (Bunchongchit and Wattanacharoensil, 2021), and most of the time it is used by airport professional people (not regular passengers) to assess the performance of airports. Additionally, it takes a long time to collect enough reviews from SkyTrax. It took Halpern and Mwesiumo (2021) one year to collect 4188 reviews, whereas Kiliç et al. (2021) collected only 1224 tweets between 2015 and 2020.

Regarding instrument development and validation, there is a tendency to validate machine learning and deep learning models with surveys that were physically conducted on passengers using a well-known survey scale, such as ASQ rating and SERVQUAL, as in Barakat et al. (2021), Lee et al. (2021), Lee and Yu (2018) and Tian et al. (2020). A rising number of studies provide evidence that analysing data available on social media produces similar results to analysing data collected physically. This has consequences in that the former approach saves both time and money. On the other hand, studies that employ topic modelling can learn regarding determining the number of topics from studies that used surveys because employing Structured Equation Modelling with the help of Principle Component Analysis (PCA) can accurately determine the number of topics found in the answers of passengers.

In order to overcome the problem of determining the possible number of services discussed in online passenger reviews, many studies, such as Barakat et al. (2021), Lee et al. (2021), Lee and Yu (2018) and Tian et al. (2020), used the number of dimensions on well-known service quality scales (e.g., ACI-ASQ, SERVQUAL and ASQ rating). Next, they

Table 2
Methods and algorithms used for sentimental analysis and topic model for airport service quality.

Author(s)	Domain/ dataset	Method	Validation	Feasibility	Applicability
(Lee and Yu, 2018)	Airport/ GMap	a. generates sentiment scores b. correlate sentiment scores with Google star ratings	The Deep learning model compared to ASQ ratings AirportCouncil International.	The proposed method generates 25 topics for airport service quality feedback which is good correspondence with ASQ service attributes.	25 topics may not be interesting for all passengers when they review airport services quality.
(Tian et al., 2020)	Airport service quality	a. extract tweets b. clean tweets and label by 3 students c. sentimental analysis d. map tweets to SERVQUAL dimension using word frequency	SERVQUAL	The problem is tweets must be labelled first manually. Nothing mentioned about building model. But they build quality matrix for airport service quality.	Has ability to be applied as they do not state high number of topics.
(Barakat et al., 2021)	Airport service quality	a. process tweets with LSTM and CNN models to predict sentimental values b. tweets clustering (using word frequency)	Topic modelling validated with ACI-ASQ	The 8 categories of airport services can be interacted with by passengers then they assess them.	Deep learning model can be practical and predict sentimental value with high accuracy.

tried to map every retrieved online review onto one of these dimensions: responsiveness, assurance, tangibility, reliability, security, etc. On the other hand, Martin-Domingo et al. (2019) relied on keyword frequencies (a keyword appears at least 10 times) to develop a list of topics, which were eventually found to be equivalent to ACI-ASQ. Similar work was done by Li et al. (2022) and Moro et al. (2020). Meanwhile, other studies employed unsupervised Natural Language Processing (NLP) to predict the number of topics. The popularity of NLP methods, such as LDA and other topic modelling tools, in research concerning people's online reviews of service quality in social media is due to the fact that the data provided in social media is unstructured and no pre-set scale has been used to rate service quality. Therefore, LDA and other topic modelling tools are capable of assisting research on identifying topics found in people's reviews by classifying keywords or terms. This is presented in Fig. 1.

Four tracks can be used in investigating the sentimental analysis of online passenger reviews. The first track is sentimental analysis using commercial tools, such as Semantria, Twinword and Theysay. The second track uses open source and free library, such as AFINN. The third track involves developing a specialized equation, as in Tian et al. (2020). The fourth track involves building deep learning model/machine learning, as in Barakat et al. (2021), Stojanovski et al. (2015) and Zhao et al. (2018). Regarding the fourth track, after analysing the performance of various deep learning architectures, Kamış and Goularas (2019) recommended focusing on the advantages and disadvantages of different deep neural network configurations (including word embedding word2Vec or GloVe) using a single dataset instead of testing across many datasets. However, all datasets tested by Kamış and Goularas (2019) were from Twitter, and under these circumstances, confirmation with other sources, such as GMap, TripAdvisor, etc., may be necessary.

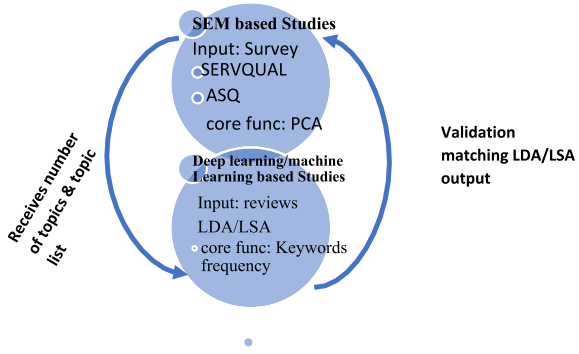


Fig. 1. The relationship between survey-based studies and AI-based studies in terms of airport service quality.

The approaches followed in the various tracks mentioned above are presented in Fig. 2. The first approach was found to run sentimental analysis, then topic modelling. The other approach employed topic modelling, categorizing data based on the number of topics detected, then used sentimental analysis. Actually, Pathak et al. (2021) argue that topic modelling and word frequency contribute to the prediction of sentimental status in the textual data as they are tightly correlated.

2.1. Gaps

First, sentimental analysis provides a polarity value (negative, positive or neutral) as the overall sentimental value for passenger feedback, which may not reflect the actual feelings of the passengers. Review rating is more accurate as the passenger expresses how positive/negative her opinion is. For instance, rate 1 means very negative. Moreover, in many cases, passenger feedback contains several sentimental values (positive, negative or neutral) tagging different airport services. Therefore, tagging the feedback with a single sentimental value may underestimate other important values that could alert airport management to drawbacks in the services provided. Li et al. (2022) found a significant relationship between review rating and some specific airport services mentioned in google map reviews. Accordingly, Aspect-Based Sentimental Analysis (ABSA) could provide more detailed information about the sentimental values that passengers want to deliver.

Second, major studies have used Twitter as a dataset source; however, the sentimental value (positive, negative, neutral) attached to tweets is determined by a human analyst not the passenger. Moreover, the sentimental values of positive/negative/neutral cannot reveal how happy/sad/angry the passenger is. Moro et al. (2020) prefer online reviews that are rated (1–5) because users "devoted attention to quantifying service quality", and this is supported by Li et al. (2022). Feedback provided in GMap with a star rating (set by the passenger) gives a good clue as to how happy/sad/angry the passenger is. Moreover, Zhu et al. (2021) consider the low ratings given by people (<4) as critical because they "are associated with problematic or undelivered services and customer dissatisfaction" (Zhu et al., 2021, p. 865). Therefore, they recommend that the comments associated with a low rating are prioritized in order to increase service quality (Zhu et al., 2021).

Fig. 2. Approaches found with sentimental analysis for airport service quality.

Third, there was little feedback on websites such as Skytrax regarding many airports around the world. For instance, this study concerned airports in the Middle East and found a very small number of reviews regarding the airports in this region. Similarly, the number of reviews collected (all reviews posted from Jan 2011 to Mar 2019) by Shadiyar et al. (2020) from Skytrax about five airlines was only 1693. Therefore, deep-learning-based models, which require very large datasets to deliver high accuracy predictions, cannot effectively benefit from datasets extracted from Skytrax, particularly with less crowded airports.

Fourth, many studies employed sentimental analysis and topic modelling when inspecting online passenger feedback. However, those two approaches were only weakly interrelated empirically. In other words, sentimental analysis was run separately to topic modelling, meaning that the sentimental value generated only tagged the passenger's overall review, with no specific value tagged to any airport service mentioned within the passenger's review. One exception was found in the work of Martin-Domingo et al. (2019), who ran sentimental analysis (using Theysay) on every group to calculate the average number of positives and negatives.

Fifth, many studies were found to have used Twitter as a source of reviews. Tweets are known to have a limited size, and the size of the comments was found to be significant in terms of predicting the sentimental status of the reviewers (Zhu et al., 2021). Datasets collected from sources that provide a larger space for comments may enable more robust prediction for future online reviews.

Sixth, sometimes the number of online reviews is huge, and thus critical reviews can get lost and be missed by businesses (Zhu et al., 2021). Therefore, missed critical reviews could represent a gap in the studies concerning airport service quality. Zhu et al. (2021) found that when people give a low rating (<4), the comments are negative and indicate that the service needs improvement.

To overcome the above-mentioned gaps, this study mainly focused on quantifying passengers' feedback rating with the help of Google Maps as it allows passengers to rate airport services. Predicting a value for passengers' review ratings is crucial as it reveals the level of satisfaction and helps in predicting a review rating for many posts published online on platforms such as Twitter, where no rating is given.

3. Methodology

3.1. Architecture

Balakrishnan et al. (2022) and Puh et al. (2022) conclude that deep learning is better for sentimental analysis within the textual data compared to supervised machine learning. This study followed the recommendations of Barakat et al. (2021) and Kamış and Goularas (2019), using Long-Short Term Memory neural network (LSTM) and Convolutional Neural Network (CNN) with word embedding (GloVe). The input layer consists of users' reviews and meta-data (this time only the number of words and number of likes that the comment received). The selection of word number was based on a recommendation by Zhu et al. (2021). The number of likes the comment received means this comment was approved by many passengers, some of whom may not leave a comment. The same architecture of the network that was used to predict the sentimental value (positive/negative) was used with a different output (review rating 1–5). The two types of input, textual and numerical, are fed into two separate layers of the model. The first layer receives a maximum of 200 words of the feedback to feed the embedding layer, which then feeds the LSTM layer. Similarly, numerical data (number of words in the review, number of likes) are fed into the two shallow layers. Then, the architecture concatenates the numerical output of WHAT? with the output of LSTM in order to feed the final CNN layer before the final output layer.

Another architecture was built, but this time with BiLSTM in addition to CNN. Dropout layers and early stopping were applied in order to overcome overfitting (which was found during evaluation of the proposed architecture). The architecture of (BiLSTM + CNN).

The novelty of such an architecture is that it subjects the number of likes received by passenger's feedback and the number of words in the feedback to an empirical assessment because, as discussed earlier, there was a recommendation to consider doing that depending on the knowledge of authors. There is currently a lack of studies that empirically assess these with airport service quality domain.

3.2. Datasets

Datasets were collected from Google Maps, and the tool provided by outscraper.com was used to collect the data of two famous airports in the Arabic peninsula: Dubai International Airport and Hamad International Airport, Doha. The number of reviews collected related to Doha and Dubai airports were 11,400 and 16170, respectively. No specific dates were set, but most of the reviews were done during the Covid-19 outbreak. The review rate 5 was the highest in both datasets (Doha and Dubai), as seen in Fig. 3. In addition, the data items used were passenger reviews, the number of likes their comments received and the number of words in the review. Other items were removed because they either revealed the personal information of the reviewer (name, image, etc.) or were related to the time and date.

The data was then pre-processed and cleaned. Punctuation, numerical values, special characters and stopwords were eliminated, and the data was lemmatized. Moreover, the column polarity was created based on the following: if the rating is greater than or equal to 3, the polarity is positive; otherwise, it is negative (Ahmed and Ghabayen, 2022).

4. Results and discussion

The results are presented in Table 3. It shows many variations in terms of loss of function and optimizer. Accuracy (training and testing) barely changes when datasets change. It can be seen that accuracy in predicting positivity or negativity is higher than accuracy in predicting review rating.

The high accuracy of LSTM + CNN architecture with polarity (positive and negative) was expected because this was reported in prior studies, such as Barakat et al. (2021). However, this architecture did not deliver high accuracy in terms of the prediction of review rating (~0.75). This also occurred in other domains. As far as the authors are aware, no similar study related to airport service quality predicted review rating. For instance, when consumers' rating of products was evaluated by Liu (2020) using several machine learning methods and Transformer-based models, the accuracy reported was 0.70. Similar results (highest precision was 0.71) were reported by Ahmed and Ghabayen (2022), but with Bi-GRU, Bi-LSTM, MLP, GNB, OneVsRest, K-NN, DT, RF, and LR. BiLSTM was found to give high accuracy of 0.72 with hotels' guests' rating. Finally, Balakrishnan et al. (2022) reported the

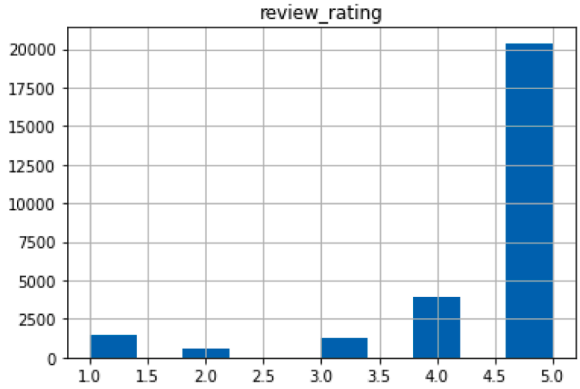


Fig. 3. Number of reviews per every rate (1–5) found in airports of Doha and Dubai.

Table 3
The results of LSTM + CNN architecture with GlOVE embedding.

Database	Training Accuracy	test accuracy	Loss function	Optimizer	Input	Prediction (output)
Dubai	0.93/0.93	0.93/0.93	Softmax/sigmoid	Adam/SGD	Review, No. of words and No. of likes	Positive/negative
Dubai	0.96/0.96	0.95/0.95	Softmax/sigmoid	Adam/SGD	Review	Positive/negative
Dubai	0.72/0.72	0.73/0.73	Softmax/Sigmoid	Adam/SGD	Review, No. of words and No. of likes	Rating (1–5)
Doha	0.93/0.93	0.93/0.92	Softmax/sigmoid	Adam/SGD	Review, No. of words and No. of likes	Positive/negative
Doha	0.97/0.97	0.95/0.95	Softmax/sigmoid	Adam/SGD	Review	Positive/negative
Doha	0.74/0.74	0.72/0.75	Softmax/Sigmoid	Adam/SGD	Review, No. of words and No. of likes	Rating (1–5)
Doha + Dubai	0.96/0.96	0.95/0.96	Softmax/sigmoid	Adam/SGD	Review	Positive/negative
Doha + Dubai	0.73/0.73	0.73/0.73	Softmax/sigmoid	Adam/SGD	Review	Rating (1–5)
Doha + Dubai	0.73/0.74	0.75/0.75	Softmax/sigmoid	Adam/SGD	Review, No. of words and No. of likes	Rating (1–5)
Doha + Dubai	0.91/0.90	0.91/0.90	Softmax/sigmoid	Adam/SGD	Review, No. of words and No. of likes	Positive/negative

results of many deep learning models in the prediction of consumer's 5-scale-rating with original datasets (other results after augmenting datasets). The highest accuracy was 0.63 when deep learning was used with Word2Vec, and 0.70 with RoBERTa models. Yet, this study recorded higher accuracy with a rating of (~0.75) when combined architecture (LSTM and CNN) was used.

There appears to be a rational explanation for this, which is that the reviews associated with 5,4 and 3 ratings probably used quite similar words to praise airports. Consequently, word embedding matrices led to similar results for many reviews, making it difficult to differentiate between ratings 5, 4 and 3. The developed architecture could achieve higher accuracy if the dataset contained more variation in terms of review rating, i.e., if more negative ratings existed. Finally, this study can confirm the work of Lee and Yu (2018) and Li et al. (2022) in terms of the feasibility of collecting data from Google Maps. On the other hand, including meta data in predictions was found to be insignificant in this study. This was due to the limited number of reviews in data collected from GMap in contrast to other resources. Accordingly, the findings of this study are not in line with the recommendation of Zhu et al. (2021) because the number of words and likes did not contribute to increasing the accuracy of predictions in all variations of the proposed architecture.

The architecture of BiLSTM with CNN was found to have higher accuracy when compared to LSTM + CNN. The accuracy while training was 0.84 and during testing was 0.75. However, regardless of adding more data (3000 records from GMap about Heathrow airport) and many hyperparameters tunings taken place, overfitting still exists.

In order to better understand the reasons behind review ratings set by passengers, some statistical analyses took place. One of those statistical analyses was correlation of the number of words, number of likes and review ratings. Pearson correlation and 2-tailed were used with the three datasets (Heathrow, Doha and Dubai). Tables 4–6 and Figs. 4-9 show moderate correlation. There is a negative relationship between the number of words and review rating, and the number of likes and words are always greater when the rating is negative (i.e., 1 or 2).

In an attempt to identify the reason for the low accuracies reported in this study, several investigations were done. Analysing the wrong predictions by the architectures above reveals that 71 % of wrong predictions are due to the existence of positive ratings. The positive ratings 3, 4 and 5 were behind the wrong prediction; i.e., when the rating set by passenger was 5, the prediction is 4 or 3 and vice versa. This is possibly due to the quite similar words used by passengers to describe their experience.

Table 4
Correlations in Heathrow dataset (N = 2390).

		review_rating	wordcount
Wordcount	Pearson Correlation	−0.216**	
	Sig. (2-tailed)	<0.001	
Review likes	Pearson Correlation	−0.069**	0.183**
	Sig. (2-tailed)	<0.001	<0.001

Table 5
Correlations in Dubai dataset (N = 16170).

		review_rating	wordcount
wordcount	Pearson Correlation	−0.221**	
	Sig. (2-tailed)	<0.001	
review_likes	Pearson Correlation	−0.067**	0.263**
	Sig. (2-tailed)	<0.001	<0.001

Table 6
Correlations in Doha dataset (N = 11381).

		review_rating	wordcount
wordcount	Pearson Correlation	−0.268**	
	Sig. (2-tailed)	<0.001	
review_likes	Pearson Correlation	−0.026**	0.113**
	Sig. (2-tailed)	0.005	<0.001

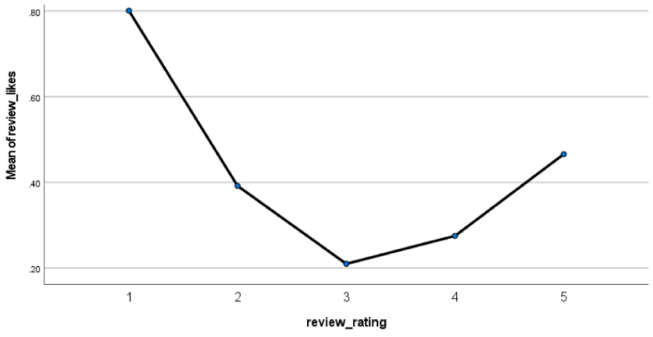


Fig. 4. The tendency to give like to specific rating (Heathrow dataset).

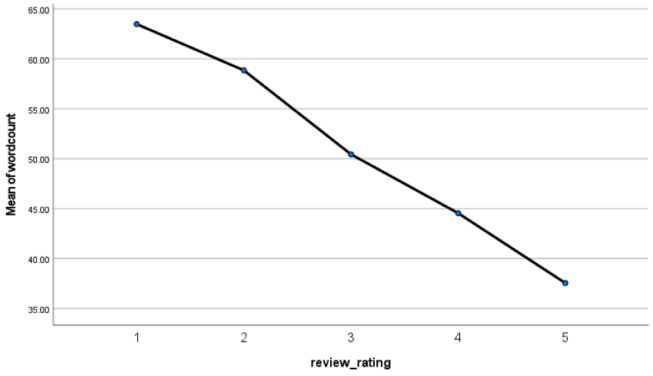


Fig. 5. The average number of words in the review per review rating (Heathrow dataset).

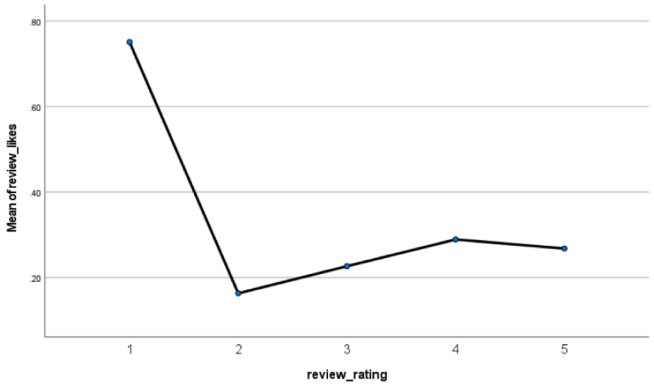


Fig. 6. The tendency to give like to specific rating (Dubai dataset).

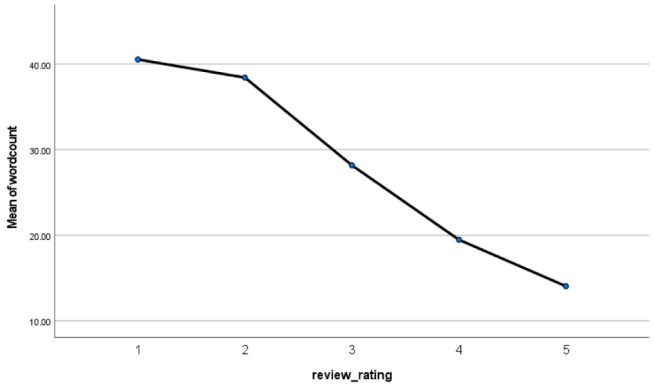


Fig. 7. The average number of words in the review per review rating (Dubai dataset).

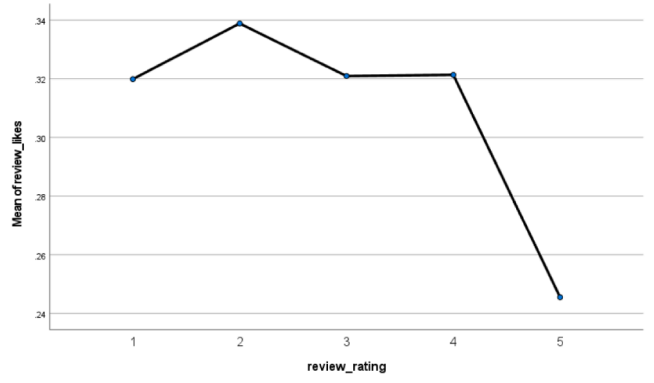


Fig. 8. The tendency to give like to specific rating (Doha dataset).

5. Conclusion and future work

Passengers' reviews on many social media platforms are not tagged by service rating, which makes it difficult to know precisely the level of satisfaction with airport services. In order to predict passengers' review rating for such data, there was a need to use a platform providing a review rating. This study used data collected from Google Maps in developing a prediction model. Focusing on this data, the study found that there is significant value in considering passenger reviews posted on Google Maps. LSTM + CNN architecture was used to develop a prediction model. This architecture was found to deliver high accuracy in terms of the prediction of polarity (positive/negative) and acceptable accuracy in relation to the prediction of review rating (1–5). This study

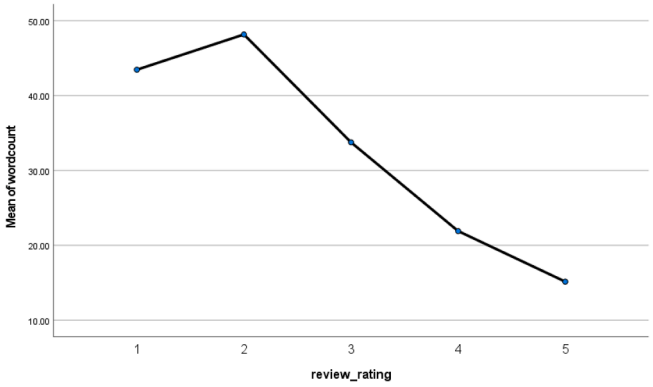


Fig. 9. The average number of words in the review per review rating (Doha dataset).

worked on datasets of passenger reviews for Doha and Dubai airports. These datasets were large but unbalanced, with positive reviews being more abundant than negative reviews. This study considered two metadata features, i.e., the number of words in the review and the number of likes, but these were found to be insignificant. Moreover, future work will include aspect-based sentimental analysis in order to identify the airport services that are explicitly described negatively in passenger reviews. Future work will also focus on balancing the dataset. Aspect-based sentimental analysis could reveal more information than topic modelling; however, it needs datasets to be labelled, which is not the case with topic modelling. Regarding limitations, this study used data related to three airports only, and therefore, the results cannot be generalized with confidence. Another limitation is that Google Maps does not provide detailed data regarding the reviewers' characteristics, which could potentially reveal information that would impact the findings of this study.

Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Mohammed Saad M. Alanazi: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Software, Visualization, Writing – original draft. **Karl Jenkins:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Supervision, Validation. **Jun Li:** Conceptualization, Investigation, Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ahmed, B.H., Ghabayen, A.S., 2022. Review rating prediction framework using deep learning. J. Ambient Intell. Human. Comput. 13, 3423–3432. https://doi.org/10.1007/s12652-020-01807-4.

Allen, J., Bellizzi, M., Grazia, E.L., Forciniti, C., Mazzulla, G., 2020. Latent factors on the assessment of service quality in an Italian peripheral airport. Transp. Res. Procedia 47, 91–98. https://doi.org/10.1016/j.trpro.2020.03.083.

Almuqren, L., Alrayes, F.S., Cristea, A.I., 2021. An empirical study on customer churn behaviours prediction using Arabic twitter mining approach. Future Internet 13 (175). https://doi.org/10.3390/fi13070175.

Bakır, M., Ozdemir, E., Akan, S.A., Atalık, O., 2022. A bibliometric analysis of airport service quality. J. Air Transp. Manag. 104. https://doi.org/10.1016/j.jairtraman.2022.102273.

Balakrishnan, V., Shi, Z., Law, C., Lim, R., Teh, L., Fan, Y., 2022. A deep learning approach in predicting products' sentiment ratings: a comparative analysis. J. Supercomput. 78 (5), 7206–7226. https://doi.org/10.1007/s11227-021-04169-6.

Barakat, H., Yeniterzi, R., Martín-Domingo, L., 2021. Applying deep learning models to twitter data to detect airport service quality. J. Air Transp. Manag. 91, 102003. https://doi.org/10.1016/j.jairtraman.2020.102003.

Brochado, A., Rita, P., Oliveira, C., Oliveira, F., 2019. Airline passengers' perceptions of service quality: themes in online reviews. Int. J. Contemp. Hosp. Manag. 31 (2), 855–873. https://doi.org/10.1108/IJCHM-09-2017-0572.

Bunchongchit, K., Wattanacharoensil, W., 2021. Data analytics of Skytrax's airport review and ratings: views of airport quality by passengers types. Res. Transp. Bus. Manag. https://doi.org/10.1016/j.rtbm.2021.100688.

Da Rocha, P., Medeiros, Costa, H., Gomes, da Silva, G., Barbosa, 2022. Gaps, trends and challenges in assessing quality of service at airport terminals: a systematic review and bibliometric analysis. Sustainability https://doi.org/https://doi.org/10.3390/su14073796.

Gitto, S., Mancuso, P., 2017. Improving airport services using sentiment analysis of the websites. Tour. Manage. Perspect 22, 132–136. https://doi.org/10.1016/j.tmp.2017.03.008.

Halpern, N., Mwesiumo, D., 2021. Airport service quality and passenger satisfaction: the impact of service failure on the likelihood of promoting an airport online. Res. Transp. Bus. Manag. https://doi.org/10.1016/j.rtbm.2021.100667.

Jain, P., Kumar, P.R., Srivastava, G., 2021. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Computer Science Review 41, 100413. https://doi.org/10.1016/j.cosrev.2021.100413.

Kamış, S., Goularas, D. (2019). Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML),.

Kiliç, S., Çadirci, T., Ozansoy, 2021. An evaluation of airport service experience: an identification of service improvement opportunities based on topic modeling and sentiment analysis. Res. Transp. Bus. Manag. https://doi.org/10.1016/j.rtbm.2021.100744.

Lee, H., Joo, L.M., Lee, H., Cruz Angelie, R., 2021. Mining service quality feedback from social media: a computational analytics method. Gov. Inf. Q. 38, 101571. https://doi.org/10.1016/j.giq.2021.101571.

Lee, K., Yu, C., 2018. Assessment of airport service quality: a complementary approach to measure perceived service quality based on Google reviews. J. Air Transp. Manag. 71, 28–44. https://doi.org/10.1016/j.jairtraman.2018.05.004.

Li, L., Mao, Y., Wang, Y., Ma, Z., 2022. How has airport service quality changed in the context of COVID-19: a data-driven crowdsourcing approach based on sentiment analysis. J. Air Transp. Manage. 105. https://doi.org/10.1016/j.jairtraman.2022.102298.

Liu, Z., 2020. Yelp Review Rating Prediction: Machine Learning and Deep Learning Models. arXiv. https://doi.org/10.48550/arXiv.2012.06690.

Martin-Domingo, L., Martín, J., Carlos, M.G., 2019. Social media as a resource for sentiment analysis of airport service quality (ASQ). J. Air Transp. Manag. 78, 106–115. https://doi.org/10.1016/j.jairtraman.2019.01.004.

Moro, S., Lopes, R.J., Esmerado, J., Botelho, M., 2020. Service quality in airport hotel chains through the lens of online reviewers. J. Retail. Consum. Serv. 56, 102193. https://doi.org/10.1016/j.jretconser.2020.102193.

Pathak, A., Ram, P.M., Rautaray, S., 2021. Topic-level sentiment analysis of social media data using deep learning. Appl. Soft Comput. 108 (2021). https://doi.org/10.1016/j.asoc.2021.107440.

Priyadarshini, I., Cotton, C., 2021. A novel LSTM–CNN–GRID search-based deep neural network for sentiment analysis. J. Supercomput. 77, 13911–13932. https://doi.org/10.1007/s11227-021-03838-w.

Puh, K., Babac, M., Bagic', 2022. Predicting sentiment and rating of tourist reviews using machine learning. J. Hospit. Tour. Insights. https://doi.org/10.1108/JHTI-02-2022-0078.

Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., Nappi, M., 2021. Discrepancy detection between actual user reviews and numeric ratings of Google App store using deep learning. Expert Syst. Appl. 181 (2021). https://doi.org/10.1016/j.eswa.2021.115111.

Saut, M., Song, V., 2022. Influences of airport service quality, satisfaction, and image on behavioral intention towards destination visit. Urban, Plann. Transp. Res. 10 (1), 82–109. https://doi.org/10.1080/21650020.2022.2054857.

Sezgen, E., Mason, K.J., Mayer, R., 2019. Voice of airline passenger: a text mining approach to understand customer satisfaction. J. Air Transp. Manag. 77, 65–74. https://doi.org/10.1016/j.jairtraman.2019.04.001.

Shadiyar, A., Ban, H.-J., Kim, H.-S., 2020. Extracting key drivers of air passenger's experience and satisfaction through online review analysis. Sustainability. https://doi.org/10.3390/su12219188.

Stojanovski, D., Strezoski, G., Madjarov, G., Dimitrovski, I., 2015. Twitter sentiment analysis using deep convolutional neural network HAIS 2015. Bilbao, Spain.

Tian, X., He, W., Tang, C., Li, L.C., Xu, H., Selover, D., 2020. A new approach of social media analytics to predict service quality: evidence from the airline industry. J. Enterp. Inf. Manag. 33 (1), 51–70. https://doi.org/10.1108/JEIM-03-2019-0086.

Usman, A., Azis, Y., Harsanto, B., Azis, A., Mulyono. (2022). Airport service quality dimension and measurement: a systematic literature review and future research agenda. Int. J. Qual. Reliab. Manage., 39(10), 2302-2322. https://doi.org/10.1108/IJQRM-07-2021-0198.

Zhang, H., Zang, Z., Zhu, H., Uddin, I., Amin, A., 2022. Big data-assisted social media analytics for business model for business decision making system competitive analysis. Inf. Process. Manag. 2022 (59), 102762. https://doi.org/10.1016/j.ipm.2021.102762.

Zhao, J., Gui, X., Zhang, X., 2018. Deep convolution neural networks for twitter sentiment analysis. IEEE Access 6, 23253–23260.

Zhu, J., Jianjun, Chang, Y.-C., Ku, C.-H., Li, S., Yiyan, Chen, C.-J., 2021. Online critical review classification in response strategy and service provider rating: algorithms from heuristic processing, sentiment analysis to deep learning. J. Bus. Res. 129, 860–877. https://doi.org/10.1016/j.jbusres.2020.11.007.

Predicting passengers' feedback rate for airport service quality

Alanazi, Mohammed Saad M.

2024-03-01

Alanazi MS, Jenkins K, Li J. (2024) Predicting passengers' feedback rate for airport service quality. Transportation Research Interdisciplinary Perspectives, Volume 24, March 2024, Article number 101046

https://doi.org/10.1016/j.trip.2024.101046