Analysing the Sentiment of Air-Traveller: A Comparative Analysis

Mohammed Salih Homaid, Desmond Bala Bisandu, Irene Moulitsas, and Karl Jenkins

***Abstract*—Airport service quality is considered to be an indicator of passenger satisfaction. However, assessing this by conventional methods requires continuous observation and monitoring. Therefore, during the past few years, the use of machine learning techniques for this purpose has attracted considerable attention for analysing the sentiment of the air traveller. A sentiment analysis system for textual data analytics leverages the natural language processing and machine learning techniques in order to determine whether a piece of writing is positive, negative or neutral. Numerous methods exist for estimating sentiments which include lexical-based methodologies and directed artificial intelligence strategies. Despite the wide use and ubiquity of certain strategies, it remains unclear which is the best strategy for recognising the intensity of the sentiments of a message. It is necessary to compare these techniques in order to understand their advantages, disadvantages and limitations. In this paper, we compared the Valence Aware Dictionary and sentiment Reasoner, a sentiment analysis technique specifically attuned and well known for performing good on social media data, with the conventional machine learning techniques of handling the textual data by converting it into numerical form. We used the review data obtained from the SKYTRAX website for each airport. The machine learning algorithms evaluated in this paper are VADER sentiment and logistic regression. The term frequency-inverse document frequency is used in order to convert the textual review data into the resulting numerical columns. This was formulated as a classification problem, whereby the prediction of the algorithm was compared with the actual recommendation of the passenger in the dataset. The results were analysed according to the accuracy, precision, recall and F1-score. From the analysis of the results, we observed that logistic regression outperformed the VADER sentiment analysis.**

***Index Terms*—Airport service quality, data analytics, machine learning, sentiment analysis, text mining, regression.**

I. Introduction

Air travel is now one of the most common and popular ways of travelling around the world. According to the International Civil Aviation Organisation, the statistic number of air travellers had reached 4.2 billion in 2018 [1]. This proves that most air travel is the preference of most people. Many travellers have encountered various challenges and problems during their journeys, whereas digital solutions play a major role in solving them. As reported by [2], airport investment increased by 40% in 2020 in order to improve the capacity and operations of airports, and most importantly, to give an enhanced passenger experience. According to [3], passenger satisfaction is one of the three key drivers of business components that make airports operationally and commercially successful. The others are real-time information distribution and ideal airport experience. An airport is not only a place where aircraft take off and land, but is a complicated place where numerous services are provided for passengers. Therefore, developing an effective tool to measure airport service quality is a salient issue, not only in the literature, but also for practitioners [4]. Airports do not usually satisfy passengers' expectations; therefore, passengers rarely forget if they experience poor service [5]. Consequently, airport operators should invest in enhancing airport services, be they aeronautical or non-aeronautical [6]. Also, to provide a safe environment for passengers. In study [7], they stated the importance of applying security measures and proposed a set of quantitative and qualitative methodologies to assess the risk in the security of air travel sector. Furthermore, airport management should focus on improving operations and businesses because they are competitive environments [2]. Airport service quality (ASQ) is an indicator of passenger satisfaction. The evaluation of ASQ requires continuous observation and monitoring to ensure the provision of high-standard services [8]. Assessing ASQ according to passenger satisfaction provides valuable feedback for airport management. Nowadays, the growth of web-based opinion platforms, such as SKYTRAX [9], Google reviews [10], allows travellers to express their opinion and rate their experience. These platforms attract a high volume of reviews. Therefore, an automated tool based on machine learning is an ideal candidate for understanding these data in order to evaluate ASQ.

In any machine learning model, predictions are made by analysing the historical data trends. The most critical part of any machine learning model is understanding the data. This will contribute to an effective pre-processing step of cleaning the data. We applied the stop-words removal on the textual data as a first step. The second step of the pre-processing was to convert the textual data into a numerical format by using TF-IDF technique [11]. Then, we implemented machine learning algorithms on that processed dataset. Many techniques that use sentiment analysis can be performed, but the key to success is identifying which technique to implement on specific data. In this paper, we implemented two techniques. The first of these is VADER (**V**alence **A**ware **D**ictation s**E**ntiment **R**easoner) Sentiment, which is well established for its performance on social-media data, and the second one is logistic regression [12].

Manuscript received February 2, 2021; revised March 4, 2022. This work was partially funded by Saudi Arabian Cultural Bureau in the United Kingdom.

Mohammed Salih Homaid, Desmond Bala Bisandu, Irene Moulitsas, and Karl Jenkins are with the Department of Computational Engineering Sciences, Cranfield University, Mk43 0AL, United Kingdom.

Mohammed Salih Homaid, Desmond Bala Bisandu and Irene Moulitsas are also with Machine Learning and Data Analytics Laboratory, Digital Aviation Research and Technology Centre (DARTeC), Cranfield University, Mk43 0AL, United Kingdom (e-mail: m.homaid@cranfield.ac.uk, i.moulitsas@cranfield.ac.uk).

DOI: 10.7763/IJCTE.2022.V14.1309

Our study has shown a comparison of the two popular approaches to sentiment analysis. Furthermore, it seeks to assess and to identify which machine learning algorithm should be implemented for such dataset, to create an effective model that will enable organisations to make decisions and business plans. We used the metrics of accuracy, recall, precision and F1-score to measure the performance of our methods. The comparison has shown that the logistic regression outperformed the VADER sentiment.

The rest of this paper is structured as follows. Section II of this paper presents the literature review and Section III introduces the methodology. Section IV presents the experimental results and Section V provides the conclusion.

II. Literature Review

During the past two decades, literature has directed considerable attention to assessing airport and airline service quality, and numerous studies have elucidated such evaluation. They generally applied user surveys and expert opinions in order to obtain and analyse data for the measurement of Airport Service Quality (ASQ). Studies from [8] and [13] used conventional methods such as user survey and experts' opinions in order to evaluate the services provided services in such an airport. A study conducted by [14], proposed an airline quality rating scale to assess U.S. airlines. The airline quality rating scale developed provides a way to compare between the quality of airlines by offering quantitative factors to provide a more reliable and objective result in assessing service quality. The study presented in [8] evaluated the passenger service quality for 14 Asia-Pacific international airports by using a fuzzy multi attributes decision making approach. Six attributes were used to rank the selected airports. This ranking helps airport management to understand the provided services for passengers. [13] evaluated the service quality at Melbourne airport by designing a questionnaire which was inspired by [15], and they used the service quality attributes from the Airports Council International (ACI), the global representative of the airports around the world. The findings indicated that there are remarkable differences between passengers' perception and expectations. Another study conducted by [16], adopted a model for airport satisfaction risk. This model quantifies passenger's experience with eight facets such as security checks, food services, amenities, waiting areas and baggage claim [17]. They also designed an algorithm to calculate the airport satisfaction risk index which added an original concept of quantification to the existing model.

Machine learning techniques, particularly natural language processing in the field of ASQ, have been applied over the past few years. Furthermore, user-generated content has become a common source for measuring user satisfaction in this area. A study conducted by [18] assessed ASQ by applying sentiment analysis on Google maps' reviews using AFINN sentiment lexicon [19]. They evaluated the airport service quality from passengers' perspective and examined how ACI service attributes match the service attributes in Google reviews. They extracted the topics from the textual reviews by using the Latent Dirichlet Allocation (LDA) [20] model and compared it with the ACI service attributes. When compared with the well-measured ASQ ratings conducted by ACI, the results of the study exhibited a high correlation. Twenty-five latent topics were extracted from the reviews, and the correspondence between these and the ACI service attributes was highly accurate. Also, the result indicated that Google star ratings and sentiment analysis are good predictors of ASQ ratings. The study also indicated that not all the services attributes are equally important for different airports size. For example, transportation to and from the airport, cleanness of airport, are more important for small airports. Whereas, customs inceptions, nice ambiance are more important in large airports. That means there are different priority for each airport to improve the quality of passengers' services.

Another study [21] investigated the customer satisfaction level of airport services by using sentiment analysis. Passengers' reviews were extracted and analysed from the SKYTRAX website. Only five international airports were considered in this study, and the data were collected from the website for the period from September 2013 to February 2014. Services have been divided into two groups, aviation and non-aviation services. Two open-source software have been employed to identify passenger' perceptions of the services provided in airports, namely KNIME and Semantria. KNIME is a software that is designed to analayse web forums and social media, whereas Semantria used an automated sentiment analysis.

Another study that utilised social media as a data source for evaluating airport service quality was presented in [22]. They assessed the services at Heathrow Airport by analysing sentiment from the Heathrow Twitter account where passengers may write comments. They used the "Theysay" tool, which was developed by computational linguists at Oxford University. 23 attributes have been extracted and compared with other ASQ scales.

Two studies from [3] and [23] measured the airport service quality by identifying and analysing the key drivers for passenger satisfaction and dissatisfaction. Herzberg's two-factor motivation theory was applied in the former study, and a new model based on Latent Dirichlet allocation (LDA) was proposed in the latter.

III. Methodology

A. The Employed Methods

In this section, we present a brief review of the methodology used to perform our experiments and extract the results. Our study proposed a novel approach to extract results from a reviews' dataset. We utilised both textual and numerical columns as inputs to the machine learning models, and compared two existing techniques in order to extract the results and to compare them. In this paper, we compared VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner) with our approach, which is essentially coupling textual data with numerical data in conventional machine learning algorithms. We employed logistic regression as part of our classification machine learning algorithm. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media; such as Twitter and google reviews where there are only limited number of characters. however, it also works well on texts from other domains [24]. Sentiment analysis is an index investigation technique that recognises extremity within the content, irrespective of whether or not this is an entire report, passage, sentence, or condition [25]. Sentiment analysis quantifies the disposition, notions, assessments, perspectives and feelings of a speaker/author according to the computational treatment of subjectivity in a content. VADER sentiment analysis depends on a word reference that maps lexical highlights to feeling intensities, which is known as estimation scores. The overall score of a book can be calculated by summarising the power of each word in its content. The resulting score for each text may be obtained by calculating a weighted sum of the score of each word. For instance, words such as "love", "appreciate", "glad" and "like" all pass on a positive supposition. Additionally, VADER is sufficiently sophisticated to understand the fundamental setting of these words; for example, "didn't cherish" as a negative assertion. Recently, we have observed that opinion-based postings on social media are reshaping many key aspects of life, such as business and public sentiments. Lexicon mapping-based approaches, such as VADER, are proving to be particularly effective. In this paper, we present the airport reviews' dataset which has multiple columns, both textual and numerical. For VADER sentiment analysis, only the textual data was used to extract the sentiment score from that review data, whereas for the logistic regression we used the TF-IDF technique to convert the textual data into numerical data to be applied in logistic regression. In the logistic regression, can have a quantitative approach to the content of the dataset and we can take advantages of all the numerical columns within it. In this particular dataset, we have applied the chosen methods to see which one works well to predict passenger's satisfaction.

B. Pre-processing the Dataset

Initially, the dataset was pre-processed, a procedure that involved the removal of stopwords by using the NLTK (**N**atural **L**anguage **T**ool**K**it) stopwords for the English language. In addition to previous pre-processing, a further step was required for the use of logistic regression. Logistic regression is a machine learning technique that expects the input variables to be in numerical form. Therefore, we made two copies of data, one of which was textual, to be used in its existing form for VADER sentiment analysis, and the other for further processing for logistic regression. In order to convert the textual data into a numerical format, we applied the TF-IDF (term frequency – inverse document frequency) technique. This is a statistical measure which evaluates the relevance of a word to a document in a collection of documents. This is attained by multiplying two metrics: the number of times a word appears in a document, and the inverse document frequency of the word across a set of documents. The resulting data express each word as a single feature with an important numerical value. This produced 39,981 attributes in our dataset which will be subjected to logistic regression for training and testing purposes.

C. Performance Measures

In order to assess the performance of both techniques, we formulated this into a classification problem. Furthermore, for the purpose of evaluating the performance of VADER sentiment and logistic regression, we used the metrics of accuracy, recall, precision and F1-score. The definitions of these metrics are given below:

Accuracy is defined as the closeness of the measurements to a specific value [26].

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (1)$$

Although accuracy provides an insight into the algorithm's overall performance, it may be insufficient to determine an algorithm's performance. Therefore, we also used the notions of recall and precision. Recall attempts to determine the parts of actual positive that were identified correctly [26].

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Recall also gives an estimate of the accuracy of the performance of our model for the positive classes. Precision is also used in an attempt to determine the proportion of positive predictions that were correct [27].

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

In addition to the precision and recall, the F1-score is also applied, which essentially conveys the balance between precision and recall. The F1-score is applied to measure the accuracy of the test [26].

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

IV. Experimental Results and Analysis

We perform our experiment on a Personal Computer (PC) with Intel(R) Core (TM) i7-9700 CPU with a processor speed of 3.00 GHz and 32 GHz RAM. Python Programming Language implemented in Jupyter Notebook is used for these experiments. We utilised libraires such as NLTK, Scikit-Learn for the analysis, and Beautiful Soup library for scraping the reviews.

A. Dataset

In order to extract the sentiment analysis and recommendation of an airport based on the review text, we scraped the reviews for all the airports from the SKYTRAX [8], a dataset that is formatted by the airport name. This dataset comprised a total of 38,584 reviews together with 20 attributes of reviews. It contained 25.19% positive reviews and 74.81% negative, covering the time span from July 2004 to November 2020. Table I presents the actual attributes of our dataset.

TABLE I: Dataset Description

Attribute Name	Data Type	Description
Airport Name	String	Name of the airport to which the review belongs
Reviewer Name	String	Name of the person who provided the review
Review Date	Date	The date on which the review was provided
Reviewer Country	String	Name of the country to which the reviewer belongs

Review Title	String	Title of the review
Review Verified Status	String	Whether or not the review is verified
Review Text	String	The textual review provided by the user
Experience at Airport	String	Transit / arrival and departure / arrival only
Date Visit	Date	The date of the reviewer's visit
Type of Traveller	String	Solo / family etc.
Recommend	String	Contains "yes" or "no" values, whether or not the reviewer recommends the airport.

B. Model Training and Experiments

1) VADER sentiment analysis

After pre-processing the dataset, the next task was to perform sentiment analysis by using VADER sentiments from the review text. VADER sentiment is a mapping technique and does not require training any type of model. Therefore, it was applied to the dataset directly in order to obtain the sentiment scores of all the reviews. VADER sentiment, which is subjected to all the reviews' data, gave a sentiment score of -1 to 1 for each review, with a variable step size, with -1 being the extreme negative review score, and 1 being the extremely positive review score. However, the sentiment scores shown in Fig. 1 do not provide the required output for the target column "recommended" which is either "yes" or "no". Therefore, in order to transform the resulting sentiment scores into a categorical recommended column, we applied a threshold of 0.0. We labelled all the reviews with sentiment scores greater than or equal to 0.0 (>= 0.0) as recommended "yes" and all the remaining sentiment scores as recommended "no". These labelled results will later be analysed for the performance measures.

2) Training logistic regression

Unlike the VADER method, logistic regression requires to be trained. The first step in training the logistic regression was to divide the dataset into two parts: train set and test set. The ratio used for training and testing was 67:33 [11], which translated into 25,530 entries for training and 12,575 entries for testing purposes. The train set will be used to train the logistic regression, and the test set will be used to test the performance of the trained model. After training the logistic regression model, we predicted the sentiments from the testing data. These predictions were later applied in order to detect the model's performance by using the above-defined performance measures in Section IV.

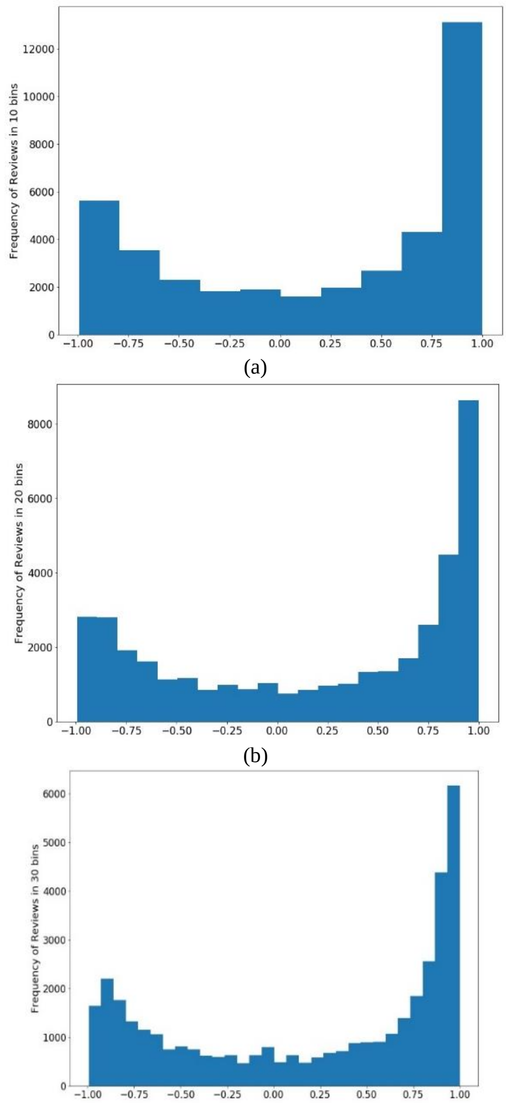
C. Results

In this paper, two techniques were presented regarding the extraction of sentiment analysis of the reviews' data from the SKYTRAX website. The methods chosen to create our models for comparison included VADER which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. However, it also works well on texts from other domains. The second technique used was logistic regression which is a conventional machine learning algorithm. When we tested our models using the reviews' data, the logistic regression proved to be outperforming the VADER sentiment. We used the recommended column in the dataset as a target column to measure the performance of both techniques. The VADER sentiment is already a rule-based analysis technique which requires no training.

To visualise the distribution of the sentiment scores returned by the VADER sentiment, Fig. 1(a), Fig. 1(b) and Fig. 1(c) depict the review frequency in each bin. The height of each bin indicates the number of reviews within it. Fig. 1(a), Fig. 1(b) and Fig. 1(c) illustrate that the VADER sentiment gave most of the reviews a positive score by the height of the bins from 0 to +1.

To ascertain whether or not the distribution of the sentiment scores follows the same trend, we created three different plots with a bin size equal to 10, 20 and 30. We used a variable step size as a parameter in the VADER sentiment represented in the x axis. The step sizes are calculated as *(upper limit of the sentiment – lower limit of the sentiment) / bin size*.

We observed the same trend in the distribution of the scores when the number of bins was increased, thereby leading us to the conclusion that this distribution of scores is stable.



(c)

Fig. 1. (a) VADER scores distribution for 10 bins (b) VADER scores distribution for 30 bins (c) VADER scores distribution for 20 bins.

Having applied the threshold of 0.0, we transformed the sentiment scores into a categorical recommended column. In order to measure the performance of both techniques, we adopted the recommended column in the dataset as a target column. The prediction results for both above-used techniques as judged by the performance measures are shown in Table II.

TABLE II: PERFORMANCE RESULTS

	Accuracy	Precision		Recall		F1-Score	
		Yes	No	Yes	No	Yes	No
VADER Sentiment	59%	0.37	0.94	0.91	0.48	0.53	0.63
Logistic Regression	87%	0.77	0.90	0.71	0.93	0.74	0.92

The data used for the scope of this paper are skewed and contain only 25% positive reviews. However, this skewness of data should not have affected the VADER sentiment as it is a pre-trained lexicon mapping approach, and widely used for sentiment analysis and social media data. On the other hand, the logistic regression requires training of the model which should have created an impact on the model performance as the data for the positive reviews are only 25% owing to the imbalance in the dataset. As for the fact that we have an imbalanced dataset, we cannot exclusively use accuracy as a metric to judge the performance of the model. Therefore, precision and recall also played their part in analysing model performance. It is evident that the logistic regression provided an accuracy of about 87%, whereas the VADER sentiment provided an accuracy of 59%. This simple fact can also be observed in Fig. 1(a), Fig. 1(b) and Fig. 1(c) where most of the reviews received a positive score from the VADER sentiment as indicated by the height of the bar in the positive region.

To gain further insight into the results, we moved towards the remaining performance measures which are precision, recall and F1-score. For this use case, we cannot exclusively rely on the accuracy score because the dataset is highly imbalanced, and accuracy is not the metric by which to judge an imbalanced dataset. The VADER sentiment classified most of the reviews as positive which resulted in a very low precision (37%) for “yes”, whereas logistic regression had a healthy precision (77%) for “yes”. We observe this because VADER sentiment is not trained on this data and work with its lexicon mapping based approach. While logistic regression was trained on this specific data.

The comparison between these two techniques was made to highlight the fact that without proper information and problem understanding, a good algorithm may not perform satisfactorily. In our use case, sentiment analysis is critical because positive reviews will be recommended to potential passengers.

V. CONCLUSION

The internet and social media have become major sources of information which enable the understanding of the sentiment and population trends over time. In 2020, 3.6 billion people were using social media which represents almost half of the planet’s population. Modern economic systems are built on data or knowledge, which will be valuable if they are leveraged to extract further insights. Domestic and international air travel has increased exponentially due to globalisation. In the meantime, if passenger feedback is not collected and utilised effectively, the possibilities of improvement will be minimal. Developing a model to correctly understand the passengers’ opinion, based on the provided services, will help the decision-making operator in airports to enhance the provided services by using alternative approaches. Furthermore, it is crucial to predict passengers’ sentiments correctly in order to provide accurate recommendations that will lead to higher satisfaction. Furthermore, it will be important to identify the major problems encountered by passengers, and attempt to rectify them in order to avoid passenger dissatisfaction. This paper used the data scraped from the SKYTRAX website. The dataset was pre-processed by removing stopwords, and for logistic regression, TF-IDF was also applied. This paper discussed two such approaches of extracting passengers’ sentiments from the review data. While VADER sentiments’ analysis has recently become popular as it requires no training of the algorithm, it is evident from the above results that the VADER sentiment may not be particularly effective when tested on the reviews’ data in comparison with conventional machine learning techniques. As the above results indicate, given the appropriate amount of data and good pre-processing, the conventional machine learning technique outperformed the VADER sentiment regarding accuracy. The results were compared according to accuracy, precision, recall and F1-score. Finally, logistic regression appeared to outperform the VADER sentiment technique for extracting sentiments in this particular study for this specific dataset. The model can be further extended to provide guidance to airport operators.

DATA ACCESSIBILITY

The code used for the simulations is freely available under the MIT license and can be downloaded from the Cranfield University repository: https://doi.org/10.17862/cranfield.rd.19375334.v1

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

MH conceptualised the methodology, undertook the research and prepared the manuscript. DB provided technical advice in the code development and edited the manuscript. IM supervised the research, provided guidance in the preparation of the manuscript and edited the manuscript. KJ co-supervised the research. All authors agree with the final version of the manuscript.

REFERENCES

[1] Air transport, passengers carried. International Civil Aviation Organization, Civil Aviation Statistics of the World and ICAO staff estimates. [Online]. Available: https://data.worldbank.org/indicator/IS.AIR.PSGR?end=2018&start=1970

[2] S. E. Zaharia and C. V. Pietreanu, “Challenges in airport digital transformation,” *Transp. Res. Procedia*, vol. 35, pp. 90–99, 2018.

[3] F. Okumus *et al*., “Airport service quality drivers of passenger satisfaction,” *Tour. Rev.*, 2013.

[4] G. C. L. Bezerra and C. F. Gomes, "Measuring airport service quality: A multidimensional approach," *J. Air Transp. Manag.*, vol. 53, pp. 85–93, 2016.

[5] V. Bogicevic, M. Bujisic, A. Bilgihan, W. Yang, and C. Cobanoglu, "The impact of traveler-focused airport technology on traveler satisfaction," *Technol. Forecast. Soc. Change*, vol. 123, pp. 351–361, 2017.

[6] R. Merkert and A. G. Assaf, "Using DEA models to jointly estimate service quality perception and profitability–evidence from international airports," *Transp. Res. Part Policy Pract.*, vol. 75, pp. 42–50, 2015.

[7] G. Tamasi and M. Demichela, "Risk assessment techniques for civil aviation security," *Reliability Engineering & System Safety*, vol. 96, no. 8, pp. 892-899, 2011.

[8] C.-H. Yeh and Y.-L. Kuo, "Evaluating passenger services of Asia-Pacific international airports," *Transp. Res. Part E. Logist. Transp. Rev.*, vol. 39, no. 1, pp. 35–48, 2003.

[9] A-Z Airport Reviews. SKYTRAX. [Online]. Available: https://www.airlinequality.com/review-pages/a-z-airport-reviews/

[10] Write reviews and add ratings of places - Computer - Google Maps Help. [Online]. Available: https://support.google.com/maps/answer/6230175?co=GENIE.Platform%3DDesktop&hl=en

[11] D. A. Culp, "Using TF-IDF to Determine Word Relevance in Document Queries.," *Urol. Clin. North Am.*, vol. 2, no. 1, pp. 29–48, 1975.

[12] A. Rane and A. Kumar, "Sentiment classification system of Twitter data for US airline service analysis," in *Proc. 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 2018, vol. 1, pp. 769–773.

[13] H. Jiang and Y. Zhang, "An assessment of passenger experience at Melbourne airport," *J. Air Transp. Manag.*, vol. 54, pp. 88–92, 2016.

[14] B. Bowen, D. Headley, and J. Luedtke, "A quantitative methodology for measuring airline quality," J*ournal of Aviation/Aerospace Education & Research*, vol. 2, no. 2, p. 1, 1992.

[15] D. Fodness and B. Murray, "Passengers' expectations of airport service quality," *J. Serv. Mark.*, 2007.

[16] M. Sahinoglu, D. Ang, and S. Morton, *Quantitative Risk Assessment & Management Metrics for Airport Services*, LAP: LAMBERT Academic Publishing, 2016, pp. 1-84.

[17] M. Sahinoglu, D. Ang, S. Morton, and P. Vasudev, "Quantitative metrics for airport services satisfaction risk assessment and management with risk-o-meter software," in *Proc. the 2016 Taiwan International Conference on Innovation and Management (ICIM)*, 2016, pp. 68-75.

[18] K. Lee and C. Yu, "Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews," *J. Air Transp. Manag.*, vol. 71, pp. 28–44, 2018.

[19] F. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," ArXiv Prepr, ArXiv11032903, 2011.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[21] S. Gitto and P. Mancuso, "Improving airport services using sentiment analysis of the websites," *Tour. Manag. Perspect.*, vol. 22, pp. 132–136, 2017.

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

[23] K. Mizufune and S. Katsumata, "Joint classification model of topic and polarity: Finding satisfaction and dissatisfaction factors from airport service review," in *Proc. 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2018, pp. 856–863.

[24] J. Wilson and C. Hernández-Hall, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Proc. Eighth Int AAAI Conf. Weblogs Soc. Media*, 2014, p. 18.

[25] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," in *Proc. the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 190–199.

[26] J. Brownlee, "Machine learning mastery with python understand your data," *Create Accurate Model Sand Work Proj. End--End*, pp. 123–145, 2017.

[27] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Am. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, 1994.

Mohammed Salih Homaid is a PhD candidate at Cranfield University and researcher in the Digital Aviation Research and Technology Center. He is working in the area of machine learning and passenger experience. He holds an MSc in information system from Portsmouth University and a BSc in information system from Al-imam Mohammed bin Saud University.

Desmond Bala Bisandu is a PhD candidate at Cranfield University and researcher in the Digital Aviation Research and Technology Center. He is working in the area of machine learning and passenger experience. He holds an MSc in computer science from American University of Nigeria and BSc in computer science from University of Jos, Nigeria.

Irene Moulitsas holds a PhD in scientific computation from the University of Minnesota. She is currently a senior lecturer in Cranfield University and course director for the MSc in computational software techniques in engineering. Her research has focused on developing novel algorithms for enabling the efficient execution of large scientific computations on parallel processing platforms. She has developed highly efficient serial and parallel algorithms and software, that are publicly available for use by numerous universities, research laboratories, and companies.

Karl Jenkins completed a PhD at the University of Manchester in computational engineering in 1996 leading to the Sir Arthur Marshal Research Fellow at Cambridge University for 7 years. He joined Cranfield University in 2004 and is currently a professor of computational engineering and the head of Centre for Computational Engineering Sciences. He leads a team of 7 academics and is the programme director for 4 MSc courses. His current research is developing high fidelity simulation codes for complex multiscale flows.