

Article

Multiclass Sentiment Prediction of Airport Service Online Reviews Using Aspect-Based Sentimental Analysis and Machine Learning

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

School of Aerospace, Transport and Manufacturing (SATM), Cranfield University, Cranfield MK43 0AL, UK; m.s.alanazi@cranfield.ac.uk (M.S.M.A.); k.w.jenkins@cranfield.ac.uk (K.W.J.)

* Correspondence: jun.li@cranfield.ac.uk

Abstract: Airport service quality ratings found on social media such as Airline Quality and Google Maps offer invaluable insights for airport management to improve their quality of services. However, there is currently a lack of research analysing these reviews by airport services using sentimental analysis approaches. This research applies multiclass models based on Aspect-Based Sentimental Analysis to conduct a comprehensive analysis of travellers' reviews, in which the major airport services are tagged by positive, negative, and non-existent sentiments. Seven airport services commonly utilised in previous studies are also introduced. Subsequently, various Deep Learning architectures and Machine Learning classification algorithms are developed, tested, and compared using data collected from Twitter, Google Maps, and Airline Quality, encompassing travellers' feedback on airport service quality. The results show that the traditional Machine Learning algorithms such as the Random Forest algorithm outperform Deep Learning models in the multiclass prediction of airport service quality using travellers' feedback. The findings of this study offer concrete justifications for utilising multiclass Machine Learning models to understand the travellers' sentiments and therefore identify airport services required for improvement.



Citation: Alanazi, M.S.M.; Li, J.; Jenkins, K.W. Multiclass Sentiment Prediction of Airport Service Online Reviews Using Aspect-Based Sentimental Analysis and Machine Learning. *Mathematics* **2024**, *12*, 781. <https://doi.org/10.3390/math12050781>

Academic Editors: Debo Cheng, Junbo Ma and Rongyao Hu

Received: 7 December 2023

Revised: 26 February 2024

Accepted: 2 March 2024

Published: 6 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: airport service quality; Deep Learning; Twitter; Google Maps; Airline Quality

MSC: 68T07; 68T50

1. Introduction

The term “airport services” encompasses a multifaceted array of offerings that collectively contribute to the overall experience of air travellers. The current literature provides various definitions and perspectives on the constituents of airport services, but often emphasises the comprehensive nature of the travellers' experience within airport facilities. For instance, researchers (e.g., [1]) asserted that airport services encompass a spectrum of amenities and processes, including check-in procedures, security protocols, baggage handling, and terminal facilities. In contrast, other scholars (e.g., [2]), adopted a more nuanced approach to defining airport services. Their focus extended beyond the procedural elements to include the quality of passenger interactions, emphasising customer service, staff responsiveness, and the overall ambiance of the airport environment. This perspective underscores the significance of human-centric factors in shaping travellers' experience. Additionally, Dhini and Kusumaningrum [3] delved into the concept of airport services as a holistic system that encompassed both tangible and intangible elements. Tangible aspects involve physical facilities and infrastructure, while intangible elements include service reliability, efficiency, and responsiveness.

Travellers' reviews in the aviation industry are important. These reviews wield significant influence over travellers' decisions when choosing an airport [1–5]. Even minor improvements in airport services can have a positive impact on travellers' perceptions and

enhance their overall airport experience [6–9]. Additionally, positive sentiments expressed by travellers contribute to the competitive features of airports [9]. As travellers readily access and consult online reviews, airport management must prioritise their airport service quality (ASQ). To identify key areas for enhancing positive reviews, researchers have developed methods for extracting and analysing travellers' reviews [2]. Currently, most studies conduct sentiment extraction, either collectively or at the sentence level. The collective sentiment analysis approach (also called topic modelling) may lack specificity, while the sentence-level approach requires additional human analysis to determine the polarity of a traveller's feedback. Consequently, there is a need for a more comprehensive sentiment analysis approach to identify all aspects of airport services with polarities in a traveller's review.

Moreover, current studies on the sentiment analysis of travellers' feedback have predominantly focused on binary classifications, categorising responses into broad 'positive' or 'negative' sentiments. While this approach provides a high-level view of traveller satisfaction, it significantly narrows the scope of analysis and overlooks the complexity inherent in individual feedback. Another significant gap in the current literature is the lack of depth in understanding the specific services being referenced in feedback. Travellers often comment on various aspects of their airport experience, each potentially carrying a different sentiment. For instance, a single review might express satisfaction with the efficiency of security checks, but dissatisfaction with the amenities in the lounge area. Traditional binary analysis methods fail to capture these nuances, leading to a generalised and potentially misleading interpretation of traveller feedback. It is common that a traveller's review often encompasses a range of sentiments tied to various aspects of their experience, such as check-in services, lounge facilities, or security processes. Current methodologies fall short in dissecting these multifaceted reviews to attribute sentiment to specific airport services. This oversight is critical, as it masks the complexity and diversity of travellers' experiences.

This study aims to address these gaps by introducing a multiclass approach to sentiment analysis. Unlike the binary model, the method seeks to dissect traveller feedback into multiple categories, corresponding to different services or experiences within an airport. This research also introduces and develops unique Deep Learning architectures and Machine Learning classification algorithms for a comprehensive multiclass sentiment analysis of travellers' online reviews. The results are ultimately compared for optimal solutions. In practice, we collect travellers' reviews from multiple social media and employ Aspect-Based Sentiment Analysis (ABSA) to categorise positive/negative/non-existent sentiments by airport services stated in travellers' reviews. The collected dataset is utilised to train and test designed Deep Learning and Machine Learning models, enabling the automation of identifying the sentimental aspects of airport services from travellers' feedback. By doing so, the research seeks to pinpoint specific service areas within airports, thereby offering a more detailed and actionable insight into traveller satisfaction and service quality, enabling stakeholders to identify specific areas of strength and improvement.

The remainder of this paper is organised as follows: Section 2 provides an overview of the current work in the area; Section 3 introduces the methodology, with a particular focus on the dataset used and the architectures of the Deep Learning models designed; Section 4 presents the model results and findings; Section 5 offers the interpretation of findings and their connections with previous studies; and Section 6 concludes the work.

2. Literature Review

Multiclass prediction refers to the process of classifying a given input into one of several categories, which, in the context of this research, pertains to the various aspects of airport services such as check-in, security, and lounges. Unlike binary classification, which limits the output to two categories (i.e., positive and negative), multiclass prediction allows for a more nuanced analysis by categorising feedback into multiple distinct service areas. This approach is particularly pertinent in the context of airport service analysis, where travellers' feedback often spans a wide range of experiences and services.

The methodology for multiclass prediction in the research area is deeply rooted in relevant literature. It builds upon the foundational principles of sentiment analysis, as outlined in seminal works such as [1,10]. Various methodologies such as topic modelling [11,12], Machine Learning [3], and sentimental analysis [10] are proposed to illuminate the aspects of airport services in travellers' feedback, with a prevalent focus on topic modelling and sentiment analysis, particularly those utilising secondary data from platforms such as Twitter, Google Review, Airline Quality, or Skytrax. Conversely, other studies, primarily relying on primary datasets collected directly from travellers, employ statistical analysis to study the impact of airport service quality on travellers' sentiment. Topic modelling is a prevalent analytical technique for scrutinising online reviews by travellers, frequently incorporating methods such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), coupled with dimensionality reduction strategies. Subsequently, sentiment analysis is employed to ascertain the positive or negative nature of the reviews.

Lee and Yu [11] and Martin-Domingo et al. [12] applied topic modelling and sentiment analysis to Google Maps reviews and London Heathrow Airport Twitter datasets, respectively. Moro et al. [13] analysed more than twenty thousand TripAdvisor reviews and presented heatmaps for airport hotel services and the sentimental status of guests. Bae and Chi [14] applied content analysis to differentiate between contented and discontented travellers. Other methods such as frequency analysis, linear regression, multinomial logit, text mining, semantic network analysis, and emotion recognition to identify airport services with positive/negative sentiments based on online reviews such as Skytrax were also utilised [8,15–17]. Nevertheless, relying solely on sentimental classes (i.e., positive, negative, and neutral) proves insufficient for accurately elucidating the specific sentiments of individuals and discerning the precise airport service [18]. For example, Lee and Yu [11] attempted to forecast the star ratings instead of sentimental classes of airport services.

Recently, Machine Learning [3,10], particularly Deep Learning [1,19,20], has gained considerable traction for predicting travellers' sentiment values. Dhini and Kusumaningrum [3] employed Support Vector Machine (SVM) and the Naïve Bayes classifier to discern positive/negative sentiments from Google reviews. Taecharungroj and Mathayomchan [10] demonstrated that the quality of airport services could be assessed through sentimental values linked to various services by leveraging LDA and Naïve Bayes modelling techniques. Kaniş and Goularas [20] explored various Deep Learning architectures with diverse datasets, concluding that the optimal performance was achieved through the combination of the long short-term memory (LSTM) network and convolutional neural network (CNN). Barakat et al. [1] leveraged thousands of English and Arabic tweets to train CNN and LSTM models to categorise travellers' feedback regarding airport services into positive or negative classes based on the US Airline Sentiment and AraSenTi datasets. CNN and LSTM neural networks were demonstrated by Barakat et al. [1] to be effective in handling multi-dimensional data and extracting nuanced patterns. However, despite the models demonstrating superior predictive capabilities, the observed difference was statistically insignificant. Nevertheless, studies focusing on Machine Learning and Deep Learning applications in assessing airport service quality and travellers' sentiment values have been limited.

Moreover, although various studies showed that certain airport services are more likely to garner positive reviews when managed effectively, there is a lack of standardised criteria for enumerating the specific airport services that necessitate attention. For instance, Gajewicz et al. [7] assessed facility attributes like cleanliness and efficiency individually, while others considered these attributes holistically or more broadly, which may encompass amenities such as food and restaurants. Consequently, the absence of uniformity in categorising airport services results in varied lists across studies [21–26]. To address this, Alaydaa et al. [27] presented a two-level category of airport services covering all explicit facilities within the airport based on the conducted review, including access, check-in and security, facilities, wayfinding, and airport environment and their sub-categories, which may well suit the Aspect-Based Sentiment Analysis of travellers' online reviews.

Overall, previous studies typically predict a single sentiment polarity (i.e., negative, positive, or neutral) for each traveller's review, which may overlook important information for improving airport services, as traveller reviews often encompass comments on multiple aspects of airport services. Aspect-Based Sentiment Analysis, which examines traveller reviews according to airport services at the sentence level, may offer more detailed insights into traveller sentiments; however, this still requires the development of bespoke Machine Learning models to provide a more comprehensive sentimental analysis of travellers' online reviews. Deep Learning models allow for a more granular understanding and classification of traveller feedback, aligning with the latest advancements in Natural Language Processing. By integrating this well-established methodology with the Aspect-Based Sentiment Analysis of customers' online reviews of airport services, this study not only adheres to the rigorous standards of current research, but also introduces a novel perspective to the field of sentiment analysis.

3. Methodology

3.1. Dataset

In this study, a total of 319,000 reviews were collected from social media using Outscraper 2023 online version, with the majority of the reviews being extracted from the COVID-19 outbreak period between 2020 and 2021. Specifically, approximately 100,000 reviews were sourced from Google Maps, 80,000 from Twitter, and the remainder from Airline Quality. It is important to note that no personal data were gathered. Feedback pertaining to flights and other topics besides airport services was excluded from the dataset. The aim of this study can be illustrated based on the snippet below from the dataset collected:

"...Didn't bother using the 'canteen' in the cupboard (departure lounge 1) and there was a big crush fighting fellow passengers to get on the aircraft. The toilets in the lounge were filthy, stinking, and graffitied..."

Previous research may tag this feedback as either neutral due to the mention of the "canteen" without specific negative or positive associations, or as negative because of the negatively described "toilets". However, such approaches would not tag the feedback based on these two specific services mentioned.

3.2. Method

Figure 1 shows the multiclass sentiment analysis process of travellers' online reviews for airport services based on Aspect-Based Sentimental Analysis and Machine Learning.

In this study, travellers' reviews underwent initial preprocessing, including tokenisation and misspelling checking, utilising the tokenizer of NLTK v2.8.1 and the TextBlob v0.18.0 library, respectively. Then, a list of aspects was extracted according to the two-level category of airport services defined in the literature above. The related terms such as food, seats, toilets, and Wi-Fi facilities were searched by sentence tokenisation. The tokens were then analysed using Aspect-Based Sentiment Analysis when an aspect was found, and the sentence was tagged based on the highest positive or negative score. Subsequently, a matrix of aspects was generated with the polarities of a traveller's feedback by airport services stated in the feedback. The polarity for each airport service was determined based on the positive/negative occurrences in the feedback. For instance, if the majority of the aspects and the average of the polarity values showed a tendency toward negativity, the airport service was tagged as -1 . A database was then created, containing approximately 300,000 records with the travellers' feedback, keywords, and service tags. Around 19,000 records were excluded due to the limited number of words. These datasets were subsequently utilised to train, validate, and test the Machine Learning models developed in this study.

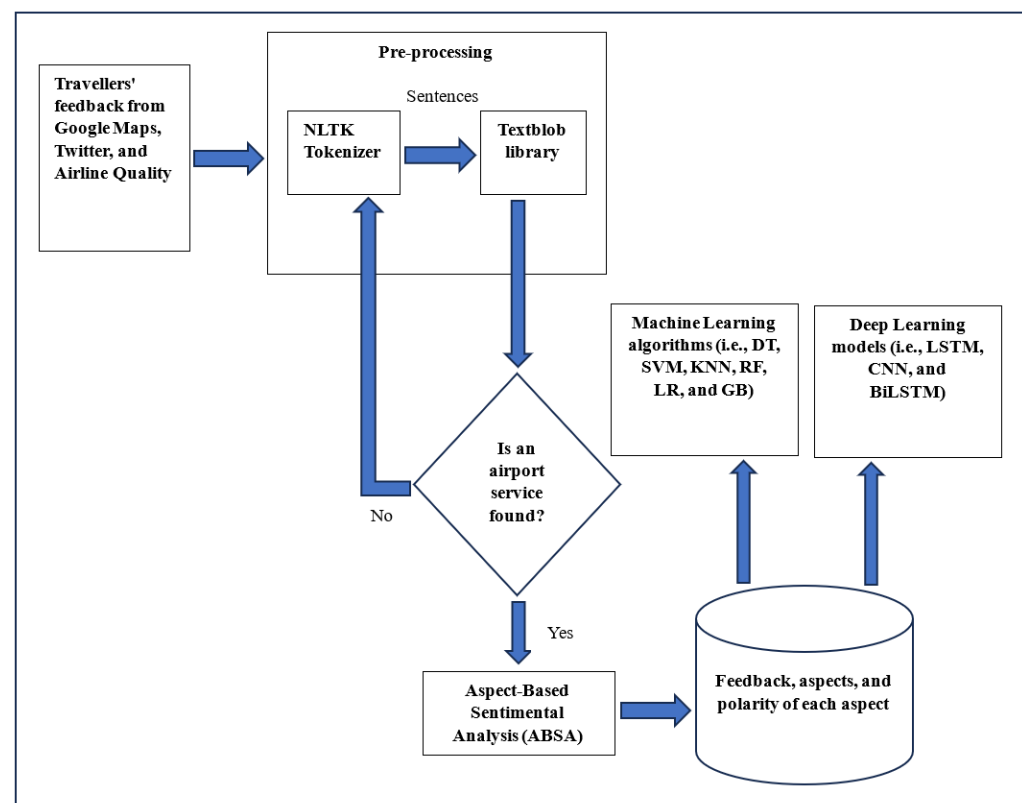


Figure 1. Multiclass sentiment analysis process of online reviews for airport services.

The CNN-based Deep Learning model is outlined in Table 1. The input, including the passengers' feedback and keywords, was fed into the embedding layer, transforming textual inputs into continuous vector representations. Its hidden layers consist of a flatten layer and two dense layers with 64 and 32 neurons, respectively, all configured with the ReLU activation function. The final dense layer comprises 7×3 neurons using the softmax activation function. The model was trained using the Adam optimiser based on categorical cross-entropy loss function and evaluated using an accuracy metric. The dataset was split into 70% for training, 20% for validation, and 10% for testing.

Table 1. CNN-based model.

Layer	Param#	Output Shape
Embedding	1,280,000	(100, 128)
Flatten	0	(None, 12,800)
Dense	819,264	(None, 64)
Dense	2080	(None, 32)
Dense	693	(None, 21)
Reshape	0	(None, 7, 3)

The LSTM-based model is shown in Table 2, which comprises an embedding layer, two LSTMs, and a pair of fully connected dense layers, configured with the ReLU activation function for non-linearity and a softmax activation function for multiclass classification. A similar model is presented in Table 3, with the only difference being the addition of dropout and recurrent dropout with a rate of 0.2 in the two LSTM layers to mitigate overfitting. Additionally, a dropout layer with a rate of 0.2 was inserted between the two fully connected dense layers. Table 4 presents an LSTM model with GloVe embeddings. The “trainable” parameter was set to “False,” ensuring that the pre-trained embeddings were not fine-tuned during training. The incorporation of pre-trained word embeddings

enhanced the model's ability to capture the semantic meaning of the words in the input data, thereby improving its overall performance.

Table 2. LSTM-based model.

Layer	Param#	Output Shape
Embedding	1,000,000	(None, 100, 100)
LSTM	42,240	(None, 100, 64)
LSTM	12,416	(None, 32)
Dense	2112	(None, 64)
Dense	1365	(None, 21)
Reshape	0	(None, 7, 3)

Table 3. LSTM-based model with dropout layers.

Layer	Param#	Output Shape
Embedding	1,000,000	(None, 100, 100)
LSTM	42,240	(None, 100, 64)
LSTM	12,416	(None, 32)
Dense	2112	(None, 64)
Dropout	0	(None, 64)
Dense	1365	(None, 21)
Reshape	0	(None, 7, 3)

Table 4. LSTM-based model with GloVe embedding and dropout layers.

Layer	Param#	Output Shape
Embedding	1,000,000	(None, 100, 100)
LSTM	117,248	(None, 100, 128)
LSTM	49,408	(None, 64)
Dense	4160	(None, 64)
Dropout	0	(None, 64)
Dense	1365	(None, 21)
Reshape	0	(None, 7, 3)

Table 5 presents a BiLSTM model featuring a bidirectional LSTM with 512 units and multiple CNN layers. The bidirectional LSTM captures both forward and backward sequential information, thereby enhancing the model's understanding of context and dependencies. The first layer comprises 256 filters with a 5-unit kernel and the ReLU activation function, while the second layer consists of 128 filters, also employing the ReLU activation function. The global max-pooling layer is used to extract the most relevant features from the previous layers, making the model robust to variations in the input data. The model contains a fully connected dense layer with 64 neurons and the ReLU activation function. The dropout layer with a rate of 0.2 helps mitigate overfitting. The output layer consists of 7×3 neurons and the softmax activation function. The reshape layer formats the output into a 7×3 matrix. The model was trained using the Adam optimiser based on categorical cross-entropy loss function and evaluated by the accuracy metric.

In addition to the advanced Deep Learning techniques of CNN and LSTM, this study also employed a suite of traditional Machine Learning algorithms to provide a comprehensive analysis. These include Decision Trees, Support Vector Machines (SVMs), K-Nearest Neighbours (KNNs), Random Forest, Logistic Regression, and Gradient Boosting. Decision Tree is a non-parametric supervised learning method used for classification and regression. The algorithm creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Support Vector Machine (SVM) is renowned for its ability to handle high-dimensional data and perform classification tasks by finding the optimal hyperplane that maximises the margin between different classes.

K-Nearest Neighbours (KNNs) is a simple, yet effective algorithm that classifies a new data point based on the majority vote of its k-nearest neighbours, thereby assigning it to the most common class among those neighbours. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of the individual trees, providing a more robust and accurate model. Logistic Regression is a linear model used for classification tasks. It estimates the probabilities, particularly useful for cases where one needs to provide a probability score for observations. Gradient Boosting is an ensemble technique that builds the model in a stage-wise fashion. It constructs new models that predict the residuals or errors of prior models and then combines them to make the final prediction, thereby enhancing accuracy. These traditional Machine Learning algorithms complemented the CNN and LSTM models in this research by providing a comparison approach to analysing the dataset. Their inclusion ensured a robust and comprehensive analysis, leveraging the strengths of each method to enhance the overall predictive performance of the study.

Table 5. BiLSTM-based model with dropout layers.

Layer	Param#	Output Shape
Embedding	1,000,000	(None, 100, 100)
Bidirectional	2,510,848	(None, 100, 1024)
Conv1D	1,310,976	(None, 96, 256)
MaxPooling1D	0	(None, 19, 256)
Conv1D	163,968	None, 15, 128)
GlobalMaxPooling1D	0	(None, 128)
Dense	8256	(None, 64)
Dropout	0	(None, 64)
Dense	1365	(None, 21)
Reshape	0	(None, 7, 3)

4. Results

In this research, the model training stops when either the training loss or validation loss stabilises. The CNN-based model exhibits an overall accuracy of 0.94, as depicted in Figures 2 and 3. However, there is a potential concern regarding overfitting, given the high training accuracy of 0.99. The area under the curve (AUC) is a crucial metric in Machine Learning for assessing the overall performance of a classification model. It measures the model's ability to differentiate between different classes, with a higher AUC value indicating better accuracy and reliability. The AUC, as shown in Figure 4, is above 0.75, indicating good discrimination among classes, where class 0, 1, and 2 represent negative, non-existent, and positive sentimental values, respectively, assigned to specific airport aspects from the passengers' feedback.

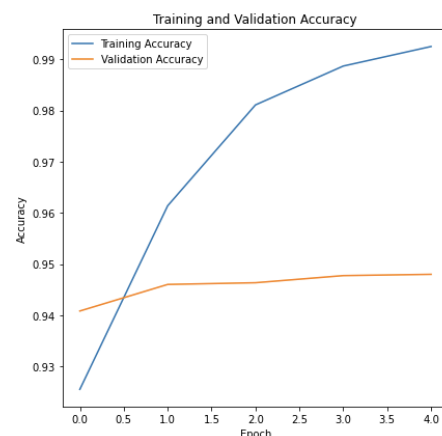


Figure 2. Accuracy of CNN-based model.

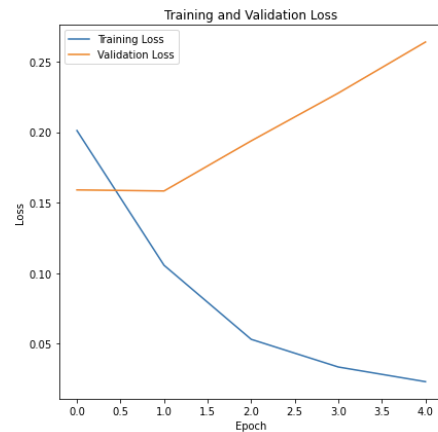


Figure 3. Validation loss of CNN-based model.

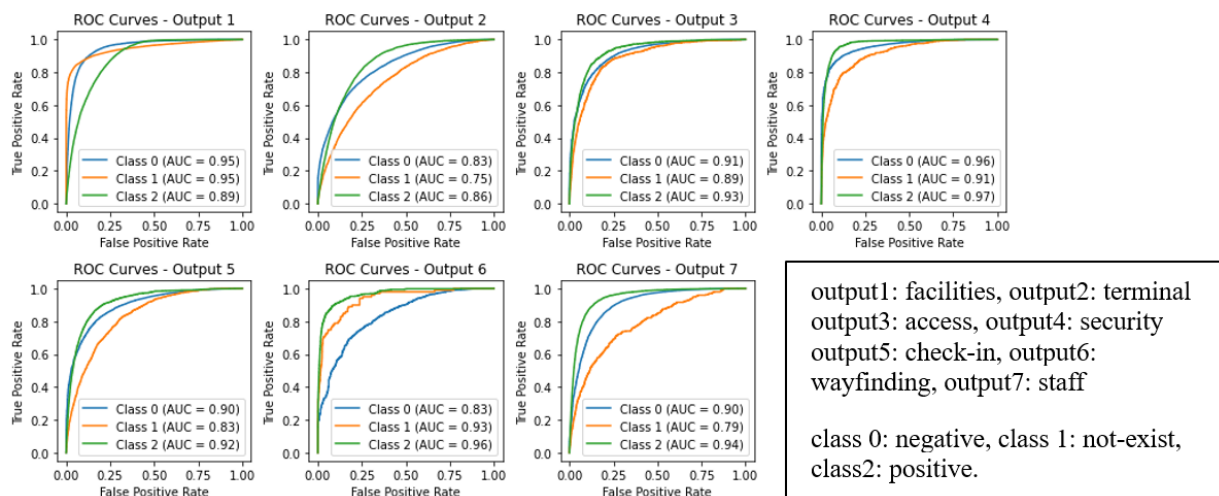


Figure 4. ROC curves of CNN-based model.

The accuracy of the LSTM-based model is 0.94, as shown in Figures 5 and 6. As the model hyperparameters in Tables 3–5 are particularly designed to overcome overfitting, the results only exhibit slight potential overfitting against the training accuracy of 0.97. The AUC in Figure 7 is above 0.77, indicating good discrimination among the classes.

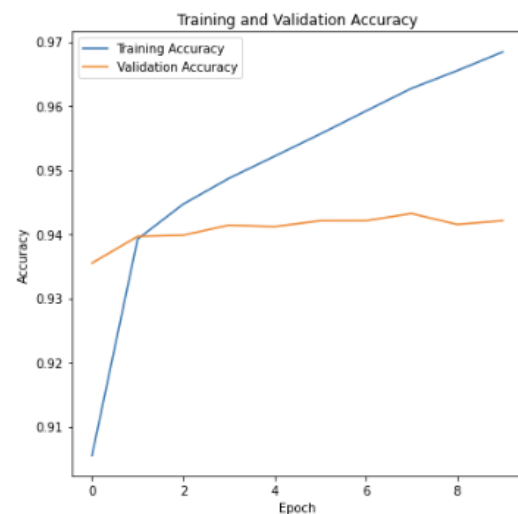


Figure 5. Accuracy of LSTM-based model.

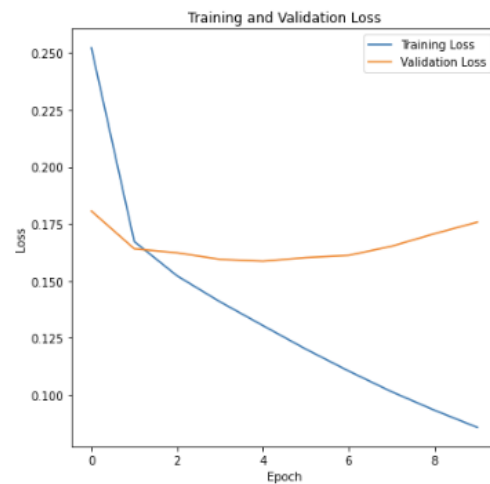


Figure 6. Validation loss of LSTM-based model.

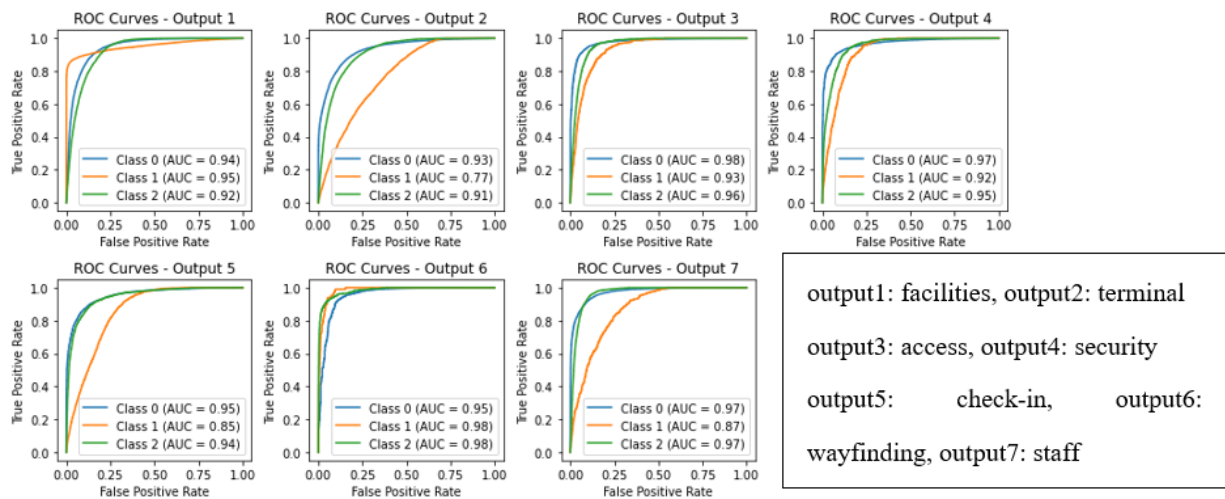


Figure 7. ROC curves of LSTM-based model.

The accuracy of the LSTM-based model with dropout layers in Figures 8 and 9 is 0.94, compared to a training accuracy of 0.95, suggesting that the model performs well on both training and testing datasets without evident signs of overfitting. This is indicative that the model generalises well to new and previously unseen data. The AUC in Figure 10 is within the range of 0.65–0.97, suggesting that the model's ability to distinguish one class from other classes varies across the seven outputs.

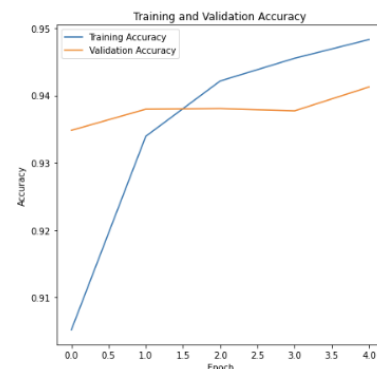


Figure 8. Accuracy of LSTM-based model with dropout layers.

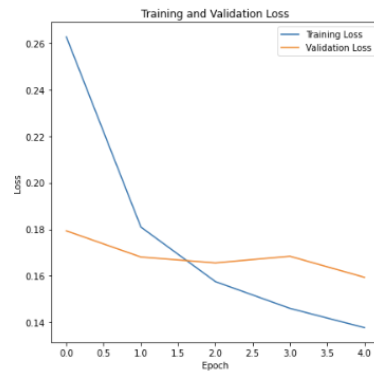


Figure 9. Validation loss of LSTM-based model with dropout layers.

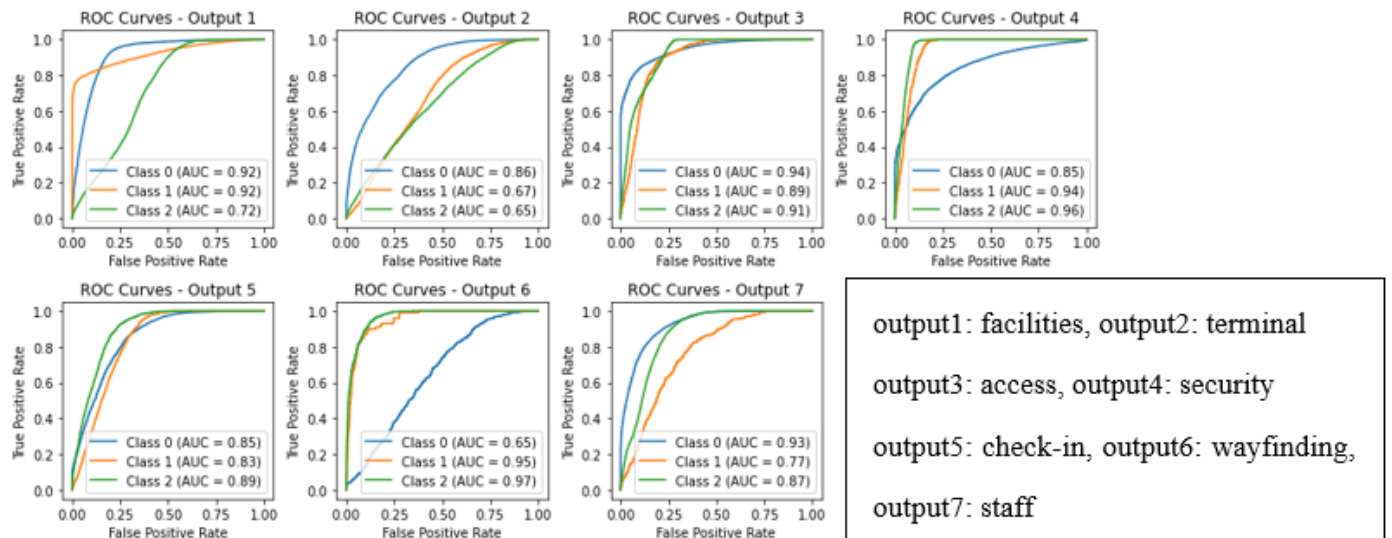


Figure 10. ROC curves of LSTM-based model with dropout layers.

The accuracy of the LSTM-based model with GloVe embedding and dropout layers shown in Figures 11 and 12 is 0.94, which is the same as the one obtained from training, suggesting that the model performs well on both the training and testing datasets without evident signs of overfitting. The AUC in Figure 13 is within the range of 0.65–0.97, suggesting that the model's ability to distinguish one class from other classes varies across the seven outputs.

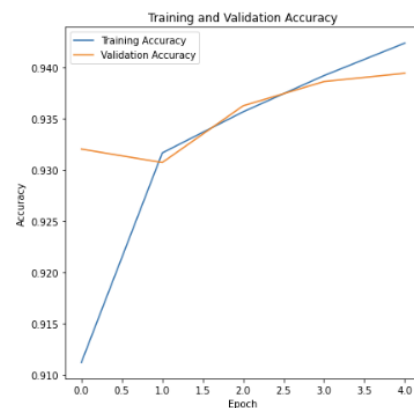


Figure 11. Accuracy of LSTM-based model with GloVe embedding and dropout layers.

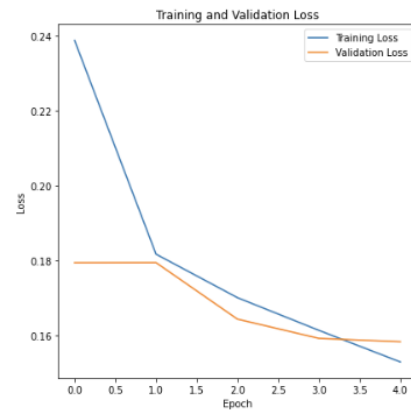


Figure 12. Validation loss of LSTM-based model with GloVe embedding and dropout layers.

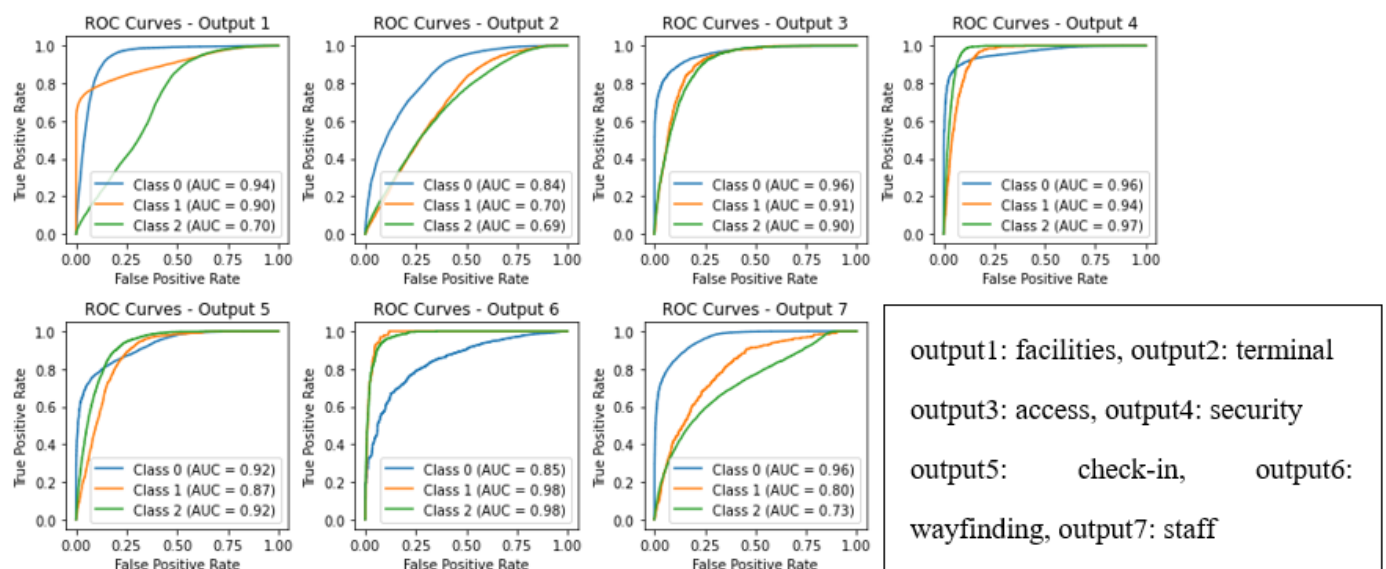


Figure 13. ROC curves of LSTM-based model with GloVe embedding and dropout layers.

The accuracy of the BiLSTM-based model with dropout layers, as shown in Figures 14 and 15, is 0.80, suggesting the possibility of overfitting against its training accuracy of 0.95. The ROC curves of the model with various outputs are shown in Figure 16. This architecture produces the poorest performance among all Deep Learning models.

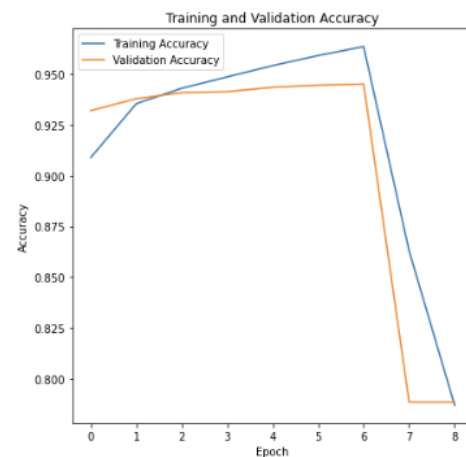


Figure 14. Accuracy of BiLSTM-based model with dropout layers.

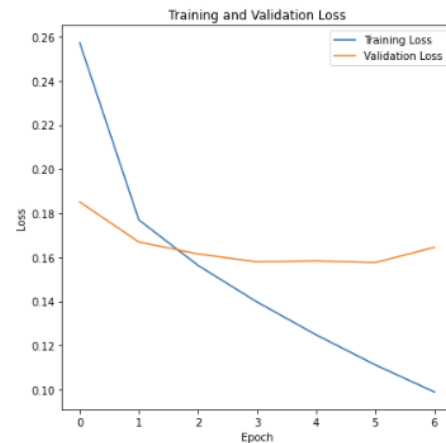


Figure 15. Validation loss of BiLSTM-based model with dropout layers.

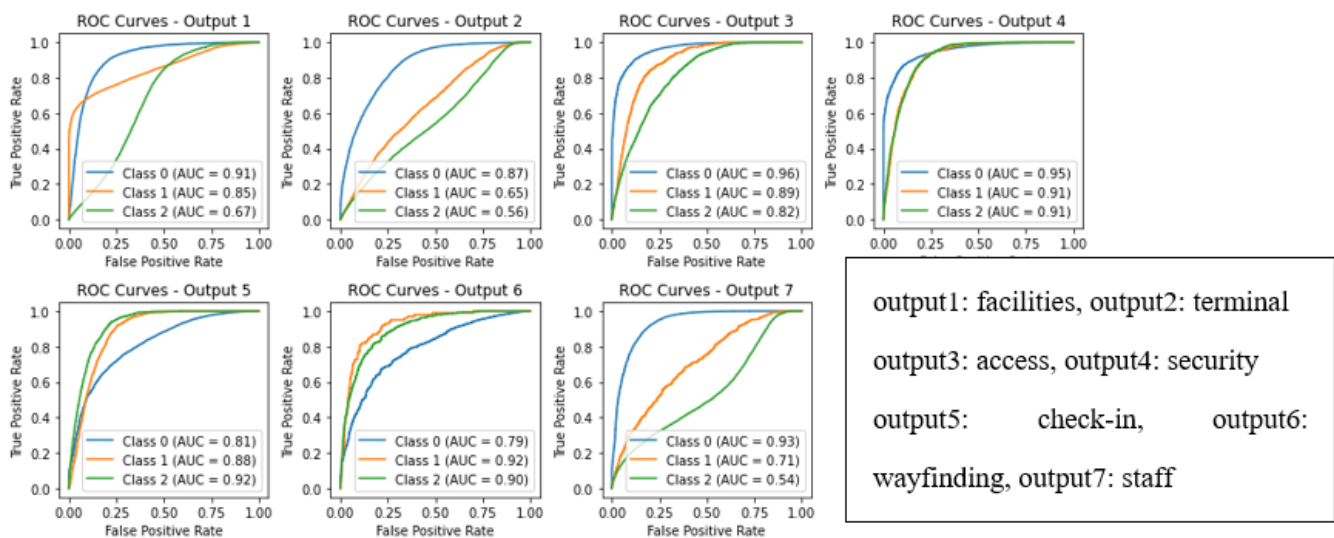


Figure 16. ROC curves of BiLSTM-based model with dropout layers.

The performance of the Machine Learning classification algorithms is presented in Table 6. Overall, the algorithms perform well, with the exception of the GB and KNN algorithms. Notably, the Random Forest algorithm exhibits the highest performance. This research highlights that Machine Learning algorithms such as Random Forest are excellent choices for the multiclass prediction of airport service quality based on travellers' feedback.

Table 6. Accuracy of Machine Learning algorithms for airport services polarity classification.

ML	Facilities Class	Terminal Class	Access Class	Security Class	Check-in Class	Wayfinding Class	Staff Class	Overall Accuracy
DT	0.88	0.87	0.96	0.97	0.94	0.99	0.95	0.94
SVM	0.83	0.82	0.95	0.96	0.92	0.99	0.94	0.92
KNN	0.63	0.72	0.93	0.95	0.91	0.99	0.87	0.86
RF	0.92	0.90	0.97	0.97	0.95	0.99	0.97	0.95
LR	0.82	0.82	0.95	0.96	0.92	0.99	0.94	0.91
GB	0.28	0.68	0.81	0.83	0.80	0.85	0.97	0.75

5. Discussion

In this study, two Deep Learning architectures, i.e., an LSTM-based model with dropout layers and an LSTM-based model with GloVe embedding and dropout layers, show good performance compared to the other architectures without displaying signs of overfitting. This indicates that the models generalise well to new and previously unseen data. The areas under the curve (AUCs) of the Deep Learning models all fall within the range of 0.65–0.97, suggesting that the models’ ability to distinguish one class from other classes varies across the seven outputs. An AUC indicates a model’s ability to correctly classify instances of any class while minimising false positives. A higher ROC curve value (e.g., 0.97) generally indicates that the model has good discrimination capability for that class. A value of 0.65 is still acceptable, but may indicate some degree of overlap between certain classes. The range of 0.77 to 0.97 for the classes is a strong indication that the model can effectively discriminate between these classes and the remaining classes. This range shows that the models have a good balance between true positives and false positives for these classes and are capable of correctly classifying instances belonging to these classes while keeping a relatively low rate of false positives. The high ROC values for classes also suggest that the model generalises well to unseen data, which is an important aspect of model performance. In cases where class distributions are imbalanced, the ROC values may need to be interpreted carefully. Figure 17 shows that there exists a data imbalance, in which the positive values are considerably less than the negative values. The extremely high ROC values could be a result of class imbalances or other factors, and it is important to consider other metrics and domain knowledge.

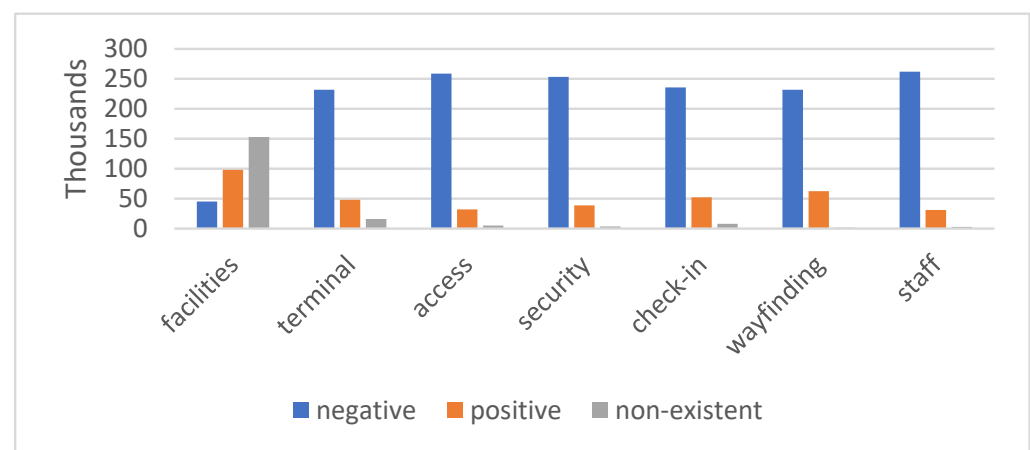


Figure 17. Sentimental classes of tagged airport services (unbalanced).

Aspect-Based Sentimental Analysis provides more insight into travellers’ sentiments than a single tag on a traveller’s review. Nevertheless, the method should only be adopted when the aspects are already known, different from traditional LSA and LDA methods. Another shortcoming of LDA is that one has to predetermine the topic numbers to associate a group of words with a specific topic. The current ABSA methods trained on various datasets such as restaurants, mobile phones, and computer sales produce an approximate accuracy of 80% from the literature, consistent with the performance of the models in this study. It was noticed that the iterative training on airport services based on the feedback may generalise well, but may not accurately reflect the individual sentiments. Further travellers’ feedback datasets may be needed for training using the current ABSA methods.

This work makes a significant theoretical contribution by proposing a comparative approach to quantitatively analyse the sentimental values within travellers’ feedback based on various Deep Learning and Machine Learning techniques. In contrast to current studies that predominantly employ topic modelling to assign an overarching sentimental value (i.e., positive or negative) to the entirety of travellers’ feedback, our approach fills a gap by providing a more granular analysis. Specifically, it delves into the individual feed-

back regarding each airport service and the sentiments expressed by distinct travellers. Moreover, unlike previous works that focus on sentiment analysis at the sentence level, often based on brief Twitter feedback, this study offers a more comprehensive analysis and captures nuanced sentiments towards specific airport services. This distinction is crucial, as many airport services are outsourced to third-party entities, necessitating a detailed understanding of traveller sentiment for effective feedback management. The practical contribution of this work lies in its empirical feasibility to accurately determine airport services and the associated sentimental feedback. By showcasing the effectiveness of the approach, the research establishes that it is practically viable to discern sentiments related to specific airport services. The methodology employed in this study can be replicated for other types of services, given a comprehensive list of defined services.

In summary, the research contributes both theoretically and practically by introducing a comparative method to quantitatively assess sentimental values in travellers' feedback, addressing the gaps left by current studies. The nuanced approach taken in the analysis provides valuable insights into the sentiments associated with individual airport services, offering a practical framework that can be adapted for broader applications beyond the aviation industry.

6. Conclusions

This study develops a multiclass model that categorises traveller feedback based on specific airport services, moving beyond the general sentiment polarity of positive/negative. Contrasting with the traditional sentence-level aspect-based models, which are often inefficient and require iterative application, the research develops and employs both Deep Learning and traditional Machine Learning techniques for a more efficient and accurate multiclass sentiment prediction in online airport service reviews. This research also categorises seven airport services through conducting a comprehensive literature review.

While qualitative Natural Language Processing (NLP) methodologies like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) offer insights into topic modelling and sentiment analysis, they often necessitate human interpretation to fully understand the underlying sentiments and topics. This study addresses this limitation by conducting a distinct analysis of the polarity associated with each airport service with a traveller's feedback. It utilises seven frequently mentioned services identified as the basis for sentiment extraction. The comparative analysis of traditional Machine Learning algorithms and Deep Learning models in the study reveals an intriguing outcome: traditional algorithms, particularly the Random Forest algorithm, demonstrate superior performance in the multiclass prediction of airport service quality using traveller reviews. This finding underscores the potential of integrating different approaches for enhanced analysis in sentiment prediction.

This approach assists airport management in pinpointing key areas of traveller concern and enables more targeted improvements in service areas. Future work aims to incorporate more social media data and airline quality data into the multiclass models, which may enable a more precise and comprehensive prediction of airport service quality and refine the methods available for airport authorities to enhance traveller experience and service efficiency.

Author Contributions: Methodology, M.S.M.A. and J.L.; Software, M.S.M.A.; Data curation, M.S.M.A.; Writing—original draft, M.S.M.A.; Writing—review & editing, J.L.; Supervision, J.L. and K.W.J.; Project administration, J.L. and K.W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by UKRI Innovate UK [Application number: 10003388; PCR—32368; Project title: ONEHeart (Out of Cycle Next Generation Highly Efficient Air Transport), WP4].

Data Availability Statement: The data will be made available by the authors upon the completion of the project.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Barakat, H.; Yeniterzi, R.; Martín-Domingo, L. Applying deep learning models to twitter data to detect airport service quality. *J. Air Transp. Manag.* **2021**, *91*, 102003. [\[CrossRef\]](#)
- da Rocha, P.M.; Costa, H.G.; da Silva, G.B. Gaps, Trends and Challenges in Assessing Quality of Service at Airport Terminals: A Systematic Review and Bibliometric Analysis. *Sustainability* **2022**, *14*, 3796. [\[CrossRef\]](#)
- Dhini, A.; Kusumaningrum, D.A. Sentiment Analysis of Airport Customer Reviews. In Proceedings of the 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Bangkok, Thailand, 16–19 December 2018.
- Li, L.; Mao, Y.; Wang, Y.; Ma, Z. How has airport service quality changed in the context of COVID-19: A data-driven crowdsourcing approach based on sentiment analysis. *J. Air Transp. Manag.* **2022**, *105*, 102298. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mirghafoori, S.H.; Izadi, M.R.; Daei, A. An integrated approach for prioritizing the barriers to airport service quality in an intuitionistic-fuzzy environment. *Cogent Bus. Manag.* **2018**, *5*, 1532277. [\[CrossRef\]](#)
- Aydogan, S. Non-Generic Measurement Structure of Airport Service Quality: A Literature Review. *Suleyman Demirel Univ. J. Fac. Econ. Adm. Sci.* **2021**, *26*, 97–117.
- Gajewicz, Ł.; Walaszczyk, E.; Nadolny, M.; Nowosielski, K. Criteria of quality assessment of regional airport services—A very last picture before the COVID-19 pandemic. *J. Air Transp. Manag.* **2022**, *103*, 102231. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kiliç, S.; Çadirci, T.O. An evaluation of airport service experience: An identification of service improvement opportunities based on topic modeling and sentiment analysis. *Res. Transp. Bus. Manag.* **2021**, *43*, 100744. [\[CrossRef\]](#)
- Lapcin, H.T. Airport Competitive Strengths in Turkey: Primary, Secondary, and Regional Airports. *Transp. Res. Procedia* **2021**, *59*, 300–309. [\[CrossRef\]](#)
- Taecharungroj, V.; Mathayomchan, B. Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tour. Manag.* **2019**, *75*, 550–568. [\[CrossRef\]](#)
- Lee, K.; Yu, C. Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews. *J. Air Transp. Manag.* **2018**, *71*, 28–44. [\[CrossRef\]](#)
- Martin-Domingo, L.; Martín, J.C.; Mandsberg, G. Social media as a resource for sentiment analysis of Airport Service Quality (ASQ). *J. Air Transp. Manag.* **2019**, *78*, 106–115. [\[CrossRef\]](#)
- Moro, S.; Lopes, R.J.; Esmerado, J.; Botelho, M. Service quality in airport hotel chains through the lens of online reviewers. *J. Retail. Consum. Serv.* **2020**, *56*, 102193. [\[CrossRef\]](#)
- Bae, W.; Chi, J. Content Analysis of Passengers' Perceptions of Airport Service Quality: The Case of Honolulu International Airport. *J. Risk Financ. Manag.* **2022**, *15*, 5. [\[CrossRef\]](#)
- Bunchongchit, K.; Wattanacharoensil, W. Data analytics of Skytrax's airport review and ratings: Views of airport quality by passengers types. *Res. Transp. Bus. Manag.* **2021**, *41*, 100688. [\[CrossRef\]](#)
- Halpern, N.; Mwesiumo, D. Airport service quality and passenger satisfaction: The impact of service failure on the likelihood of promoting an airport online. *Res. Transp. Bus. Manag.* **2021**, *41*, 100667. [\[CrossRef\]](#)
- Shadiyar, A.; Ban, H.-J.; Kim, H.-S. Extracting Key Drivers of Air Passenger's Experience and Satisfaction through Online Review Analysis. *Sustainability* **2020**, *12*, 9188. [\[CrossRef\]](#)
- Tian, X.; He, W.; Tang, C.; Li, L.; Xu, H.; Selover, D. A new approach of social media analytics to predict service quality: Evidence from the airline industry. *J. Enterp. Inf. Manag.* **2020**, *33*, 51–70. [\[CrossRef\]](#)
- Ahmed, B.H.; Ghabayen, A.S. Review rating prediction framework using deep learning. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 3423–3432. [\[CrossRef\]](#)
- Kamış, S.; Goularas, D. Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data. In Proceedings of the 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey, 26–28 August 2019.
- AlKheder, S. Passengers intentions towards self-services check-in, Kuwait airport as a case study. *Technol. Forecast. Soc. Chang.* **2021**, *169*, 120864. [\[CrossRef\]](#)
- Antwi, C.O.; Fan, C.-J.; Ihnatushchenko, N.; Aboagye, M.O.; Xu, H. Does the nature of airport terminal service activities matter? Processing and non-processing service quality, passenger affective image and satisfaction. *J. Air Transp. Manag.* **2020**, *89*, 101869. [\[CrossRef\]](#)
- Chonsalasin, D.; Jomnonkwao, S.; Ratanavaraha, V. Measurement model of passengers' expectations of airport service quality. *Int. J. Transp. Sci. Technol.* **2021**, *10*, 342–352. [\[CrossRef\]](#)
- Liu, X.; Zheng, W. Study on Passenger Satisfaction about Service Quality at Terminals of Wuhan Tianhe International Airport. In Proceedings of the EBIMCS 2021: 2021 4th International Conference on E-Business, Information Management and Computer, Hong Kong, China, 29–31 December 2021.
- Saut, M.; Song, V. Influences of airport service quality, satisfaction, and image on behavioral intention towards destination visit. *Urban Plan. Transp. Res.* **2022**, *10*, 82–109. [\[CrossRef\]](#)

26. Usman, A.; Azis, Y.; Harsanto, B.; Azis, A.M. Airport service quality dimension and measurement: A systematic literature review and future research agenda. *Int. J. Qual. Reliab. Manag.* **2022**, *39*, 2302–2322. [[CrossRef](#)]
27. Alaydaa, M.S.M.; Li, J.; Jenkins, K. Aspect-Based Sentimental Analysis for Travellers' Reviews. *Comput. Sci. Inf. Technol. (CS IT)* **2023**, *13*, 477–483. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



DSPACE

<https://dspace.org/>

Multiclass sentiment prediction of airport service online reviews using aspect-based sentimental analysis and machine learning

Alanazi, Mohammed Saad M.; Li, Jun; Jenkins, Karl W.

2024-03-01

MDPI

<https://doi.org/10.3390/math12050781>

Alanazi MS, Li J, Jenkins KW. (2024) Multiclass sentiment prediction of airport service online reviews using aspect-based sentimental analysis and machine learning. Mathematics, Volume 12, Issue 5, March 2024, Article number 781