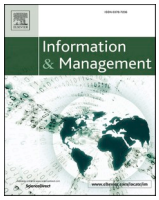
Information & Management 59 (2022) 103587

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im



Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry

Yung-Chun Chang [b], Chih-Hao Ku [a,*], Duy-Duc Le Nguyen [b]

[a] Monte Ahuja College of Business, Cleveland State University, Cleveland, USA
[b] Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

ARTICLE INFO

Keyword:
Aspect-based Sentiment Analysis
Social Media Analysis
Natural Language Processing
Deep Learning
Information Visualization
Bidirectional Encoder Representations from Transformers

ABSTRACT

This study investigates customer satisfaction through aspect-level sentiment analysis and visual analytics. We collected and examined the flight reviews on TripAdvisor from January 2016 to August 2020 to gauge the impact of COVID-19 on passenger travel sentiment in several aspects. Till now, information systems, management, and tourism research have paid little attention to the use of deep learning and word embedding techniques, such as bidirectional encoder representations from transformers, especially for aspect-level sentiment analysis. This paper aims to identify perceived aspect-based sentiments and predict unrated sentiments for various categories to address this research gap. Ultimately, this study complements existing sentiment analysis methods and extends the use of data-driven and visual analytics approaches to better understand customer satisfaction in the airline industry and within the context of the COVID-19. Our proposed method outperforms baseline comparisons and therefore contributes to the theoretical and managerial literature.

1. Introduction

On March 11, 2020, the world changed when the World Health Organization (WHO) declared COVID-19 a global pandemic. It also triggered a precipitous drop in air travel and government-imposed travel restrictions [1]. As the pandemic intensified, so did the impact on the airline industry [2] because flight delays and cancellations are considered core airline service failures [3]. Pere et al. [4] conducted a study showing a lack of consumer confidence, even when flights started to resume. They further pointed out that airline reputations are difficult to build but easy to lose. Therefore, there is a need to understand consumer perspectives because early assessment can help the airline industry prepare to recover and regain consumer confidence.

Customer purchasing decisions are strongly influenced by opinions shared on social media platforms, such as Yelp! and TripAdvisor [5,6]. These platforms reduce the risk and information asymmetries of a product or service before it is purchased [7]. In the tourism and hospitality industry, it is crucial to understand customer preferences through online reviews to improve service quality and the competitiveness of service or product offerings [8, 9].

Most related studies have relied on expert input [10, 11] and surveys [9, 12] to study service quality. Sentiment analysis can also be used to measure user satisfaction with products or services, which can help overcome a firm's weakness and complement existing approaches [13]. Sezgen et al. [14] point out that only a few studies have conducted sentiment analysis of airline reviews in social media, such as Twitter [15], Skytrax [16], and TripAdvisor [8]. However, user satisfaction is driven by several factors, which pose challenges to aspect-based sentiment analysis (ABSA) [17, 18]. Further, the application of ABSA focuses mainly on the extraction of product features [19], and insufficient attention has been given to analyzing tourism reviews using ABSA, especially for air flight traveler reviews [20, 21].

This study fills the above research gap by focusing on various dimensions of airline service and explores how ABSA can gain insights into service quality. In particular, we develop visualizations to observe the relations between aspect ratings and customer satisfaction (overall ratings) before and during the pandemic. Based on the insights learned from the interactive visualizations, we developed a deep learning-based natural language processing (NLP) model with various embedding techniques, e.g., bidirectional encoder representations from transformers (BERT), to capture latent linguistic features from airline reviews. To the best of our knowledge, no previous information systems (IS) research has used BERT and deep learning techniques to analyze airline reviews on TripAdvisor and conduct ABSA in an airline and

* Corresponding author.
E-mail addresses: changyc@tmu.edu.tw (Y.-C. Chang), c.ku17@csuohio.edu (C.-H. Ku), m946106007@tmu.edu.tw (D.-D.L. Nguyen).

https://doi.org/10.1016/j.im.2021.103587
Received 3 November 2020; Received in revised form 25 December 2021; Accepted 26 December 2021
Available online 29 December 2021

COVID-19 context. Using deep learning and information visualization techniques, this study investigates how the pandemic has affected customer satisfaction through different aspects of airline service. We then use the rated aspects to predict the unrated aspects of airline reviews.

2. Related work

This study focuses on three streams of literature: online reviews and customer satisfaction, ASBA, and deep learning-based NLP for ASBA. We review recent studies and applications while highlighting the research gaps, which the current study aims to fill.

2.1. Online reviews and customer satisfaction

The airline industry is vulnerable to global events as COVID-19 has clearly demonstrated [22]. Airline service failures and failed recovery attempts can profoundly influence an airline's financial performance [23] due to loss of customer trust and loyalty [16]. Empirical studies [24,25] have revealed that customer emotion directly affects customer satisfaction and loyalty. Ou and Verhoef [25], for example, analyzed 10,527 customer responses from 102 leading firms in the Netherlands, including four airline firms, with the results showing that positive and negative emotions can influence customer loyalty and purchase intentions. Similarly, Khan and Urolagin [24] collected and analyzed over 50,000 tweets from 18 airlines using sentiment analysis and information visualization techniques. They found that customer sentiments can predict customer loyalty with high accuracy (>90%).

In line with appraisal theory, customers are likely to evaluate a product or service performance when they are exposed to an environmental stimulus [26]. Positive and negative sentiments can be evoked in this evaluation process. According to expectancy disconfirmation theory, when a customer's expectation and perceived service are different, this can affect customer satisfaction [14,27]. Although the extant literature has examined the relation between rating scores and customer satisfaction [28,29], a customer's expectation of service quality depends on several dimensions, such as price, cabin class, and implicit service promises [16]. Hence, using a generic score to measure the aggregated dimensions of satisfaction has limitations because the primary drivers of customer satisfaction can be individual rating categories [8]. Further, little is known about customer satisfaction and service quality during the present pandemic. For this reason, we measured customers' multidimensional sentiments before and after the start of COVID-19 by developing a deep learning-based model for ABSA.

2.2. ABSA

Sentiment analysis or opinion mining uses NLP, text analysis, and computational linguistics to extract meaning and polarity from a text [30]. Such analysis can be carried out at the document, sentence, and aspect levels to detect positive, negative, or neutral opinions conveyed in a text [31]. For instance, sentiment analysis algorithms proposed by Mohammad et al. [32] have been widely used to extract polarity at the sentence level. Although prior studies have examined sentiment analysis on hotels [31,33,34] and restaurant reviews [35,36], scant effort has been spent on conducting sentiment analysis of airline reviews and services [14].

To identify the opinion providers' multidimensional sentiments of the service or product, ABSA can be used. ABSA is a text analysis technique that explores aspects within a document and then allots each one with a sentiment level. ABSA involves three steps: aspect identification or extraction, sentiment classification, and sentiment aggregation or summarization [37,38]. The results are detailed, engaging, and accurate since ABSA investigates more closely the information behind a text. Based on SemEval-2016 Task 5 [39], ABSA comprises three sub-tasks: sentence-level, text-level, and out-of-domain ABSA. The goal of the first subtask is to identify entities, entities' attributes, offsets, and their polarity for a given sentence. As for text-level ABSA, it is a classification that summarizes the opinions expressed in a customer review. The last subtask develops a system to test in domains for which no training data are made available. This study focuses on the second subtask, or text-level ABSA, in organizing given aspects into suitable categories for each review; incidentally, this subtask was called aspect category sentiment analysis in the SemEval-2014 Task 4 [40]. A document may have different sentiment polarities for different aspects and overall polarities. Take the example the review "Moderate ticket prices compared with other airlines, but the legroom (LR) is insufficient," which has an overall *positive* sentiment; however, if the aspects are {VALUE_FOR_MONEY, LEGROOM}, its polarities will be {*positive, negative*}.

ABSA has been recently leveraged to understand service failure in the hospitality industry [33,41]. For instance, Korfiatis et al. [8], who collected 557,208 airline passenger reviews from TripAdvisor, used structural topic models (STM) and latent Dirichlet allocation (LDA) with a probabilistic extension to measure aspect-level service quality. Their findings show that customer service (CS) and value for money (VM) are critical factors that lead to increased customer satisfaction. Similarly, Sezgen et al. [14] used latent semantic analysis (LSA) to analyze over 5,000 passenger reviews from 50 airlines on TripAdvisor and found that customer satisfaction was related to the cabin class and the cost of an airline. Specifically, friendly and helpful staff was the critical satisfaction driver for economy-class passengers, while product value was vital to premium-cabin passengers. For a low-cost airline, low ticket price was the key to customer satisfaction. However, this study had a couple of limitations: first, LSA does not consider the sentence-level meaning of an individual document, which is an inherent limitation of bag-of-words analysis methods [14,42]; and second, the sample size of 5,000 reviews from 50 airlines is relatively small.

2.3. Deep learning-based NLP for sentiment analysis

Recently, IS research has applied deep learning techniques in process prediction [43], healthcare analytics [44], as well as product reviews [45], and has emerged as a new approach for ABSA [15, 46, 47]. Automated sentiment analysis can be classified into lexicon-based and machine learning methods [48]. A lexicon-based method requires existing sentiment dictionaries, which may not meet the requirement of specific domains [49]. Moreover, a self-constructed lexicon is labor-intensive and inefficient. As for traditional machine learning methods, they are incapable of extracting semantic relationships between aspects and relative content words [50]. On the other hand, word embedding is a machine learning technique that transforms words into an n-dimensional vector space [48], thus enabling semantically similar words to have similar vector representations, such as "*strong*" being close to "*powerful*" and "*boat*" being close to "*water*," which depend on training data [51]. Word embedding, typically used in sentiment analysis, therefore overcomes the drawback of a bag-of-words analysis and contributes to the success of deep learning-based NLP [51]. Pre-trained word embeddings have become an integral part of modern NLP systems [52] due to the steady and robust development of language models, such as Word2Vec [53], GloVe [54], and ELMo [55].

In 2018, Devlin et al. [52] from Google introduced BERT, a language model based on attention [56]. Liu et al. [57] then used attention-based deep learning and word embedding techniques to predict shareholder litigation in insider trading from financial texts, while Brahma et al. [58] applied BERT and text analytics techniques to predict mortgage origination delays from textual conversations. For sentiment analysis, the use of attention [59,60] and convolution neural networks (CNN) [47] has emerged as a popular strategy. Wang et al. [61], for example, used attention to focus on an essential part of the sentence, which demonstrates the efficiency of the attention mechanism for ABSA tasks. Similarly, Poria et al. [62] used a seven-layer deep CNN to identify aspects and tagged them with corresponding sentiments. This transfer learning

technique with language models clearly has potential and can overcome lengthy and costly data collection and data labeling [63].

The use of BERT for ABSA is new and has primarily found mostly in conference proceedings. We have reviewed recent ABSA studies using BERT and their datasets and summarized their findings below. Table 1 shows representative BERT-based studies, including the main topic and novelty, the use of corpus, and the experimental results for each study. Unlike previous language models integrated with additional features that carried human knowledge of the end task, BERT adopts a fine-tuning approach that requires almost no specific architecture for each end task [64]. This strategy enables BERT to extract knowledge directly from data. Modem ABSA researchers have been encouraged by the progress made possible by BERT, and several of their performance evaluation datasets use the corpora of SemEval competitions [40] and SentiHood [65]. An example of this is the SemEval-2014 Task 4 dataset, which comprises two domain-specific datasets: laptops and restaurants. This dataset contains the five aspects of food, service, price, ambiance, and anecdotes/miscellaneous. Each aspect has its set of sentiments, including positive, negative, conflicting, and neutral. SentiHood, which is a targeted ABSA dataset collected from the question-and-answer platform of Yahoo!, includes ten aspects: life, safety, price, quietness, dining, nightlife, transit-location, touristy-ness, shopping, green-culture, and multi-culture. Each aspect contains positive and negative sentiment labels.

Table 1
Summary of previous BERT-based works on aspect sentiment analysis

Study	Main topic and novelty	Corpus	Result
[65]	Develop an auxiliary sentence for extracting sentiment polarity and entity information from a text and transforming ABSA into a sentence-pair classification task.	SemEval-2014 Task 4&SentiHood	The fine-tuned BERT model captured crucial information related to ABSA from the constructed auxiliary sentences. The overall performance can achieve F_1-score of 92.18% and 87.90% on SemEval-2014 and SentiHood datasets, respectively.
[66]	Using a gating mechanism with context-aware aspect embeddings to enhance the BERT representation for ABSA.	SemEval-2014 Task 4&SentiHood	The proposed GBCN model enhances the BERT representation for ABSA through learning context-aware aspect embeddings that encode richer ABSA correlated information. The model reached the highest F_1-score of 92.89% on the SemEval-2014 datasets and an F_1-score of 88.0% on SentiHood datasets.
[67]	Develop a method that can identify words corresponding to different targets and different aspects, and correlate aspect and sentiment.	SemEval-2014 Task 4&SentiHood	The proposed model integrates context information into the self-attention calculation of BERT, which can learn the correlation between words, targets, and aspects. Thereby achieving an overall F_1-score of 92.64% and 89.70% on SemEval-2014 and SentiHood datasets, respectively.
[69]	Adopt a manually created ontology to extract the domain-based knowledge for predicting relationships among entities and their properties to benefit the prediction of the polarity value of an aspect.	SemEval 2016 Task 5&SemEval 2015 Task 12	The hybrid ABSA model can extract the polarity of an aspect through learning relationships among the aspect and its context words. The overall performance on SemEval-2015 and SemEval-2016 datasets can obtain high accuracies of 93.1% and 92.7%, respectively.
[70]	The model can capture the dual dependency of sentiments on both targets and aspects and handle implicit target cases.	SemEval-2014 Task 4&SentiHood	The BERT-based multi-task learning ABSA model can capture the dual dependency of sentiments on both targets and aspects and handle implicit target cases, it achieves the remarkable performance with an F_1-score of 65.89% and 58.09% on SemEval-2016 and SemEval-2015 datasets, respectively.

Sun et al. [65] investigated several approaches to develop an auxiliary sentence for extracting sentiment polarity and entity information from a text and transformed ABSA into a sentence-pair classification task. Their fine-tuned BERT model captured crucial information related to ABSA from the constructed auxiliary sentences, thereby achieving an overall F_1-score of 92.18% and 87.90% on SemEval-2014 and SentiHood datasets, respectively. Li et al. [66] built on Sun et al.'s approach to develop a method, which used a gating mechanism with context-aware aspect embeddings to enhance the BERT representation for ABSA, and thus resulted in the capacity to learn context-aware embeddings that encode richer ABSA correlated information. By adding these extra embeddings into the fine-tune BERT, Li et al.'s model reached the highest F_1-score of 92.89% on the SemEval-2014 datasets and an F_1-score of 88.0% on SentiHood datasets. Wu et al. [67] proposed the quasi-attention context-guided BERT that integrates context information into the self-attention calculation. This method was able to 1) identify words corresponding to different targets and different aspects and 2) correlate aspect and sentiment. Therefore, the ABSA prediction performance improved on SemEval-2014 and SentiHood datasets with F_1-scores of 92.64% and 89.70%, respectively.

Recently, most researchers have adopted the SemEval-2015 and -2016 corpora for performance evaluations. Both datasets are an extension of SemEval-2014, which aims to detect the targets, aspects, and sentiment polarities in texts. The SemEval 2015 ABSA dataset [68], collected from the laptops and restaurant domains, contains several aspects formed by the composition of entity and attribute labels. For instance, in the restaurant domain, the entity label "Food" includes five attribute labels "general," prices," "quality," "style and options," and "miscellaneous." In the SemEval 2016 ABSA dataset [39], the training dataset was scaled up by merging the training and test datasets of SemEval 2015, and a new dataset for testing was also created.

Meškelė et al. [69] proposed a hybrid ABSA method using a lexicalized domain ontology and a neural attention model. They adopted a manually created ontology to extract the domain-based knowledge for predicting relationships among entities and their properties to benefit the prediction of the polarity value of an aspect. The proposed neural attention model could extract the polarity of an aspect through learning relationships among the aspect and its context words. They integrated the above advantages into the BERT model and then evaluated performances on SemEval-2015 and SemEval-2016 datasets to obtain high accuracies of 93.1% and 92.7%, respectively. Furthermore, Wan et al.'s [70] BERT-based, multi-task learning method for ABSA predicts whether targets exist for 1) an aspect sentiment pair and 2) a tag sequence for extracting the targets, which reduces the joint detection problem to binary text classification and sequence labeling problems. Since their model can capture the dual dependency of sentiments on both targets and aspects and handle implicit target cases, it achieves the remarkable performance with an F_1-score of 65.89% and 58.09% on SemEval-2016 and SemEval-2015 datasets, respectively.

Our research stands out from recent BERT-based ABSA studies in

many respects. First, the dataset we collected contains airline reviews on one of the largest online review sites, TripAdvisor, thus containing significantly more reviews than previous works. Although each review contains up to eight aspects, some reviewers provide few or even no aspect ratings, which poses additional challenges for data analyses. Second, recent research studies adopt a two-staged method of first recognizing the aspect and then predicting the sentiment. In contrast, our approach treats ABSA as a multi-task classification problem that can enable our model to simultaneously predict aspects and the corresponding sentiments in one stage. Third, our method successfully exploits the syntactic structures, sentiment semantics, and content of review texts. As a result, it can capture discriminative textual features and eliminate undesired noise to boost sentiment classification performance. Finally, we infuse COVID-19 factors into our deep learning model to further improve the performance of the current BERT-based ABSA approach. In short, our proposed method is efficient in ABSA prediction and achieves the best performance among the compared methods.

3. Methodology

3.1. Data collection and preprocessing

Tripadvisor.com is the world's largest travel platform[1] [71] and is honored annually with the Traveler's Choice Best of the Best award based on reviews, ratings, and other relevant factors[2]. According to travelers on TripAdvisor, we selected the top 10 airlines in the world for 2019 and 2020[3]. Two airlines ranked in 2019 dropped out of the top 10 rankings in 2020, which resulted in 12 airlines being selected.

We developed a web crawler and collected data from January 2016 to August 2020. For each review, we stored the airline URL, aspect ratings, crawl time, cabin class, route (e.g., departure and destination airport), type of flight (e.g., international and domestic), review rating, review text, review date, reviewer URL, reviewer name, review title, travel date, response text, response date, and responder name. Each review on TripAdvisor contains up to eight aspects: LR, seat comfort (SC), in-flight entertainment (FE), CS, VM, cleanliness (CN), check-in and boarding (CB), and food and beverage (FB). However, not all reviewers rate all aspects, and some do not rate any, which poses additional challenges for data analyses. Fig. 1 shows a sample review of an airline on TripAdvisor. We highlighted the information extracted from each review stored in a comma-separated values file.

Data preprocessing and cleaning were conducted to enhance data quality. We first extracted numerical aspect values from each aspect, with the collected reviews, including four cabin classes: first-class, business, premium economy, and economy. However, not all 12 airlines offered these four classes. Since we were particularly interested in service ratings before and during COVID-19, we split the reviews into two groups: before and during the pandemic. Reviews that mentioned keywords like "COVID-19," "pandemic," and "refunds" were extracted for additional analyses. Reviews with date errors, e.g., June 2038, were removed from our analysis.

3.2. Data summary and visualization

Table 2 presents an overview of the data summary for each airline. The first column indicates the airline ranking in 2019 and 2020 based on the consumer voting results published by TripAdvisor. The airlines' headquartered countries are classified into the four regions of Pacific Ocean, Asia, South America, and North America, and the Middle East according to the proximity of the countries; this is because some airlines, such as Azul and Korean Air, have relatively fewer reviews. The total number of reviews was 191,123, while the average review rating value was 4.26 out of 5.

Textual reviews and aspect ratings are helpful indicators of travelers' overall satisfaction [72]. We used January 1, 2020, as a cutoff point to separate reviews into before and after the start of COVID-19. Next, the trends of average review ratings before and during COVID-19 were broken down by airline names, with blue and red, respectively, indicating the average rating before and during the pandemic. Overall, the average review rating during the pandemic is relatively lower than the rating before the pandemic for all airlines, as shown in Fig. 2.

To reveal the variation of rating changes during COVID-19, we developed a box-and-whisker plot. The average review rating for each travel month was broken down by quarter, and colors show the details about the airlines. As shown in Fig. 3, the average rating score in January and February 2020 was normal. However, after the WHO's announcement of the global pandemic in March 2020, there is a noticeable drop in average rating scores. This further confirms a significant rise in airline passenger complaints amid COVID-19[4].

Of course, an average review rating cannot tell us the complete story. We therefore developed the visualization in Fig. 4 to display the average ratings of the eight aspects (CB, LR, FB, SC, FE like Wi-Fi, TV, and movies, CS, VM, and CN). We used January 1, 2020, as a cutoff line to split the aspect ratings into before and during pandemic, and the colors show the details of the aspects. Most aspect ratings dropped after the start of the pandemic except for the LR and SC ratings, which increased slightly. The service rating presents the most significant drop (-.16) from before to during the pandemic. This further made us curious to investigate the service rating for ticket refunds and flight cancellations. To this end, we extracted the keywords of "refund," "reimburse," and "cancellations," from airline reviews during COVID-19. We then visualized the results, as shown in Fig. 5, to denote CS and review ratings as red and blue, respectively, for each airline. The numeral next to each airline icon shows the total number of reviews containing the extracted keywords. The average service rating was 2.3 out of 5 when flight reviews involved a refund, reimbursement, or cancellation. The highest service rating was given to Jet.com and ANA airlines with service rating scores of 3.9 and 3.7, respectively. This finding corresponds to the user stories found in the news media[5] and reveals that complicated refund processes, policies, and refund transparency are still the major concerns of many airline travelers during the pandemic.

To understand the passengers' opinion toward each operating airline during the pandemic, we classified the airlines into four regions: Asia, Middle East, South America, North America, and the Pacific Ocean based on their headquarters and geographical proximity (see Fig. 6). The visualization is filtered after the start of the pandemic (January 1, 2020). We use blue-teal stepped color to indicate aspect ratings, where a lighter teal represents a higher rating and darker teal a lower rating. Different passenger concerns correlated with different geographic regions: FE was the primary driver for satisfaction for airlines operating in South America, and North America; VM for airlines in the Middle East and the Pacific Ocean; and FB for airlines in Asia.

Contemporary studies have found that customer satisfaction and dissatisfaction are often related to airline cabins [14]. This study analyzed eight aspects for the four cabin classes of first-class, business, premium economy, and economy during the pandemic (see Fig. 7); note that particular cabins, such as the premium economy may not be an available option for all airlines and flights. The colors in the figure show

[1] About TripAdvisor, https://tripadvisor.mediaroom.com/us-about-us

[2] Travelers' Choice Best of the Best, https://www.tripadvisor.com/TravelersChoice

[3] Top 10 Airlines—World, https://www.tripadvisor.com/TravelersChoice-Airlines

[4] US airlines saw a 965% rise in passenger complaints: https://www.foxnews.com/travel/us-airlines-customer-service-complaints-coronavirus

[5] Airline refund story, https://www.telegraph.co.uk/travel/advice/airline-not-refunded-flights-covid-insurance/.

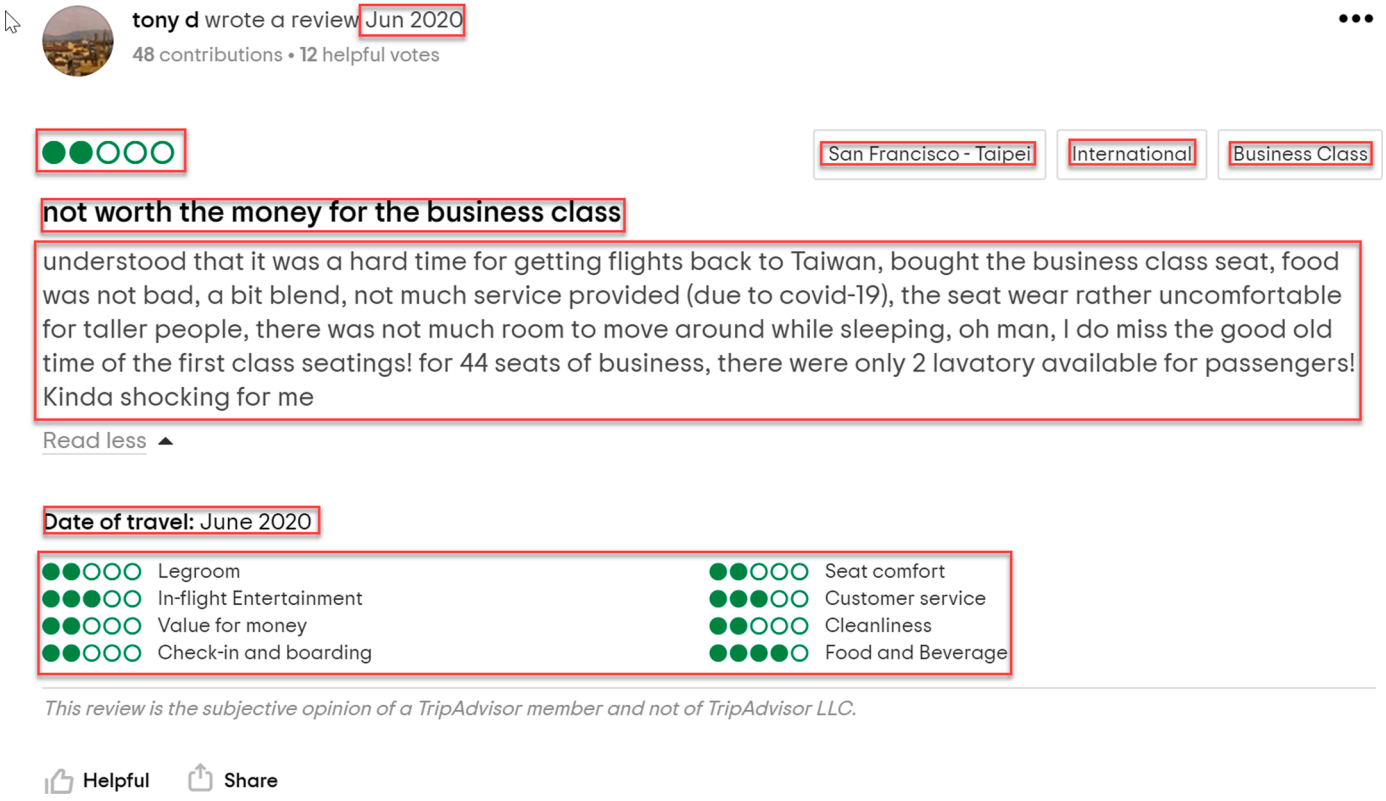


Fig. 1. An airline review on TripAdvisor

details about each cabin class, and the marks are labeled by the number of records with 16 null values excluded. The hollow and filled airline shapes represent before and during pandemic, respectively. During the pandemic, lower aspect ratings were observed for the four cabins, while slightly higher ratings on LR and SC were found for economy passengers. Fewer passengers and social distancing likely explain the slightly higher ratings of LR and SC for economy passengers. It is also noteworthy that first-class and premium economy passengers were highly unsatisfied with FB during the pandemic.

However, the overall ratings of all airlines for each cabin cannot reflect the ratings for each airline. For example, during the pandemic, EVA Air's premium economy passengers gave lower ratings on most aspects, typically on CN, check-in, FB, FE, and VM (ratings < 4), as shown in Fig. 8. Therefore, by examining aspect ratings for each cabin, airline managers can be alerted to aspects that require attention to further improve the service quality.

Next, we retrieved the reviews containing specific keywords, such as a pandemic, COVID, and virus, and explored the aspect ratings in these reviews. It should be noted that three null values were excluded, no ratings were found from premium economy passengers during the pandemic. Fig. 9 presents these results, with the number next to each airline representing the number of reviews containing these keywords. The bottom row of the figure indicates reviews containing pandemic keywords for all cabin classes. Overall, we found slightly lower ratings on reviews with pandemic keywords for most cabins. Additionally, low ratings were found on check-in, service, and VM for two first-class passengers. However, the small sample size limits our ability to interpret these findings.

We then calculated the correlation between each aspect rating and the overall rating (satisfaction) for all cabin passengers using Pearson's correlation coefficient. Fig. 10 shows the overall results, where values that more closely correlate with 1 indicate a stronger relationship between an aspect rating and overall satisfaction. The result indicates that CS and VM are essential indicators of customer satisfaction. Similarly, Rajaguru [12] conducted a survey study on 15 full-service and six low-cost carrier customers and found that VM and customer satisfaction are correlated.

However, we cannot apply this finding to all cabin classes. Instead, we used a filter to observe the difference for each cabin, as the example of Fig. 11 shows. Before the pandemic, the CS, FB, and VM had higher correlations to first-class passenger satisfaction; in contrast, CN and FB were the major concerns of first-class passengers during the pandemic. We repeated the same analysis for each cabin and found a similar result for business and premium economy passengers. This is understandable because most airlines had new pandemic policies relating to FB, such as "no snacks, no food," and "no refill of beverages" during the pandemic [73].

We applied the same type of analysis for each airline, but Fig. 12 illustrates a sample result for all cabins of Korean Air. It is interesting that during the pandemic CN and CB became even more important factors influencing passenger satisfaction, typically for ANA, EVA Air, JAL, and Korean Air. Therefore, while traveling during the pandemic, the airlines should pay special attention to improving or adapting the boarding methods to accommodate new social distancing norms without sacrificing passenger safety [73] and satisfaction.

Our visualization results show that customers have different perceptions depending on their choice of cabin class and airline. Interactive visualizations can reveal several attributes that may lead to passengers' satisfaction before and during the pandemic; however, there is no consensus in the literature about the intricate relationship between service quality and passengers' satisfaction [14,16]. Nonetheless, the aspect ratings and customer satisfaction are different before and during the pandemic. Measuring ratings of airline service is complicated because it can be both quantitative and qualitative, and includes seat reservation, ticketing, check-in process, in-flight service, baggage handling, CN, and employee courtesy or a combination of different services [74,75]. To better capture the relationship between rating scores and text reviews, we adopted deep learning-based NLP and word

Table 2
The number of reviews and average review rating for each airline: January 2016–August 2020

Year (rank)	Airline name	Country	Region	Number of reviews	Average review rating
2019 (8), 2020 (6)	Air New Zealand	New Zealand	Pacific Ocean	12,300	4.28
2019 (10)	All Nippon Airways (ANA)	Japan	Asia	3,135	4.35
2019 (7), 2020 (1)	Azul	Brazil	South America	516	3.93
2019 (4), 2020 (10)	Emirates	United Arab Emirates	Middle East	39,757	4.11
2019 (3), 2020 (7)	EVA Air	Taiwan	Asia	4,301	4.30
2019 (5), 2020 (4)	Japan Airlines (JAL)	Japan	Asia	2,809	4.35
2019 (9), 2020 (5)	Jet2.com	England	North America	27,992	4.41
2020(3)	Korean Air	South Korea	Asia	3,148	4.29
2019 (2), 2020 (9)	Qatar Airways	Qatar	Middle East	19,649	4.12
2019 (1), 2020 (2)	Singapore Airlines	Singapore	Asia	19,805	4.29
2019(6)	Southwest Airlines	United States	North America	42,975	4.37
2020(8)	Virgin Atlantic Airways	United Kingdom (England)	North America	14,706	4.11
Total				**191,123**	**4.26**

embedding techniques, and our model also considers pandemic factors.

Before feeding the collected data into our deep learning model, we needed to preprocess the rating data. Aspect rating values can range from 1 to 5 or can be blank, so we classified airline review ratings into either negative (1–3 ratings) or positive (4–5 ratings), removed reviews with no aspect label, and converted words into lower case. Table 3 shows the overall result of the rating ratio of eight aspects. Our goal was to develop a deep learning model that can effectively learn aspect ratings based on review content so that it can predict unrated aspect ratings in the collected reviews.

3.3. Discriminative Linguistic Features Fused with BERT for aspect-based sentiment prediction of airline reviews

The impact of COVID-19 on the airline industry can be evaluated via sentiment analysis of the texts relating to certain aspects of an airline, such as online reviews and social media. In this research, we model the ABSA as a multi-task classification problem and define it as follows. Let be a set of documents, $A = \{a_1, a_2, \ldots, a_m\}$ a set of aspects, with each aspect having its set of sentiments $S = \{s_1, s_2, \ldots, s_n\}$. The goal of this task is to decide the most appropriate sentiment s_i of each aspect a_j for a document d_l, where one or more sentiments can be associated with a document. In this way, we can observe the affected aspects of airlines in a quantified manner.

We constructed a BERT-based model [52] fused with discriminative linguistic features for ABSA. Fig. 13 illustrates an overview of our proposed model, called LiFeBERT, which predicts the sentiment behind different aspects of airline reviews. We first needed to learn the keywords of the sentiments from the input corpus. Then, given the learned keywords, a set of embedding vectors was generated to match several aspects of a document. After this, we further integrated these vectors with two embeddings that were already trained by the original BERT, token embeddings and positional embeddings. The fused representations (vectors) went through multi-head attention layers to predict the aspect with sentiment ratings behind the text of the airline review. We explain the functions of each layer in the following paragraphs.

Input layer—Multi-feature fusion text representation: It is vital to preprocess the raw text data to conduct machine learning efficiently. We first transformed all words to lower case for consistency and removed punctuation; stop words, such as “is” and “the,” were also filtered out. Next, we adopted the WordPiece [76] toolkit to decompose a text into token sequences. The token embedding was generated from the pre-trained BERT to represent the words in an airline review. Positional embedding was also adopted to capture the order information of tokens [56]. As is the convention of using pre-trained BERT, the first token of sequences is the [CLS], which is a unique token designated for classification tasks.

Extracting discriminative lexicons for text representation can eliminate undesired information and boost sentiment classification performance [47,77]. This study infused different embeddings that can highlight the tokens positively associated with each aspect category. For this reason, we utilized the log-likelihood ratio (LLR) method [78] to >extract polarity keywords denoting sentiments of an aspect[6] as shown in Eq. (3). More specifically, a word with a considerable LLR value was closely associated with the sentiment of an aspect. Next, all words were ranked by their LLR value in the training procedure, and the top 150 were selected as polarity keywords for each aspect. After this, the discriminative linguistic embeddings were trained by freezing the other parameters in the LiFeBERT model, which only learns the embeddings of those polarity keywords. Then, we unfroze all parameters and trained a few more epochs. Finally, we ended up with a 768-dimension vector for text representation composed of token embedding, positional embedding, and keyword embedding. We also added a normalization operation [79] before feeding a text representation to the multi-head attention blocks. The attention mechanism was then adopted to learn which parts of the representation needed to be focused on.

Multi-head Attention layer: Vaswani et al. [56] described an attention function as mapping vectors (such as a query and a set of key-value pairs) to an output. A weighted sum of the values is calculated, in which a compatibility function between the query and the key vectors denotes the weight. Specifically for our study, let $Q, K, V \in R^{d_a}$ denote the matrices of query, key, and value, respectively. The output matrix is derived using the attention in Eq. (1). In this paper, we employ multiple transformer layers with multi-head attention instead of a single attention function. The proposed model can jointly attend to information from different representation subspaces at different positions through the multi-head attention mechanism [56]. The critical point is that each attention head looks at the entire input sentence with a different focal point. Let L denote the number of transformer layers, H as the hidden size, and A as the number of attention heads per layer. We set L = 12, H = 768, and A = 12 to take advantage of the pre-trained BERT model[7].

[6] The detail for calculating the LLR value is described in the Appendix.

[7] The pre-trained BERT model is called the BERT-Base model, which contains 12 layers with 768-dims and which was retrieved from https://github.com/google-research/bert.

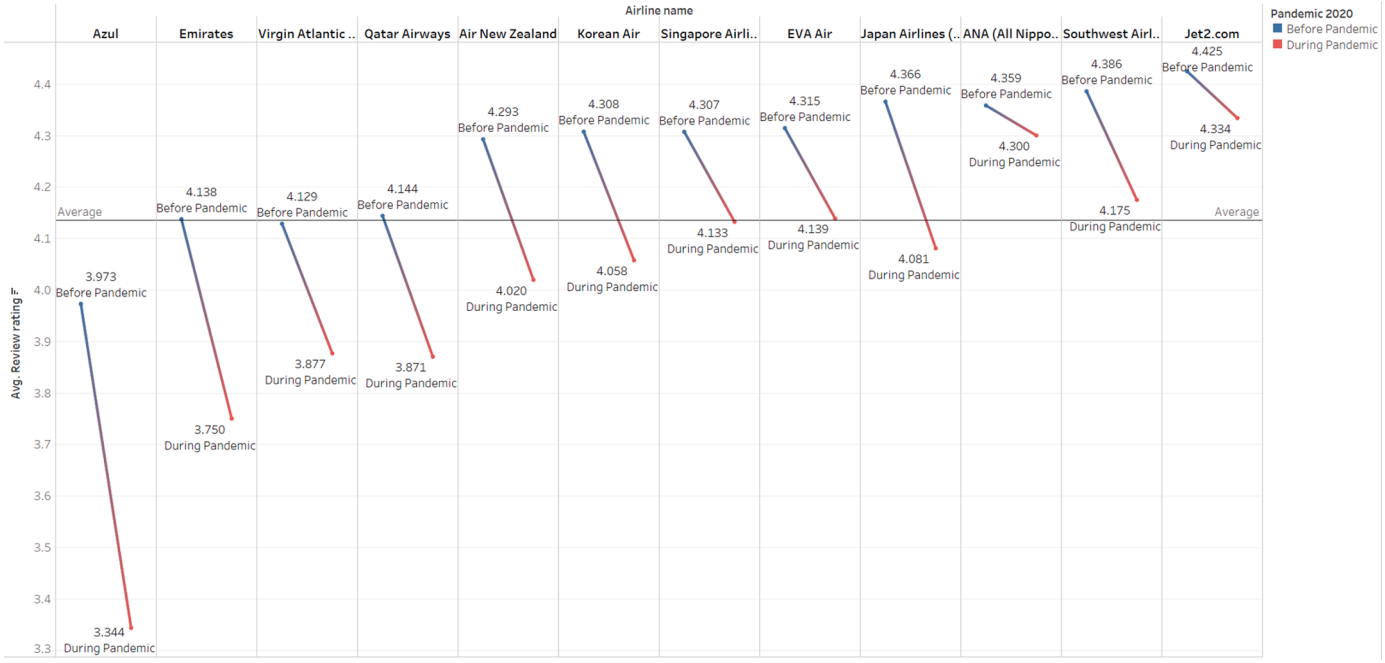


Fig. 2. Average review ratings for each airline before and during COVID-19 (January 1, 2020)

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_a}}\right)V \tag{1}$$

Output layer—Multi-task classification: Following the multi-head attention layer, the fully connected layers are adopted to reduce the output dimension to 24 gradually. Since we formulated ABSA as a multi-task classification problem, we reshaped a 24-dim 1D vector to a 2D matrix with dimensions of (3, 8), which indicates three classes and eight aspects for each aspect. It is noteworthy that two binary features are integrated into our model to consider the correlation between COVID-19 and sentiment prediction. The first feature is related to the pandemic time, and we examine whether the posting time of a review is during the COVID-19 period. The value of this feature is equal to 1 if the posting time is after January 1, 2020 (denoted as CVT); otherwise, it is 0. Next, since we considered that the refund rate would increase during the pandemic, the second feature was designed to recognize whether the review conveys a refund issue (denoted as CVR). The value of this feature is equal to 1 if the review text contains keywords "refund," "reimburse," and "cancellation;" otherwise, it is 0. Finally, we combined both COVID-19-related features into the fully connected layers.

As shown in Table 3, the airline review dataset is imbalanced. While handling an imbalanced dataset, loss calculation can be tricky. The most common approach to balancing the loss is assigning weights to the loss. The weights are calculated as the inverse of the number of class instances or the inverse of the square root of the number of class instances. This form of weighing scheme creates a problem by shifting focus entirely to the classes with very few instances. To handle this shifting focus, we adopted the class-balanced loss based on the effective number of samples [80] to reduce the impact of the imbalanced data. More specifically, we utilized the class-balanced softmax cross-entropy loss function for each aspect to normalize the relative loss across classes and reduce the drastic imbalance of weights by inverse class frequency. The final loss is the mean of all aspect losses. Given a model output y, a loss for this output with class i and its weight w can be calculated using Eq. (2), in which $\beta = \frac{\#sample-1}{\#sample}$ and ni is number of class i samples. Finally, the outputs are aspects with a sentiment that depicts 24 possible output states.

$$Loss(y,i) = -\frac{1-\beta}{1-\beta^{n_i}}\log\left(\frac{\exp(y_i)}{\sum_j \exp(y_j)}\right) \tag{2}$$

The LiFeBERT model was implemented using PyTorch[8], a Python deep learning library. We primarily followed the original settings of optimization and hyper-parameters for fine-tuning, i.e., five epochs of training time and the AdamW optimizer [81] with the learning rate set to 2e–5, but the β_2 set to 0.98 to improve stability during the training procedure. For the text representation, the pre-trained BERT model was trained with BooksCorpus (800M words) [82] and English Wikipedia (2,500M words). We loaded the weights of the pre-trained token embedding and positional embedding, and the transformer layers. For discriminative linguistic embeddings and classifier layers, their parameters were initiated by a normal distribution with a mean of 0 and standard deviation of 0.02. Also, these layers were first trained for ten epochs before unfreezing other pre-trained parameters. The maximum sequence length was 512 tokens, with padding or truncating at the end of the sequence. We ran the LiFeBERT model on a single Nvidia RTX 2080Ti, and seven sequences were trained per batch due to memory constraints.

4. Experiment results and discussion

In our experiments, the performance evaluation metrics included precision, recall, and F1-score. In general, there is a trade-off between precision and recall, and because these two metrics evaluate system performance from different perspectives, a single metric is essential to balance (average) the trade-off. That single metric is the F1-score, which is the harmonic mean of precision and recall. This score is generally close to the minimum of the two values and can thus be considered as an attempt to find the best possible compromise (balance) between precision and recall [83]. The F1-score is also deemed a conservative metric that prevents the possible overestimate of system performance because the harmonic mean is always less or equal to the arithmetic mean and geometric mean. For this reason, the F1-score is extensively used to

[8] https://pytorch.org/

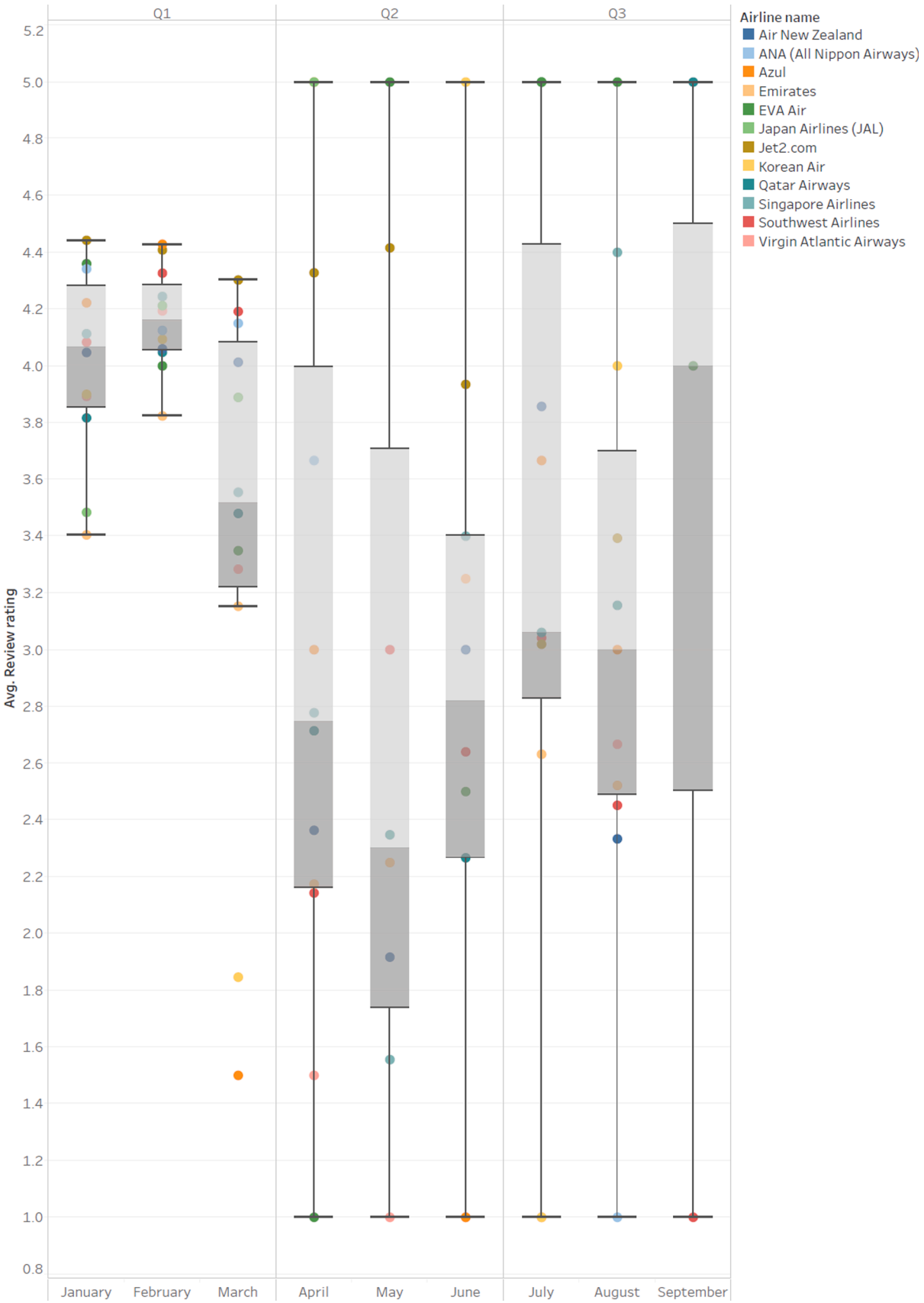


Fig. 3. Average review ratings for each month and airline during COVID-19

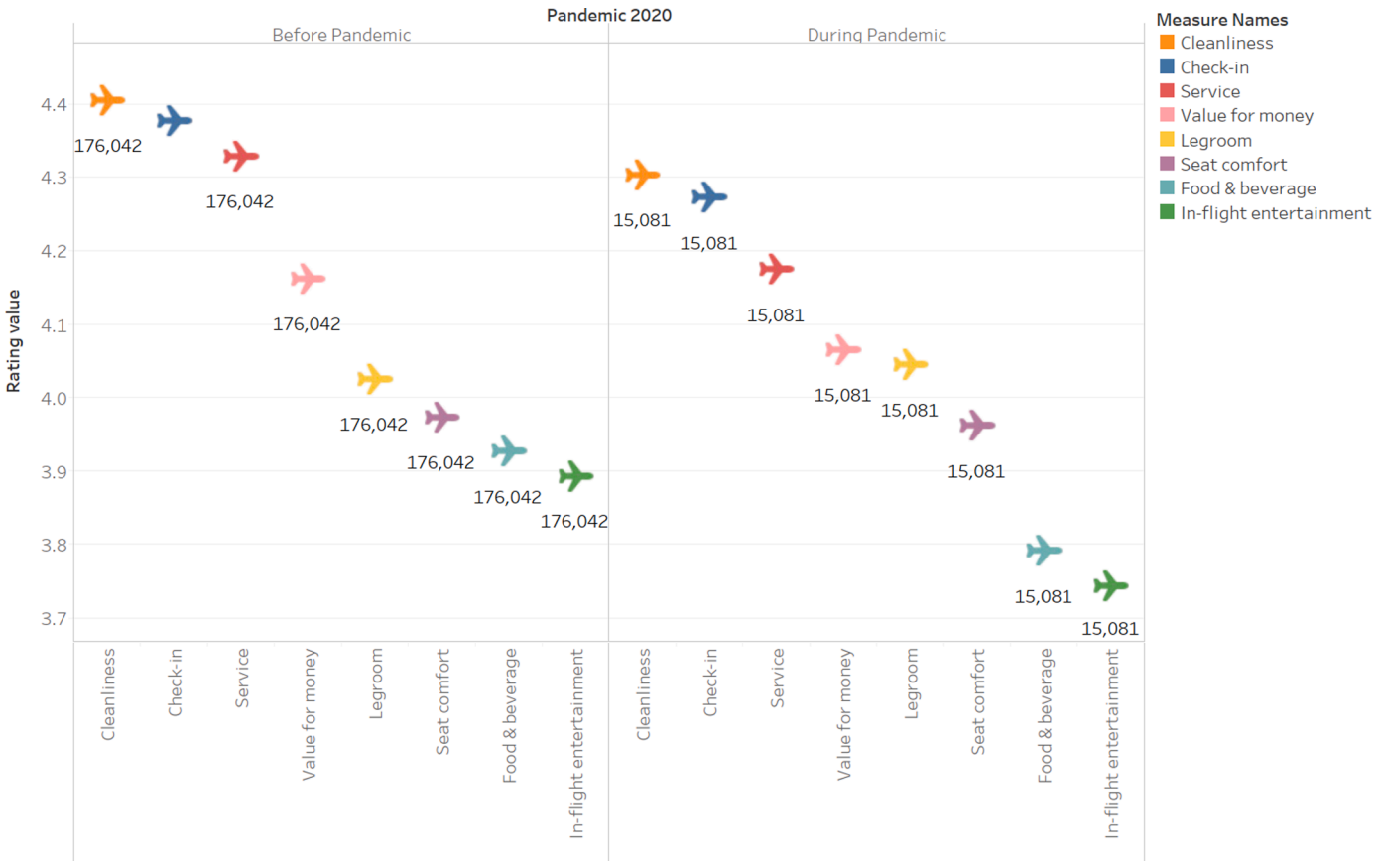


Fig. 4. Aspect ratings for all airlines—before and during COVID-19

Fig. 5. Service ratings of flight reviews during COVID-19

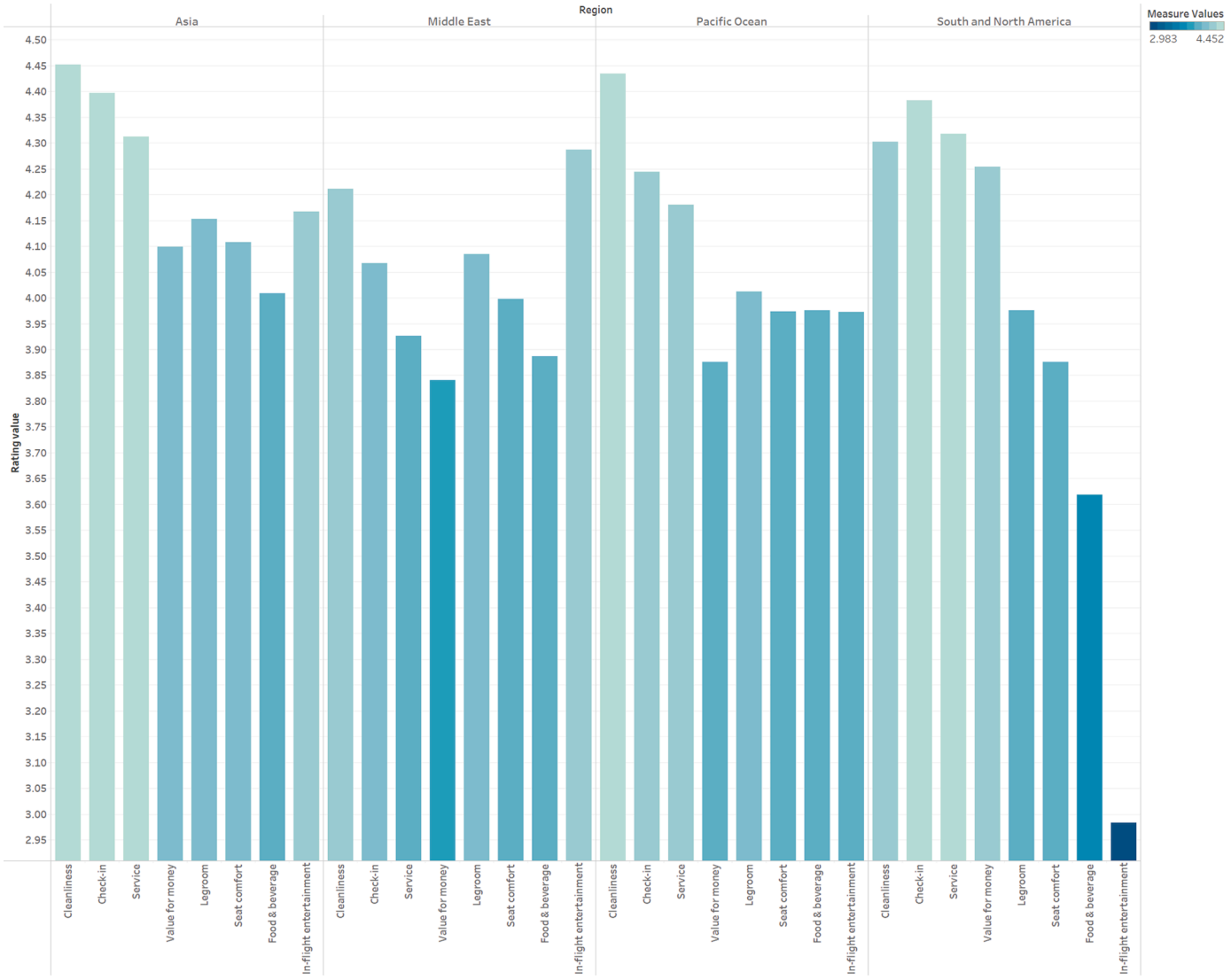


Fig. 6. Aspect ratings of airlines in Asia, the Middle East, Pacific Ocean, South America, and North America

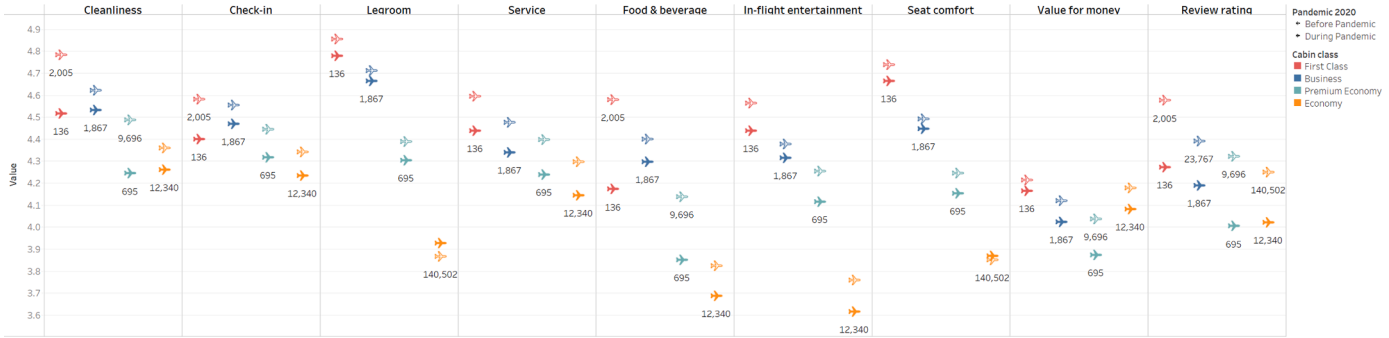


Fig. 7. Aspect ratings for the first-class, business, premium economy, and economy cabins for all airlines

judge the superiority of IS [78]. We also used the macro-average (Aμ) to compute the average performance, and to derive credible evaluation results, we used a ten-fold cross-validation approach [78]. We first investigated the effect of the COVID-19-related features that enhance the LiFeBERT on the eight aspects of LR, SC, FE, CS, VM, CN, CB, and food and beverage (FB).

Table 4 presents the performance of LiFeBERT and the results of incrementally applying both COVID-19-related features of CVT and CVR (denoted as +CVF). The results demonstrate that LiFeBERT effectively predicts airline ABSA to achieve an overall performance of 60% F1-scores across the eight aspects. This also shows that our LiFeBERT method successfully integrated token embeddings, positional embeddings, and discriminative linguistic embeddings, which were learned from raw texts to predict airline ABSA. It is evident from the table that our model can further improve the airline ABSA prediction performance by integrating the extra two COVID-19-related features. As the CVT and

Fig. 8. Aspect ratings of four cabins for EVA Air

Fig. 9. Aspect ratings of four cabins with pandemic keywords extracted

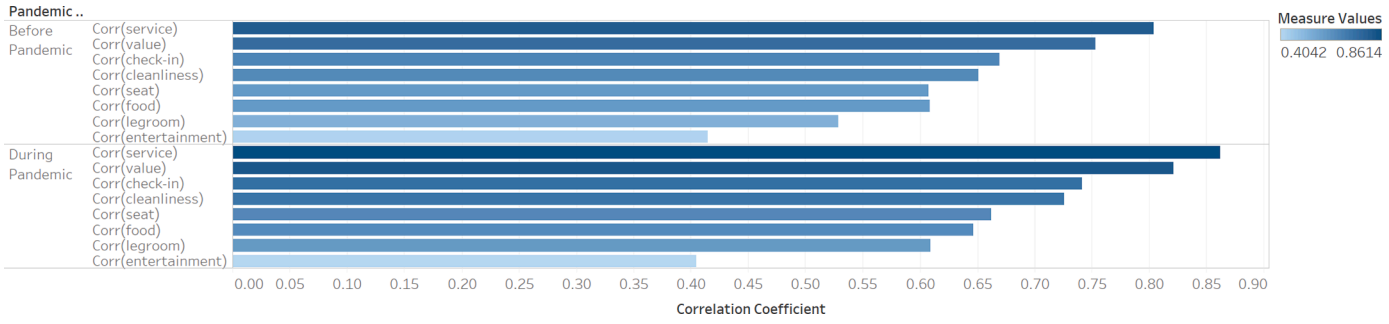


Fig. 10. Correlation between each aspect rating and overall satisfaction for all cabins

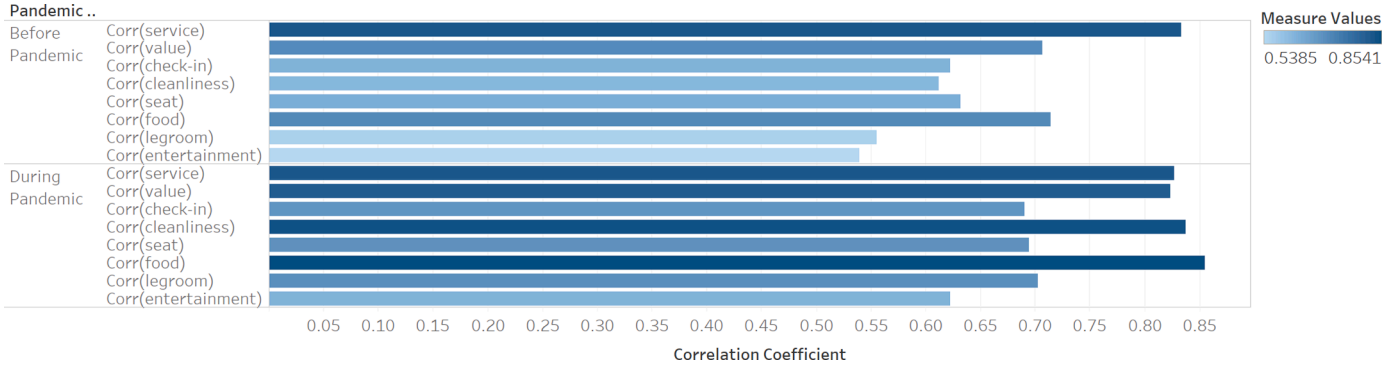


Fig. 11. Correlation between each aspect rating and overall satisfaction for first-class passengers

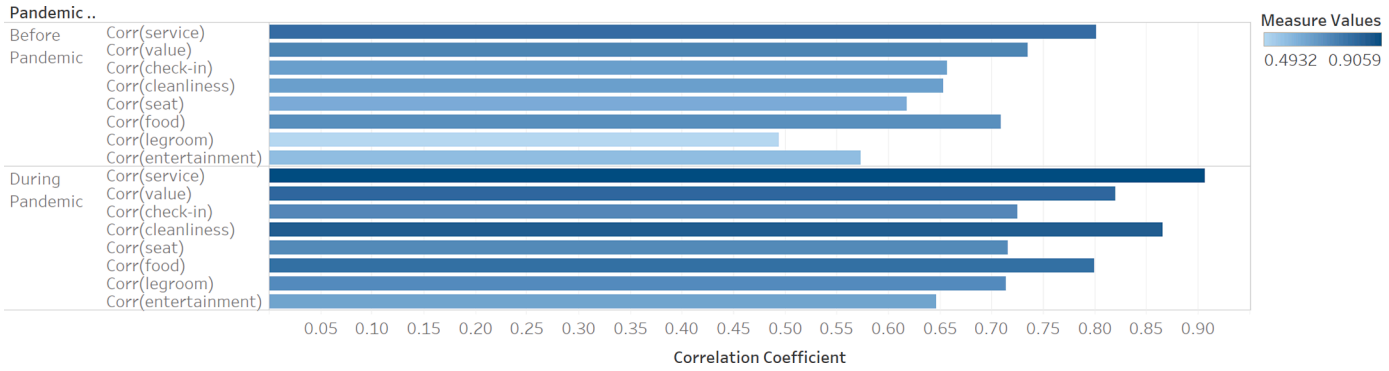


Fig. 12. Correlation between each aspect rating and overall satisfaction for all cabins of Korean Air

Table 3
Aspect ratings classified as negative, positive, and blank

Aspects/Polarity	Negative	Positive	None
Legroom	41,143/27.0%	109,509/71.8%	1,922/1.3%
Seat Comfort	41,886/27.5%	109,017/71.5%	1,671/1.1%
In-flight Entertainment	44,245/29.0%	93,024/61.0%	15,305/10.0%
Customer Service	24,753/16.2%	125,874/82.5%	1,947/1.3%
Value for Money	31,240/20.5%	116,713/76.5%	4,621/3.0%
Cleanliness	15,820/10.4%	106,232/69.6%	30,522/20.0%
Check-in and Boarding	17,618/11.5%	104,638/68.6%	30,318/19.9%
Food and Beverage	38,613/25.3%	80,718/52.9%	33,243/21.8%

CVR examine whether the airline reviews during COVID-19 relate to the refund issue, they do not conflict with the BERT model, which analyzes syntactic and semantic information in the texts. As a result, integrating CVT and CVR improves the system's performance, which indicates that airline ABSA prediction is highly associated with COVID-19 factors.

Machine learning classifiers can be classified into conventional, deep learning-based, and transformer-based learning [15]. To evaluate the proposed method, we select five popular, representative models, which have been used for sentiment analysis. Decision tree (denoted as DT), K-nearest neighbor (denoted as KNN), and random forest (denoted as RF) represent conventional models, TextCNN is a deep learning-based model, and BERT is a transformer-based model. Specifically, we compared LiFeBERT with DT [78] and KNN [78] to serve as a basis for comparisons. RF is an ensemble learning method for classification by constructing a multitude of DTs adopting the term frequency–inverse document frequency (TF–IDF) text representation. Next, we compared our method with TextCNN, a well-known CNN-based text classification approach [81]. Finally, we compared our approach with the state-of-the-art model, a BERT [52]. In order to compare the statistical significance of differences between the performances, we used McNemar's test [84] to examine whether the proposed method significantly improves the overall performance of the comparisons. The difference between LiFeBERT and the compared models was considered statistically significant if the p-value <0.001. The symbol "*" indicates that our

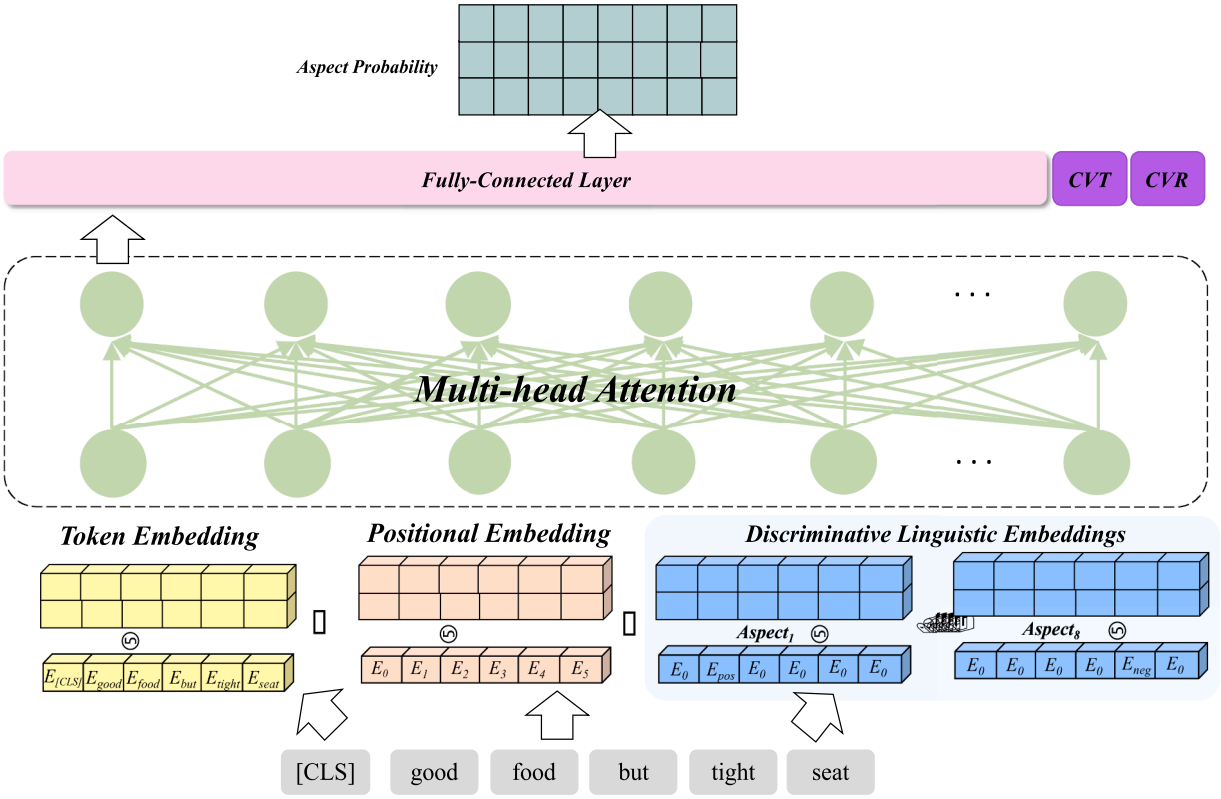


Fig. 13. Illustration of the proposed architecture for aspect-based sentiment prediction of airline reviews

Table 4
Incremental contribution of the COVID-19-related features with eight aspects

Aspect	Precision/Recall/F_1-score (%) LiFeBERT	+CVF
LR	58.9/52.6/54.9	**67.5/57.2/60.0**
SC	60.9/55.3/57.5	**69.6/60.8/63.2**
FE	60.4/54.9/56.4	**62.1/57.9/59.2**
CS	**72.8/56.3/56.8**	64.4/54.5/56.2
VM	66.4/53.1/54.8	**70.2/53.6/56.8**
CN	57.6/52.8/54.1	**69.2/54.2/57.1**
CB	58.7/53.5/55.0	**68.8/54.1/56.6**
FB	56.9/54.2/54.6	**66.0/55.1/56.5**
A^{μ}	61.6/54.1/55.5	**67.2/58.0/60.1**

method significantly outperforms the other systems.

Table 5 presents the results of our model LiFeBERT and the baseline methods for ABSA of airline reviews. First, the DT, a keyword statistics-based method, can only accomplish a mediocre performance with a 41.8% F1-score. The KNN calculates document similarity in the bag-of-words feature space with TF–IDF term weighting and achieves a similar result. Notably, the RF integrates multiple DTs through ensemble learning to optimize prediction results. However, it surprisingly has the worst performance. This may be because the multi-task classification problem is not a strength of the RF method, and the learned weighting is therefore unreliable. Interestingly, the TextCNN scored approximately 3% lower than both baseline methods and was the lowest among all neural models in this study.

Table 5
The aspect-based sentiment prediction results of the compared methods on the eight aspects and five merged aspects

Precision/Recall/F_1-score (%) *Aspect*	*DT*	*KNN*	*RF*	*TextCNN*	*BERT*	*LiFeBERT +CVF*
LR	40.5/	53.4/	46.9/	43.8/	56.8/	**67.5/57.2/**
	40.1/	39.3/	35.2/	39.7/39.8	49.0/	**60.0**
	40.3(*)	39.7(*)	31.9(*)	(*)	50.9(*)	
SC	41.8/	54.3/	48.5/	45.1/	59.0/	**69.6/60.8/**
	41.3/	40.3/	37.0/	41.1/41.6	51.5/	**63.2**
	41.5(*)	40.9(*)	35.2(*)	(*)	53.5(*)	
FE	44.8/	47.9/	51.4/	38.5/	58.2/	**62.1/57.9/**
	44.5/	42.4/	37.7/	39.0/36.9	57.3/	**59.2**
	44.6(*)	42.4(*)	34.4(*)	(*)	57.6(*)	
CS	45.8/	46.9/	57.2/	48.2/	57.0/	64.4/54.5/
	45.4/	45.8/	38.2/	42.6/44.1	**56.5**/	56.2
	45.6(*)	46.2(*)	38.9(*)	(*)	56.7(*)	
VM	42.6/	47.1/	52.6/	46.0/	59.4/	**70.2/53.6/**
	42.4/	43.0/	37.5/	41.0/41.7	52.4/	**56.8**
	42.5(*)	43.0(*)	36.7(*)	(*)	53.3(*)	
CN	39.8/	42.1/	48.5/	35.9/	57.9/	**69.2/54.2/**
	39.4/	39.8/	33.5/	37.8/34.8	51.6/	**57.1**
	39.6(*)	38.3(*)	27.6(*)	(*)	53.8(*)	
CB	40.3/	43.3/	62.7/	36.3/	54.6/	**68.8**/54.1/
	40.0/	43.3/	34.4/	37.6/34.5	**55.0**/	**56.6**
	40.2(*)	40.5(*)	29.3(*)	(*)	54.8(*)	
FB	39.9/	42.7/	51.6/	35.8/	54.9/	**66.0**/55.1/
	39.9/	41.9/	39.0/	41.7/37.2	**55.7**/	**56.5**
	39.9(*)	39.6(*)	33.8(*)	(*)	53.1(*)	
A_8^{μ}	41.9/	47.2/	52.4/	41.2/	57.2/	**67.2/58.0/**
	41.6/	42.0/	36.6/	40.1/38.8	53.6/	**60.1**
	41.8(*)	41.3(*)	33.5(*)	(*)	54.2(*)	
SQ	42.5/	54.7/	49.2/	59.3/	58.6/	**66.0/72.6/**
	41.6/	39.8/	34.6/	44.6/50.9	67.5/	**64.5**
	42.0(*)	46.1(*)	40.6(*)	(*)	61.0(*)	
SV	45.9/	47.0/	58.3/	57.1/	55.1/	**63.8/61.3/**
	45.6/	46.1/	37.2/	45.8/50.8	58.8/	**62.5**
	45.7(*)	46.5(*)	45.3(*)	(*)	56.7(*)	
VM	42.6/	47.1/	52.5/	55.8/	67.9/	**70.3/59.7/**
	42.5/	43.0/	37.4/	44.2/49.3	53.3/	**63.5**
	42.5(*)	44.9(*)	43.7(*)	(*)	53.6(*)	
CN	39.9/	42.1/	55.9/	54.7/	59.9/	**69.0**/56.5/
	39.6/	39.8/	33.5/	39.0/45.5	**56.7**/	**60.1**
	39.7(*)	40.9(*)	41.9(*)	(*)	54.2(*)	
FE	41.5/	47.8/	45.6/	57.4/	59.7/	**66.0/59.0/**
	41.1/	39.8/	35.4/	43.6/49.6	54.0/	**61.2**
	41.3(*)	43.4(*)	39.9(*)	(*)	55.2(*)	
A_5^{μ}	42.5/	47.7/	52.3/	56.9/	60.2/	**67.0/61.8/**
	42.1/	41.7/	35.6/	43.4/49.2	58.1/	**62.3**
	42.3(*)	44.5(*)	42.4(*)	(*)	56.2(*)	

However, BERT can further improve performance to reach a 54.2% F1-score. This indicates that the BERT model is efficient in representing textual information and learning the context of airline reviews for multi-task classification. It is worth mentioning that our LiFeBERT method, combined with multiple text representations and latent linguistic features, can be learned from airline reviews through additional embeddings. The token and positional embeddings are adopted for learning the syntactic and context relations. We employed the discriminative linguistic embeddings to encode the characteristics of linguistic features for identifying aspects and sentiments that are hidden within airline reviews. Consequently, our LiFeBERT outperforms the comparisons and achieves the best overall precision, recall, and F1-score.

There are different dimensions of service, which can be classified into in-flight and ground [85] and tangible and intangible [86] services. Current research has shown that the appropriate combination of categories can enhance system performance when a large training dataset is not readily available [87]. Based on our observations, we further merged the eight aspects into five. Specifically, we merged “Legroom” and SC as Seat Quality (SQ), “Check-in & Boarding” and CS as Service (SV), and “In-flight Entertainment” and “Food & Beverage” as Food and Entertainment (FE). A comprehensive evaluation was then conducted to examine the performance of our models (see Table 5). We noticed that these combinations could boost model performance, typically for the recall and F1-score. The rationale of these merging also corresponds to the aspect-rating visualization in Fig. 4. For example, in-flight service taken to comprise both FB and FE shows the lowest rating value (below 4) among eight aspects. Further, the ratings of LR and SC are close to each other (around 4), while CB, and CS are close to each other with higher rating values, above 4.3.

To obtain clearer insights into the reviews, we used word clouds to visualize the learned positive and negative keywords for each aspect and color-coded them for clarity. Each unique color represents one aspect: red: LR; orange: SC; green: FE; light blue: VM; dark blue: CS; purple: CN; pink: CB; grey: FB. The top 10 keywords of each aspect were selected for both sentiments to generate the word cloud. Its LLR value decides the size of a word in the word cloud. Therefore, we can quickly identify features within each group, and specific sentiments can be associated with their descriptions. As shown in Fig. 14, words can affect the polarity of the reviews. For the positive word cloud, the term “excellent” appears in most aspects, so that, any review containing this word will likely have a high rating. Customers who fly with Emirates or have a seat in economy class with fine food are satisfied with the FE aspect. Also, they will often leave comments containing the word “friendly” if they are happy with the CS.

In contrast, when customers leave comments with emphasized words like “worse,” “poor,” or “terrible,” the described airline aspects are generally unsatisfactory to the customer. The most common word is “uncomfortable,” which means that reviews containing this word are primarily negative in all aspects. Most words in the negative word cloud are strong emotional words, such as “worst,” “awful,” and “horrible.” It is interesting to note that when customers are not satisfied with the LR, they also do not leave a positive view regarding the SC aspect. As for the FE aspect, the word “rude” has the second highest LLR score, which shows that when customers are frustrated with some aspects, they are very likely to evaluate the other aspects negatively.

Our proposed model outperforms the baseline comparisons, which relates to our use of the deep learning model to train the flight reviews with known aspect ratings to predict the reviews without aspect ratings. Among the collected data, a total of 152,574 flight reviews had aspect ratings, compared with 38,312 flight reviews without them. Our prediction model was designed to generate positive, negative, and no

Fig. 14. Word clouds from reviews with positive (left-hand side) and negative (right-hand side) sentiments.

sentiment on each aspect, and we developed a dashboard to visualize our prediction results for each aspect, as shown in Fig. 15. We filtered out the aspects with no sentiments and separated the prediction results by the ratio of negative to positive reviews. In the figure, the colors indicate the prediction results from January 2016 to August 2020, while the number of reviews is shown next to each circle. The overall results reveal that the percentage of negative reviews had risen in recent years for all aspects, but this was especially the case in 2020. These findings are consistent with the data visualizations of aspect ratings presented in the previous section.

5. Conclusions and implications

The unprecedented nature and scale of COVID-19 pandemic has crippled the worldwide economy and the airline industry in particular. Airline service and traveler satisfaction have been precarious during the pandemic. Sezgen et al. [14] point out that few studies have used online reviews to identify critical elements of airline services by conducting sentiment analysis. Online reviews provide an alternative source for firms to understand better consumer perspectives on their products and services, such as the process for refunding flight tickets. This study sheds light on the use of deep learning-based NLP and word embedding techniques to conduct ABSA and use visual analytics to better understand customers' satisfaction before and during the pandemic. This research thus provides both theoretical and managerial implications.

5.1. Theoretical implications

Our visualization findings confirm expectancy disconfirmation theory [14]. That is, customers have expectations about services prior to their purchase and have varying expectations regarding different cabins and ticket prices. During the pandemic, the CB process can be longer, the FB service is restricted, and the ticket refund process can be frustrating and tedious. As a result, customers naturally compare the outcomes or perceptions to their prior expectations, which leads to satisfaction or dissatisfaction with each service aspect. Positive and negative sentiments are often evoked in this evaluation process.

Travelers reacted differently to airline services before and during the

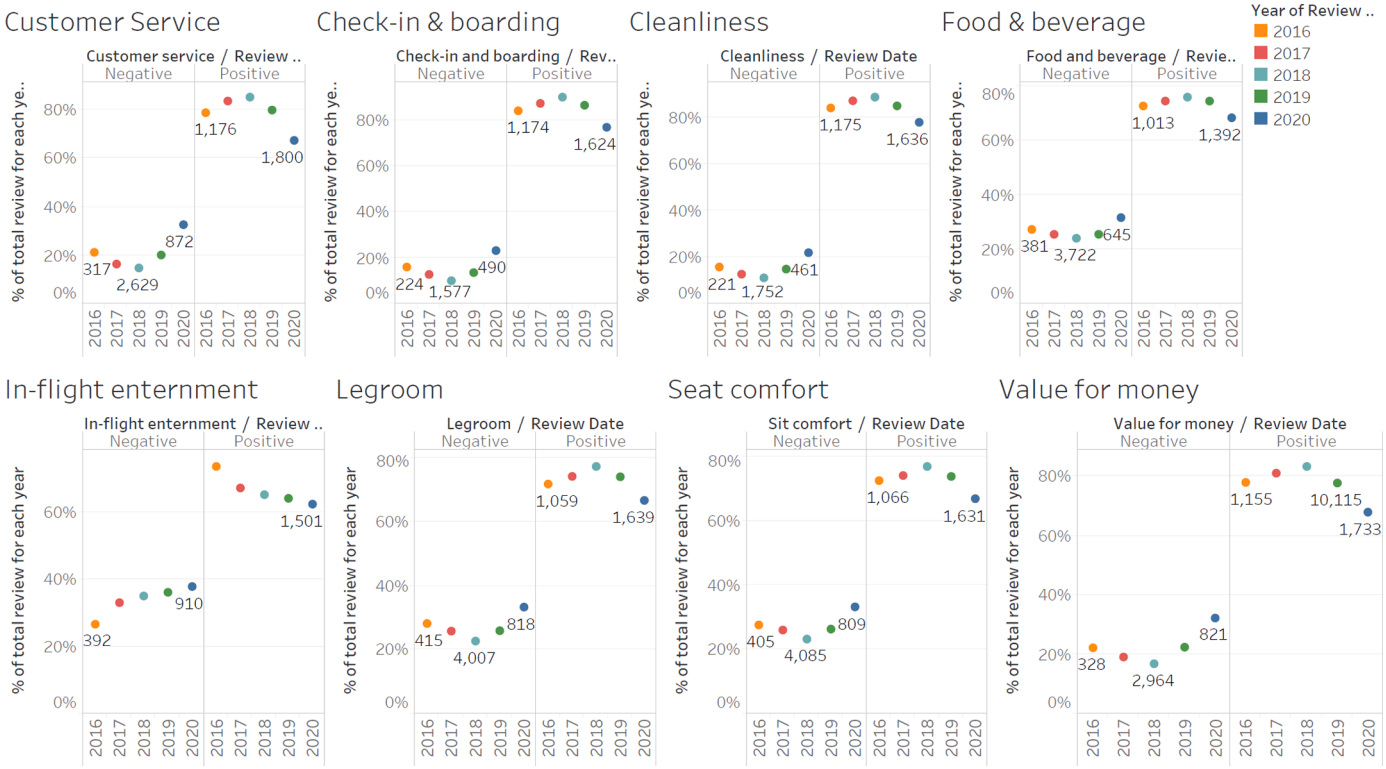


Fig. 15. Aspect rating predictions based on our deep learning model

pandemic, which is consistent with appraisal theory [16]. We further noticed that flight reviews related to refunds or cancellations during the pandemic are strong indicators of customer dissatisfaction. When the service performance is higher or lower than expected, a respective positive or negative disconfirmation may occur, as described by disconfirmation theory [88].

This study also extends knowledge about the use of deep learning techniques by exploring latent linguistic features and various service aspects to learn and predict aspect-level sentiments in the context of the airline industry and the COVID-19 pandemic. ABSA techniques can be classified into lexical-based, machine learning, and hybrid approaches [18,20]. Moreover, a domain-specific lexicon has rarely been adapted for the tourism and hospitality context, while a general lexicon is limited for this context [18]. Finally, traditional machine learning approaches cannot effortlessly capture semantic relations and implicit meanings [50], which is why Tubishat et al. [20] pointed out that implicit aspect extraction in sentiment analysis is still a challenge.

Recently, word embedding techniques have been gaining increased attention because they enable researchers to 1) identify words with similar meanings used in a particular context [89] and 2) overcome the drawback of the bag-of-words analysis [51]. Pre-trained embedding techniques, such as Word2Vec [53], GloVe [54], and FastText [90], are now actively used as new tasks for text mining, NLP, and deep learning. However, the use of word embedding techniques in management and hospitality studies remains rare [89], which is why Alaei et al. [18] encouraged future tourism research to discover the dynamics of data and gain deeper insights from different aspects of data using deep learning approaches.

In this study, the proposed model driven by deep learning and embedding techniques bridges the abovementioned gaps in the literature by including pandemic factors to detect traveler sentiments. Our data-driven framework integrates token embedding, positional bedding, discriminative logistic embedding, and BERT, which enables us to capture syntactic, semantic, and implied information between words and sentences. Compared with established and widely used models, such as DT, CNN, and KNN, the proposed deep learning model offers distinctive advantages and thus outperforms baseline comparisons.

The experimental results suggest that including the pandemic-related attributes to the deep learning model and combining similar aspect ratings can improve the model performance. More importantly, our proposed model can predict unrated aspects in airline reviews, which can identify and understand travelers' emotional reactions. This is evident in the consistent findings from our model evaluations and visual analytics. To our best knowledge, we are the first to use deep learning-based NLP, multiple word embeddings, and visual analytics techniques to conduct ABSA and focus on airline and pandemic context.

5.2. Managerial implications

This study provides important implications for managerial practice. Our visualization and deep learning findings confirm that the average review and aspect ratings in 2020 were lower than in previous years (before the pandemic). This was especially noticeable during the WHO's announcement of the global pandemic, which has led to a decline in airline service quality for most aspects as perceived by passengers. Therefore, from a managerial perspective, the strategies to improve service quality and passenger satisfaction should be different during the pandemic.

First, airline practitioners should handle passenger reviews during the pandemic differently. Our interactive visualizations reveal an apparent decline in aspect ratings and customer satisfaction during the outbreak of COVID-19. Our model performance improved by adding the pandemic factors, which provides further evidence that the driving factors of passenger satisfaction are different during the pandemic. Airlines, for example, can differentiate their service by responding to the pandemic more quickly and automatically issuing ticket refunds, flight certificates, and flight credits, which may include seat upgrades, flight tickets, travel insurance, and luggage refunds. This is also a good opportunity for airlines to review their refund policy and procedures to significantly reduce the number of customer complaints and long waits in CS phone calls. An automated, intelligent system should cope with these mundane processes more efficiently and more conveniently. A positive airline image will enhance its reputation and positively affect customer loyalty and trust [91].

Second, there are a variety of driving factors of passenger satisfaction for different cabins. Managers can prioritize and optimally allocate resources based on each cabin class, and each passenger's socio-economic and cultural information, which can be collected from flight booking and pre-boarding processes. Construal level theory describes how people process information and react to, and interpret the same event differently [92]. Chatterjee and Mandal [93], who analyzed 28,341 reviews for 345 airlines, found that drivers of ratings and satisfaction differ according to traveler's choice of cabin class and traveler goals, such as business and leisure. For example, business travelers tend to care more about business success than the travel itself (high construal level), while leisure travelers are more psychologically invested in the actual travel experience (lower construal level) [93]. Customers in a higher construal level thus tend to be more generous and rate higher in service evaluations [94]. Our visual analytics support construal level theory and previous findings because first-class and business passengers tend to assign a higher rating to most aspects and have higher satisfaction for all airlines and each airline. However, we found changes in traveler rating behavior during the pandemic, especially for first-class, business, and premium economy passengers. For example, first-class traveler ratings on FB drop significantly (lower than the ratings of business travelers) during the pandemic. We also noticed that travelers of each airline exhibit different rating patterns according to the region of the airline, and whether the airlines were low-cost or full-service, and whether the flights were long-distance or domestic. There are a number of dimensions of service that may impact a traveler's post-purchase experience, and such analysis can be used for customer segmentation and service quality improvements.

Although airlines strive for short-term survival and may deem short-term support wasteful, a long-term sustainability strategy is needed based on stakeholder engagement [4]. Another interesting finding is that the airline response rate (<0.02%) from the selected 12 airlines is significantly lower than the hotel response rate (< 70%) [47] on TripAdvisor. This suggests that airlines may need to increase their efforts to cultivate relationships with customers through their social media platforms, which is supported by the recent findings by Tubishat et al. [20].

Considering the intense competition between airlines, our predictive model and visual analytics provide valuable insights about the aspects that stimulate customer satisfaction and dissatisfaction across multiple dimensions, such as cabin classes, regions of airlines, and time. Furthermore, from a managerial perspective, the ability to differentiate service quality aspects enables an airline to attain a distinctive advantage and influence customer repurchase intention, which are key drivers of revenue growth and profitability [95]. Thus, our proposed approach, which integrates machine power and human intelligence, enables managers to monitor passenger sentiment at low cost and in real-time, anticipate their satisfaction or dissatisfaction, and act accordingly by adjusting their services in a timely manner.

5.3. Limitation and future research

While our study has demonstrated an innovative way to detect and predict user satisfaction through aspect ratings, review content, and COVID-associated factors, it also has several limitations. First, the study can be extended to include more airlines across more countries. Asian airlines, such as ANA, EVA, and Korean Air, tend to have fewer online reviews on TripAdvisor since travelers may feel more comfortable writing reviews in their own language, and TripAdvisor is not yet the

primary travel platform in Asia. Therefore, a multi-language analysis can be conducted in later studies. Future research can also collect data from multiple platforms, like Skytrax [91], though the data attributes and aspects can differ on different platforms, which would require additional effort to detect and extract common aspect information.

While the service level in our study is proxied by cabin classes, additional information about travelers, such as demographics, review contributions, etc., may enhance the performance of the predictive model. Cross-cultural studies of airline reviews have remained limited and rely heavily on survey methods [93]. Future studies can collect reviewer profiles on TripAdvisor to conduct cross-cultural analyses. By doing this, the managers can better understand traveler requirements and the factors driving their satisfaction based on different cultures, which would be valuable information for marketing and customer segmentation. A further investigation lens could be extended to different flight routes as the distance of a trip and its route is likely to influence traveler satisfaction in different ways. An airline may improve its service based on specific routes.

Finally, our experimental results reveal that COVID-19 and ticket refund factors can be used to improve the performance of machine learning models for recent airline research. Future studies may conduct discourse or deeper language analysis to gain a more profound insight into the socio-psychological characteristics in written reviews, which may advance the algorithm to the next level.

Funding

This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 107-2410-H-038 -017 -MY3, MOST 107-2634-F-001-005, and MOST 109-2410-H-038 -012 -MY2. The Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan also provided financial support for our work.

Author Statement

Chih-Hao Ku and **Yung-Chun Chang** jointly work on the research conceptualization, methodology, data curation, data analysis, investigation, information visualization, experiments, paper review, revision, and editing. **Yung-Chun Chang** acquires research funds to support this research. **Duy-Duc Le Nguyen** works on model development and experiments under supervision.

Appendix

In this research, we used the following equation to calculate LLR value to extract polarity keywords denoting sentiments of an aspect. Let w indicate a word and S denote a sentiment of a certain aspect. $N(S)$ and $N(S)$ are the numbers of reviews that respectively contain this sentiment or not. $N(wS)$, which is shortened as k, is the number of reviews containing w and S simultaneously, while $N(wS)$ is the number of reviews without this sentiment but including w, denoted as l. To further simplify the formula, we also define $m = N(S) - k$ as the number of reviews containing S without the word w, and $n = N(S) - l$ as those with neither S nor w. A maximum likelihood estimation is conducted to obtain probabilities $p(w)$, $p(w|S)$, and $p(w|S)$ through calculating the log-likelihood of the hypothesis that the presence of w in set S is beyond chance. A word with a large LLR value is closely associated with the sentiment of an aspect. Finally, the entire set of words are ranked by their LLR value in the training data, and the top 150 are selected as polarity keywords for each aspect.

$$LLR(t, S) = 2\log\frac{p(w|S)^k(1-p(w|S))^m p(w|S)^l(1-p(w|S))^n}{p(t)^{k+l}(1-p(w))^{m+n}} \quad (3)$$

References

[1] H. Nakamura, S. Managi, Airport Risk of Importation and Exportation of the COVID-19 Pandemic, Transport Policy 96 (2020) 40–47, https://doi.org/10.1016/j.tranpol.2020.06.018.
[2] T. Powley, J. Reed, K. Inagaki, P. Riordan, M. Badkar, J. Smyth, L. Lewis, N. Rovnick, Airlines slash flights to cut costs as coronavirus hits travel demand, (2020). https://www.ft.com/content/c28b5790-62c6-11ea-a6cd-df28cc3c6a68 (accessed September 13, 2020).
[3] S.W. Anderson, L.S. Baggett, S.K. Widener, The Impact of Service Operations Failures on Customer Satisfaction: Evidence on How Failures and Their Source Affect What Matters to Customers, M&SOM. 11 (2009) 52–69. 10.1287/msom.1070.0193.
[4] S.-S. Pere, V.-D. Augusto, C.-E. Natàlia, An Early Assessment of the Impact of COVID-19 on Air Transport: Just Another Crisis or the End of Aviation as We Know It? J Transp Geogr 86 (2020), 102749 https://doi.org/10.1016/j.jtrangeo.2020.102749.
[5] R. Filieri, What Makes an Online Consumer Review Trustworthy? Annals of Tourism Research 58 (2016) 46–64, https://doi.org/10.1016/j.annals.2015.12.019.
[6] I. Pentina, A.A. Bailey, L. Zhang, Exploring Effects of Source Similarity, Message Valence, and Receiver Regulatory Focus on Yelp Review Persuasiveness and Purchase Intentions, Journal of Marketing Communications 24 (2018) 125–145, https://doi.org/10.1080/13527266.2015.1005115.
[7] Z. Yan, M. Xing, D. Zhang, B. Ma, EXPRS: An Extended Pagerank Method for Product Feature Extraction from Online Consumer Reviews, Information & Management 52 (2015) 850–858, https://doi.org/10.1016/j.im.2015.02.002.
[8] N. Korfiatis, P. Stamolampros, P. Kourouthanassis, V. Sagiadinos, Measuring Service Quality from Unstructured Data: A Topic Modeling Application on Airline Passengers' Online Reviews, Expert Systems with Applications. 116 (2019) 472–486. 10.1016/j.eswa.2018.09.037.
[9] A.J.D.V.T. Melo, R.M. Hernández-Maestro, P.A. Muñoz-Gallego, Service Quality Perceptions, Online Visibility, and Business Performance in Rural Lodging Establishments, Journal of Travel Research 56 (2017) 250–262, https://doi.org/10.1177/0047287516635822.
[10] H. Gupta, Evaluating Service Quality of Airline Industry Using Hybrid Best Worst Method and VIKOR, Journal of Air Transport Management 68 (2018) 35–47, https://doi.org/10.1016/j.jairtraman.2017.06.001.
[11] T.A.- Quintana, S.M.- Gil, P.P.- Peral, How Could Traditional Travel Agencies Improve Their Competitiveness and Survive?, in: A Qualitative Study in Spain, Tourism Management Perspectives, 20, 2016, pp. 98–108, https://doi.org/10.1016/j.tmp.2016.07.011.
[12] R. Rajaguru, Role of Value for Money and Service Quality on Behavioural Intention: A Study of Full Service and Low Cost Airlines, Journal of Air Transport Management 53 (2016) 114–122, https://doi.org/10.1016/j.jairtraman.2016.02.008.
[13] D. Kang, Y. Park, Review-Based Measurement of Customer Satisfaction in Mobile Service: Sentiment Analysis and Vikor Approach, Expert Systems with Applications. 41 (2014) 1041–1050. 10.1016/j.eswa.2013.07.101.
[14] E. Sezgen, K.J. Mason, R. Mayer, Voice of Airline Passenger: A Text Mining Approach to Understand Customer Satisfaction, Journal of Air Transport Management 77 (2019) 65–74, https://doi.org/10.1016/j.jairtraman.2019.04.001.
[15] N.C. Dang, M.N. Moreno-García, F. De la Prieta, Sentiment Analysis Based on Deep Learning, A Comparative Study, Electronics 9 (2020) 483, https://doi.org/10.3390/electronics9030483.
[16] X. Xu, W. Liu, D. Gursoy, The Impacts of Service Failure and Recovery Efforts on Airline Customers' Emotions and Satisfaction, Journal of Travel Research 58 (2019) 1034–1051, https://doi.org/10.1177/0047287518789285.
[17] M. Afzaal, M. Usman, A. Fong, Predictive Aspect-Based Sentiment Classification of Online Tourist Reviews, Journal of Information Science 45 (2019) 341–363, https://doi.org/10.1177/0165551518789872.
[18] A.R. Alaei, S. Becken, B. Stantic, Sentiment Analysis in Tourism: Capitalizing on Big Data, Journal of Travel Research 58 (2019) 175–191, https://doi.org/10.1177/0047287517747753.
[19] M. Siering, A.V. Deokar, C. Janze, Disentangling Consumer Recommendations: Explaining and Predicting Airline Recommendations Based on Online Reviews,

Decision Support Systems 107 (2018) 52–63, https://doi.org/10.1016/j.dss.2018.01.002.
[20] M. Tubishat, N. Idris, M.A.M. Abushariah, Implicit Aspect Extraction in Sentiment Analysis: Review, Taxonomy, Oppportunities, and Open Challenges, Information Processing & Management 54 (2018) 545–563, https://doi.org/10.1016/j.ipm.2018.03.008.
[21] Dr.S. Chatterjee, Explaining Customer Ratings and Recommendations by Combining Qualitative and Quantitative User Generated Contents, Decision Support Systems 119 (2019) 14–22, https://doi.org/10.1016/j.dss.2019.02.008.
[22] S.V. Gudmundsson, M. Cattaneo, R. Redondi, Forecasting Recovery Time in Air Transport Markets in the Presence of Large Economic Shocks: COVID-19, Social Science Research Network, Rochester (2020), https://doi.org/10.2139/ssrn.3623040. NY.
[23] L. Shen, United Airlines Stock Drops $1.4 Billion After Passenger-Removal Controversy, Fortune (2017). https://fortune.com/2017/04/11/united-airlines-stock-drop/(accessed. September 13, 2020.
[24] R. Khan, S. Urolagin, Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data, International Journal of Advanced Computer Science and Applications (IJACSA) 9 (2018), https://doi.org/10.14569/IJACSA.2018.090652.
[25] Y.-C. Ou, P.C. Verhoef, The Impact of Positive and Negative Emotions on Loyalty Intentions and Their Interactions with Customer Equity Drivers, Journal of Business Research 80 (2017) 106–115, https://doi.org/10.1016/j.jbusres.2017.07.011.
[26] R. Raine Cai, L. Lu, D. Gursoy, Effect of Disruptive Customer Behaviors on Others' Overall Service Experience: An Appraisal Theory Perspective, Tourism Management 69 (2018) 330–344, https://doi.org/10.1016/j.tourman.2018.06.013.
[27] N. Elkhani, S. Soltani, M.H.M. Jamshidi, Examining a Hybrid Model for E-Satisfaction and E-Loyalty to E-Ticketing on Airline Websites, Journal of Air Transport Management 37 (2014) 36–44, https://doi.org/10.1016/j.jairtraman.2014.01.006.
[28] T. Radojevic, N. Stanisic, N. Stanic, Inside the Rating Scores: A Multilevel Analysis of the Factors Influencing Customer Satisfaction in the Hotel Industry, Cornell Hospitality Quarterly (2017), https://doi.org/10.1177/1938965516686114.
[29] G. Viglia, R. Minazzi, D. Buhalis, The Influence of E-Word-of-Mouth on Hotel Occupancy Rate, International Journal of Contemporary Hospitality Management 28 (2016) 2035–2051, https://doi.org/10.1108/IJCHM-05-2015-0238.
[30] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment Analysis Is a Big Suitcase, IEEE Intelligent Systems 32 (2017) 74–80, https://doi.org/10.1109/MIS.2017.4531228.
[31] M.L. Yadav, B. Roychoudhury, Effect of Trip Mode on Opinion About Hotel Aspects: A Social Media Analysis Approach, International Journal of Hospitality Management 80 (2019) 155–165, https://doi.org/10.1016/j.ijhm.2019.02.002.
[32] S.M. Mohammad, S. Kiritchenko, X. Zhu, NRC-Canada, Building the State-of-the-Art in Sentiment Analysis of Tweets (2013) 321–327. ArXiv:1308.6242 [Cs].
[33] Y.-C. Chang, C.-H. Ku, C.-H. Chen, Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor, International Journal of Information Management 48 (2019) 263–279, https://doi.org/10.1016/j.ijinfomgt.2017.11.001.
[34] S. Chatterjee, Drivers of Helpfulness of Online Hotel Reviews: A Sentiment and Emotion Mining Approach, International Journal of Hospitality Management 85 (2020), 102356, https://doi.org/10.1016/j.ijhm.2019.102356.
[35] M. Nakayama, Y. Wan, The Cultural Impact on Social Commerce: A Sentiment Analysis on Yelp Ethnic Restaurant Reviews, Information & Management. 56 (2019) 271–279, https://doi.org/10.1016/j.im.2018.09.004.
[36] H.Q. Vu, G. Li, R. Law, Y. Zhang, Exploring Tourist Dining Preferences Based on Restaurant Reviews, Journal of Travel Research 58 (2019) 149–167, https://doi.org/10.1177/0047287517744672.
[37] Y.-H. Hu, Y.-L. Chen, H.-L. Chou, Opinion Mining from Online Hotel Reviews – a Text Summarization Approach, Information Processing & Management 53 (2017) 436–449, https://doi.org/10.1016/j.ipm.2016.12.002.
[38] K. Schouten, F. Frasincar, Survey on Aspect-Level Sentiment Analysis, IEEE Transactions on Knowledge and Data Engineering 28 (2016) 813–830, https://doi.org/10.1109/TKDE.2015.2485209.
[39] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S.M. Jiménez-Zafra, G. Eryiğit, SemEval-2016 Task 5: Aspect Based Sentiment Analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 19–30, https://doi.org/10.18653/v1/S16-1002.
[40] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 Task 4: Aspect Based Sentiment Analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland (2014) 27–35, https://doi.org/10.3115/v1/S14-2004.
[41] R. Sann, P.-C. Lai, Understanding Homophily of Service Failure Within the Hotel Guest Cycle: Applying Nlp-Aspect-Based Sentiment Analysis to the Hospitality Industry, International Journal of Hospitality Management 91 (2020), 102678, https://doi.org/10.1016/j.ijhm.2020.102678.
[42] N.E. Evangelopoulos, Latent Semantic Analysis, WIREs Cognitive Science 4 (2013) 683–692, https://doi.org/10.1002/wcs.1254.
[43] J. Evermann, J.-R. Rehse, P. Fettke, Predicting Process Behaviour Using Deep Learning, Decision Support Systems 100 (2017) 129–140, https://doi.org/10.1016/j.dss.2017.04.003.
[44] X. Liu, B. Zhang, A. Susarlia, R. Padman, Go to You Tube and Call Me in the Morning: Use of Social Media for Chronic Conditions, MISQ 44 (2020) 257–283, https://doi.org/10.25300/MISQ/2020/15107.
[45] S. Zhou, Z. Qiao, Q. Du, G.A. Wang, W. Fan, X. Yan, Measuring Customer Agility from Online Reviews Using Big Data Text Analytics, Journal of Management Information Systems 35 (2018) 510–539, https://doi.org/10.1080/07421222.2018.1451956.
[46] O. Araque, I. Corcuera-Platas, J.F. Sánchez-Rada, C.A. Iglesias, Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications, Expert Systems with Applications 77 (2017) 236–246, https://doi.org/10.1016/j.eswa.2017.02.002.
[47] Y.-C. Chang, C.-H. Ku, C.-H. Chen, Using Deep Learning and Visual Analytics to Explore Hotel Reviews and Responses, Tourism Management 80 (2020), 104129, https://doi.org/10.1016/j.tourman.2020.104129.
[48] R. Nie, Z. Tian, J. Wang, K.S. Chin, Hotel Selection Driven by Online Textual Reviews: Applying a Semantic Partitioned Sentiment Dictionary and Evidence Theory, International Journal of Hospitality Management 88 (2020), 102495, https://doi.org/10.1016/j.ijhm.2020.102495.
[49] Y. Liu, K. Huang, J. Bao, K. Chen, Listen to the voices from home: An analysis of Chinese tourists' sentiments regarding Australian destinations, Tourism Management 71 (2019) 337–347, https://doi.org/10.1016/j.tourman.2018.10.004.
[50] S. Wu, Y. Xu, F. Wu, Z. Yuan, Y. Huang, X. Li, Aspect-Based Sentiment Analysis Via Fusing Multiple Sources of Textual Knowledge, Knowledge-Based Systems. 183 (2019) 104868. 10.1016/j.knosys.2019.104868.
[51] H.T. Nguyen, M.L. Nguyen, Multilingual Opinion Mining on YouTube – A Convolutional N-gram BiLSTM Word Embedding, Information Processing & Management. 54 (2018) 451–462. 10.1016/j.ipm.2018.02.001.
[52] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT, Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota.
[53] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, ArXiv:1301.3781 [Cs]. (2013). http://arxiv.org/abs/1301.3781 (accessed December 19, 2018).
[54] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014: pp. 1532–1543. http://www.aclweb.org/anthology/D14-1162 (accessed December 19, 2018).
[55] M.E. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-Supervised Sequence Tagging with Bidirectional Language Models, ArXiv:1705.00108 [Cs]. (2017). http://arxiv.org/abs/1705.00108 (accessed October 13, 2020).
[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 6000–6010.
[57] R. Liu, F. Mai, Z. Shan, Y. Wu, Predicting Shareholder Litigation on Insider Trading from Financial Text: An Interpretable Deep Learning Approach, Information & Management (2020), 103387, https://doi.org/10.1016/j.im.2020.103387.
[58] A. Brahma, D.M. Goldberg, N. Zaman, M. Aloiso, Automated Mortgage Origination Delay Detection from Textual Conversations, Decision Support Systems. (2020) 113433. 10.1016/j.dss.2020.113433.
[59] B. Huang, Y. Ou, K.M. Carley, Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks, (2018). https://arxiv.org/abs/1804.06536v1 (accessed October 4, 2020).
[60] X. Li, L. Bing, P. Li, W. Lam, Z. Yang, Aspect Term Extraction with History Attention and Selective Transformation, (2018). https://arxiv.org/abs/1805.00760v1 (accessed October 4, 2020).
[61] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for Aspect-level Sentiment Classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2016) 606–615, https://doi.org/10.18653/v1/D16-1058. Austin, Texas.
[62] S. Poria, E. Cambria, A. Gelbukh, Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network, Knowledge-Based Systems. 108 (2016) 42–49. 10.1016/j.knosys.2016.06.009.
[63] C. Sur, RBN: Enhancement in Language Attribute Prediction Using Global Representation of Natural Language Transfer Learning Technology Like Google BERT, SN Appl. Sci. 2 (2019) 22, https://doi.org/10.1007/s42452-019-1765-9.
[64] J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018: pp. 328–339. 10.18653/v1/P18-1031.
[65] C. Sun, L. Huang, X. Qiu, Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, (2019). https://arxiv.org/abs/1903.09588v1 (accessed October 4, 2020).
[66] X. Li, X. Fu, G. Xu, Y. Yang, J. Wang, L. Jin, Q. Liu, T. Xiang, Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis, IEEE Access 8 (2020) 46868–46876, https://doi.org/10.1109/ACCESS.2020.2978511.
[67] Z. Wu, D.C. Ong, Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis, in: 35th AAAI Conference on Artificial Intelligence, AAAI Press, Virtual

Conference (2021) 1–9, http://adsabs.harvard.edu/abs/2020arXiv201007523W (accessed May 19, 2021).
[68] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, SemEval-2015 Task 12: Aspect Based Sentiment Analysis, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics (2015) 486–495, https://doi.org/10.18653/v1/S15-2082. Denver, Colorado.
[69] D. Meškelė, F. Frasincar, ALDONAr: A Hybrid Solution for Sentence-Level Aspect-Based Sentiment Analysis Using a Lexicalized Domain Ontology and a Regularized Neural Attention Model, Information Processing & Management 57 (2020), 102211, https://doi.org/10.1016/j.ipm.2020.102211.
[70] H. Wan, Y. Yang, J. Du, Y. Liu, K. Qi, J.Z. Pan, Target-Aspect-Sentiment Joint Detection for Aspect-Based Sentiment Analysis, AAAI. 34 (2020) 9122–9129, https://doi.org/10.1609/aaai.v34i05.6447.
[71] J. He, X. Wang, M.B. Vandenbosch, B.R. Nault, Revealed Preference in Online Reviews: Purchase Verification in the Tablet Market, Decision Support Systems 132 (2020), 113281, https://doi.org/10.1016/j.dss.2020.113281.
[72] Y. Zhao, X. Xu, M. Wang, Predicting Overall Customer Satisfaction: Big Data Evidence from Hotel Online Textual Reviews, International Journal of Hospitality Management 76 (2019) 111–121, https://doi.org/10.1016/j.ijhm.2018.03.017.
[73] L.-A. Cotfas, C. Delcea, R.J. Milne, M. Salari, Evaluating Classical Airplane Boarding Methods Considering COVID-19 Flying Restrictions, Symmetry 12 (2020) 1087, https://doi.org/10.3390/sym12071087.
[74] W. Li, S. Yu, H. Pei, C. Zhao, B. Tian, A Hybrid Approach Based on Fuzzy Ahp and 2-Tuple Fuzzy Linguistic Method for Evaluation in-Flight Service Quality, Journal of Air Transport Management 60 (2017) 49–64, https://doi.org/10.1016/j.jairtraman.2017.01.006.
[75] A. Brochado, P. Rita, C. Oliveira, F. Oliveira, Airline Passengers' Perceptions of Service Quality: Themes in Online Reviews, International Journal of Contemporary Hospitality Management 31 (2019) 855–873, https://doi.org/10.1108/IJCHM-09-2017-0572.
[76] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, (2016). https://arxiv.org/abs/1609.08144v2 (accessed October 4, 2020).
[77] Y.-C. Chang, C.-C. Chen, Y.-L. Hsieh, C.C. Chen, W.-L. Hsu, Linguistic Template Extraction for Recognizing Reader-Emotion and Emotional Resonance Writing Assistance, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics (2015) 775–780, https://doi.org/10.3115/v1/P15-2127. Beijing, China.
[78] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, 1 edition, Cambridge University Press, New York, 2008.
[79] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-Balanced Loss Based on Effective Number of Samples, 2019, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9260–9269, https://doi.org/10.1109/CVPR.2019.00949.
[80] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, USA, 2015: pp. 19–27. 10.1109/ICCV.2015.11.
[81] Y. Kim, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics (2014) 1746–1751, https://doi.org/10.3115/v1/D14-1181. Doha, Qatar.
[82] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, ArXiv: 1711.05101 [Cs, Math]. (2019). http://arxiv.org/abs/1711.05101 (accessed October 4, 2020).
[83] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval: The concepts and technology behind search, 2nd ed., Addison-Wesley Publishing Company, USA, 2011.
[84] Q. McNemar, Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages, Psychometrika 12 (1947) 153–157, https://doi.org/10.1007/BF02295996.
[85] V. Bogicevic, W. Yang, A. Bilgihan, M. Bujisic, Airport Service Quality Drivers of Passenger Satisfaction, Tourism Review. 68 (2013) 3–18, https://doi.org/10.1108/TR-09-2013-0047.
[86] F. Ali, W.G. Kim, K. Ryu, The Effect of Physical Environment on Passenger Delight and Satisfaction: Moderating Effect of National Identity, Tourism Management 57 (2016) 213–224, https://doi.org/10.1016/j.tourman.2016.06.004.
[87] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[88] Y. Ju, K.-J. Back, Y. Choi, J.-S. Lee, Exploring Airbnb Service Quality Attributes and Their Asymmetric Effects on Customer Satisfaction, International Journal of Hospitality Management 77 (2019) 342–352, https://doi.org/10.1016/j.ijhm.2018.07.014.
[89] W. Kwon, M. Lee, K.-J. Back, Exploring the Underlying Factors of Customer Value in Restaurants: A Machine Learning Approach, International Journal of Hospitality Management 91 (2020), 102643, https://doi.org/10.1016/j.ijhm.2020.102643.
[90] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146, https://doi.org/10.1162/tacl_a_00051.
[91] A. Punel, L.A.H. Hassan, A. Ermagun, Variations in Airline Passenger Expectation of Service Quality Across the Globe, Tourism Management 75 (2019) 491–508, https://doi.org/10.1016/j.tourman.2019.06.004.
[92] A.L. Steinbach, D.L. Gamache, R.E. Johnson, Don't Get It Misconstrued: Executive Construal-Level Shifts and Flexibility in the Upper Echelons, AMR. 44 (2018) 871–895, https://doi.org/10.5465/amr.2017.0273.
[93] S. Chatterjee, P. Mandal, Traveler Preferences from Online Reviews: Role of Travel Goals, Class and Culture, Tourism Management 80 (2020), 104108, https://doi.org/10.1016/j.tourman.2020.104108.
[94] N. Huang, G. Burtch, Y. Hong, E. Polman, Effects of Multiple Psychological Distances on Construal and Consumer Evaluation: A Field Study of Online Reviews, Journal of Consumer Psychology 26 (2016) 474–482, https://doi.org/10.1016/j.jcps.2016.03.001.
[95] R. Hussain, A.A. Nasser, Y.K. Hussain, Service Quality and Customer Satisfaction of a UAE-Based Airline: An Empirical Investigation, Journal of Air Transport Management 42 (2015) 167–175, https://doi.org/10.1016/j.jairtraman.2014.10.001.

Yung-Chun Chang received the PhD degree in information management from National Taiwan University, Taiwan, in 2016. He is currently an assistant professor in the Graduate Institute of Data Science at Taipei Medical University. His papers have appeared in the *Tourism Management, JAMIA, IEEE Transactions on Knowledge and Data Engineering, Journal of Business Research, International Journal of Information Management, Bioinformatics, JASIST, and ACL*, etc. His current research interests include natural language processing, text mining, information retrieval, knowledge discovery, and question answering.

Chih-Hao Ku is an Assistant Professor in the Department of Information Systems of Monte Ahuja College of Business at Cleveland State University. Dr. Ku received his M.S. and Ph.D. in Information Systems and Technology in 2012 at Claremont Graduate University. His research currently focuses on deep learning, natural language processing, and visual analytics. His work appears in the *Journal of MIS, Tourism Management, Journal of Business Research, JASIST, Government Information Quarterly, Journal of Information Systems, and International Journal of Information Management* among others.

Nguyen Le Duy Duc received the master's degree in Data Science at the Taipei Medical University, Taiwan, in 2021. His research interests include information retrieval and text mining. His research interests include natural language processing and information retrieval. He has participated in many research projects such as sentiment analysis on social media, and micro-activity retrieval of daily living.