LUND UNIVERSITY
School of Economics and Management

Department of Economics

Mining the Skies:

An Exploration of Airline Reviews using LDA

M.Sc. Data Analytics and Business Economics

Author:

Joel Ljungström

Supervisors:

Joakim Westerlund & Rani Basna

May 24, 2023

Abstract

Airlines face a highly competitive industry requiring large capital investment at thin profit margins. In addition, the carriers are re-emerging from the COVID-19 pandemic; a time period which brought the industry to a standstill and revenue streams to an unforeseen low. Airlines are now facing their next challenge: regaining their market shares to pre-pandemic levels, and beyond. As such, reputation and recommendations are important factors for the companies to gain new customers, and online reviews presents a unique, low-cost opportunity to learn from indirect feedback and predict what areas of their products and services leads to positive recommendations. Through Latent Dirichlet Allocation (LDA), this study extracted 18 unique latent topics from 128,631 samples to identify key areas often written about in reviews. Additionally, the probabilities of these topics to occur in reviews were used to predict the outcome of recommendation and overall ratings with the use of classification trees and variations of logistic regression. The top performing model had an accuracy of 85.77% in predicting recommendation, and multiple areas were identified as opportunities for airlines to make managerial decisions on to improve their reputation online. Key dimensions relating to a positive recommendation found were Good Service, Efficiency and Cabin Crew, whereas dimensions relating to a negative recommendation were identified as Bad Customer Service, Travel Delays, and Charges.

Table of Contents

1. Introduction

Passenger air traffic is a lucrative industry generating hundreds of billions of dollars per year (IATA, 2021). The industry not only face fierce competition through its low profit margins and threat of low-cost carriers, high barriers of entry due to extensive capital investments required, and external challenges. According to Calisir, Basak and Calisir (2016), external factors facing the industry include "(1) fuel price, (2) low-cost carriers, (3) economic crisis, (4) increasing security precautions, (5) personnel shortage, (6) government regulations, etc." As the result of the COVID-19 pandemic, it can be argued that public health crisis can be added to the list of external challenges.

The pandemic drove the industry to a halt through government regulations, which, to a large extent, prohibited tourism and movement of people (Rita, Moro & Cavalcanti, 2022). The year was expected to break previous records in number of travellers, expectations which were quickly grinded to a halt overnight and resulted in a sharp 61% decline from predicted 4,723 million to 1,807 million passengers (IATA, 2022). See Figure 0-1 and Figure 0-2 in Appendix A for detailed graphs on industry revenue and passengers served. With the mixture between abandonment of government regulations throughout 2021 and 2022, and desire of wanderlust travellers to explore the world once again after heavy restrictions, airlines underwent rapid adaption and competition to regain their position within the vast half-trillion-dollar revenue sector once again (IATA, 2021; Rita et al., 2022).

Considering the multitude of challenges confronting firms within the passenger aviation industry, customer satisfaction emerges as a significant determinant of passenger loyalty, as well as the success and profitability of an airline (Calisir et al., 2016). Although their work was limited to managerial actions during the periodic event of the pandemic, recent research by Kiraci, Tanriverdi and Akan (2023) corroborated this point of view and ascertained that information sharing, specifically the exchange of relevant details with customers, assumes a pivotal role not only in financial performance, but also in fostering customer satisfaction and nurturing their tendency to recommend the airline's products and services to others (Kiraci et al., 2023). Although their work was limited to managerial actions during the periodic event of the pandemic, Boubker and Naoui (2022) further highlighted that passengers who experience satisfaction are inclined to develop affinity towards the brand and exhibit heightened levels of loyalty. Customers who hold this genuine affection for the brand possess the potential to evolve into loyal customers and actively engage in positive word-of-mouth recommendation within their social groups (Boubker & Naoui, 2022; Namukasa, 2013).

Various methods to study airline customer satisfaction have been employed in different research (Boubker & Naoui, 2022; Gao & Koo, 2014; Lucini, Tonetto, Fogliatto & Anzanello, 2020; Namukasa, 2013). Conventional statistical approaches, such as regression analysis, offers insight into the relationship between quantitative ratings of various dimensions and customer satisfaction through passenger surveys and questionnaires (Cosmina Laura, Maria & Cristina, 2022). However, upon examining existing research, it may be deduced that these methodologies are susceptible to (1) time consuming data collection, (2) limited capture of observations, (3) constrained dimensions for respondents' reflection, and (4) prone to attract potential bias in favour of more negative versus positive responses. For example, Han and Anderson (2022) found that the higher status passengers have in a loyalty membership program increases their likelihood in participating in a survey, and that there is a tendency to post negative responses rather than positive ones. Additionally, Calisir et al. (2016) identified that service quality and price to have positive effects on customer satisfaction on a specific route, but were limited to 175 questionnaires collected from face-to-face interviews. These challenges restrict the method from potentially finding emerging new preferences and may risk injecting noise into observations.

An alternative approach is to analyse Online Customer Reviews (OCRs), which can be said to be a mixture of User Generated Content (UGC) and electronic Word-of-Mouth (eWOM) as introduced by Mauri and Minazzi (2013). This alternative offers several advantages. Firstly, it enables consumers to share impressions online to create an ever-lasting impact accessible by any individual at any point in time. Secondly, they are not limited to only a product or service, but can also include organizations and destinations, and thirdly, they offer a valuable source of organic information which differs from advertisement or marketing initiatives by companies (Mauri & Minazzi, 2013). As further highlighted by Noh, Jeon and Hong (2023), the online domain presents a "never-ending stream of reviews", providing researchers with a virtually limitless repository of observations that grow on a daily basis. This highlights an opportunity for businesses to gain a deeper insight into their customers' experiences, and leverage from an additional stream of indirect feedback. To bridge the gap between this indirect source of feedback and to allow executives to optimise their business operations, topic modelling OCRs presents a unique opportunity to model quantitative ratings and employ text analysis methods on individuals' opinions.

The objective of this study is to apply machine learning methods to analyse the collective information inherent in thousands of OCRs with the goal of offering actionable

insights for airlines to enhance their competitiveness. Hence, the research question which will be investigated in this study is:

To what extent can airline recommendations be predicted using online customer reviews?

In addition, investigating this research question will allow for the following questions to be reviewed:

1. What are key dimensions of customer satisfaction expressed in OCRs?
2. How has the dimensions changed between the pre-pandemic- and post-pandemic era?
3. What are important aspects influencing customer's recommendation of an airline?
4. What are important aspects influencing customer's overall rating of airlines?

2. Previous Research

The field of text analysis has been wildly studied in a range of areas spanning from predicting IPO outcome using prospectus content (Emidi & Galan, 2022) to identifying informative topics relating to COVID-19 in tweets (Khan & Chua, 2021) and topic modelling customer feedback from online ticketing systems (Ponay, 2022). In recent years, additional research into mining text of UGCs have emerged predominantly through online social networks, such as Tweets on Twitter (Uthirapathy & Sandanam, 2023; Wan & Gao, 2015), and further studies have been conducted specifically into the airline industry (Lakshmanarao, Gupta & Kiran, 2022; Lucini et al., 2020). Specifically, LDA is an admired method utilized for topic modelling.

Lakshmanarao et al. (2022) investigated how airlines can look for ways to increase the quality of their products and services through sentiment analysis of 11,540 tweets. The researchers applied four different deep learning techniques to predict three categories of sentiment (negative, neutral, and positive) of the tweets. Results showed that the researchers were able to achieve a 93% accuracy score in one of their models, however, did not declare how the sentiment scoring was conducted.

One recent study from the hospitality industry studied travellers' reviews on hotel performance. Roy (2023) scraped reviews from TripAdvisor from the six months period leading up to their study. The data of 6,355 observations included a text-based review in addition to a categorical hotel type variable (luxury, mid-tier, low-tier), a walkable score between 0-100, and the number of nearby attractions to the hotels. They calculated sentiment scores of reviews, performed topic extraction from the UGC and employed multiple linear

regression analysis and found that hotel guests tend to be vocal about three topics in particular: Safety and Security, Health and Wellbeing, and Daily Comfort of Life. The researcher highlighted the importance for managers to take necessary actions in improving these intangible assets Roy (2023).

Yao, Yuan, Qian and Li (2015) extracted 7,466 OCRs from airline review website Airlinequality.com to research common concerns for among travellers. OCRs from 25 airlines were used, each with a sample range spanning 87 to 694. By analysing the top 20% frequent words, they found that concerns were almost similar regardless of airline, and that individual airlines only had reviews of very few distinct concerns. Additionally, Airlinequality's parent company Skytrax offers a quality rating for airlines which is widely recognized in the industry and is a status symbol used by marketing teams at airlines. They found that when comparing concerns extracted from OCRs with the quality rating system, 4-star and 5-star companies had similar concerns, and that the greater the star level difference, the larger the difference in service quality concerns were.

A similar study conducted on OCRs extracted from Airlinequality.com by Lucini et al. (2020) used a sample size of 55,775 OCRs. In contrast to analysing top-word frequencies, they utilized an LDA model to discover latent topics, a Naïve Bayes Classifier for sentiment scoring of adjectives used in the reviews, and logistic regression to model the relationship between topics and travellers' recommendation. They identified 27 unique latent topics, where "Cabin staff" (8.58), "Onboard service" (7.77), and "Value for money" (6.24) had the highest topic coefficients in airline recommendation. Additionally, their sentiment scores identified words associated with a positive recommendation and a negative recommendation. These words were "Good", "Excellent", "Great" and "Absurd", "abysmal", "dismissive" for the two outcomes respectively.

In their study about predicting reviewers' ratings of a specific product, Poushneh and Rajabi (2022) used LDA and GBDT on a sample size of 6,855 text-based reviews and product scores. 10 latent topics were discovered and grouped into a binary category of Abstract/Non-abstract topics. They found that non-abstract topics were able to predict review scores, whereas abstract topics were not. Additionally, they recommended analysis of associations between reviews and ratings be conducted prior to drawing conclusions from OCRs.

This study will use methods employed in previously discussed research to study the relationship between OCRs and recommendations and ratings of airlines to contribute to the literature. As can be reflected from previous studies, sample sizes may be considered to have

been restricted. Given the nature of LDA, the notion of "the more, the merrier" will hold true by examining a larger sample size in present study compared to previous studies (Lakshmanarao et al., 2022; Lucini et al., 2020; Yao et al., 2015). Additionally, whereas an emphasis in previous studies have been put on sentiments of the reviews, this study will focus on the concrete topics being reflected about airlines in the reviews. Finally, more insight into the attributions of customer satisfaction will be generated by deploying different machine learning methods to predict two dependent variables instead of the more common approach of restricting to one.

3. Methodology

A three-step process will be implemented to analyse the research question. Firstly, a web crawler will be built and deployed to capture as many OCRs as possible. The sample data will undergo cleaning and pre-processing to adapt for the machine learning methods utilized. Secondly, Natural Language Processing (NLPs) method Latent Dirichlet Allocation (LDA) will be applied to the full set of OCRs to discover latent topics. These topics will be designated a term with logical connection to the topic words in order to identify the underlying customer satisfaction dimensions. Thirdly, varieties of logistic regression and classification trees will be employed to predict travellers' airline rating and recommendation using the probabilities of latent topics occurring in the reviews generated from the LDA.

3.1 Data Collection

To collect the required data for this research, a custom web crawling application was developed to scrape and extract public reviews from a website known as Air Travel Review (ATR). The website, owned and operated by the widely recognized airline rewards company SkyTrax, serves as a customer forum, and can be accessed from the following url: https://www.airlinequality.com. The website provides comprehensive reviews for airports, airport lounges, airline seats, and airlines, and offers a wide range of dimensions for reviewers to evaluate and score. Two other similar websites were considered as an option for data extraction; TripAdvisor (https://www.tripadvisor.com/) and Trustpilot (https://www.trustpilot.com/). Both websites are well-established in the travel industry for their extensive databases of UGC and OCRs encompassing travel-related businesses and airlines. One notable aspect of these platforms is that they also offer a flipside for businessowners to showcase their ratings through their recognizable badges. These badges

serve as a visual representation of a business' performance and helps to instil trust and confidence in potential customers. Although a great feature for positively reviewed companies, this provides marketing teams at airlines with an opportunity to offer incentives to consumers to artificially influence their reviews and accelerate their reputation. This phenomenon, referred to as “amplified WOM” by Mauri and Minazzi (2013), may introduce bias in the collective sentiment of the OCRs. To mitigate this potential bias and capture more authentic and “organic WOM” (Mauri & Minazzi, 2013), ATR was selected in this study as the primary data source. Since ATR offer a niche website specifically within the airline industry and operate independently of third-party affiliations, it can be justified that truthful, organic UGC are more likely to be published there, providing a valuable source of reliable information for the study.

The web crawler programmed exclusively gathered individual reviews related to the airlines on ATR to facilitate data aggregation for analysis. These individual OCRs consist of 21 variables (see Table 0-1 in Appendix B for a full explanatory list of variables). It is important to note that due to the development of ATR, not all dimensions have been historically rateable. Over time, additional dimensions have been added, resulting in incomplete data. For instance, certain *ratings* variables and the *Aircraft* dimension have only been present on the site for the last seven years according to our data, as illustrated in Figure 0-1 in Appendix C which presents the duration of availability for each OCR dimension on ATR. In this study, particular focus will be given to three specific variables described in Table 3-1 in addition to the *Review* text-based variable. *Review*, *Year Published*, *Recommended* and *OverallScore* have all been present on ATR for the majority of the forum’s life, with the exception of *OverallScore* which was added in 2008 according to our data (see Figure 0-2 in Appendix C). There were 4,330 (3.37%) observations missing in this dimension, which were instead replaced by the average *OverallScore* rating, rounded to nearest integer (5). Additionally, due to class imbalance between individual scores, where some scores were represented in only 3% of the reviews, they were aggregated into three distinct categories based on the perceived logical connection to a relevant context: “Low Score” (1-3), “Neutral” (4-6), and “High Score” (7-10). Moreover, to identify time periods before and after the pandemic, the published years were binned into three categories: 2002-2015, 2016-2019, 2020-2023.

Table 3-1: Occurrence of selected variables in the sample (n = 128,631).

Variable	No. of reviews	% of total
Overall Rating		
Low Score (1-3)	63,527	49.39
Neutral (4-6)	18,872	14.67
High Score (7-10)	46,232	35.94
Recommended		
No	77,009	59.87
Yes	51,622	40.13
Year Published		
2002-2015	47,340	36.81
2016-2019	52,833	41.08
2020-2023	28,458	22.13

Furthermore, the distribution of word counts for the *Review* variable can be found in Table 3-2. Whereas there is a wide span of word counts in the reviews (ranging from 1-1,058), the average review has a length of 136 words, which brings a considerable amount of data to the LDA model.

Table 3-2: Statistic summary for word count of Review variable in OCRs.

Minimum	25th percentile	Median	Mean	75th percentile	Maximum
1	70	109	136	171	1,058

The text-based variable in the dataset has been identified to contain reviews written exclusively in the English language. The timespan of the published OCRs and dimensions are from January 2002 to May 2023. To complement the analysis and assist with the pre-processing of the OCRs, data relating to airports, IATA airport codes, cities and country names worldwide was extracted from public data sets provided by OpenFlights (https://openflights.org).

The crawlers was built in Python using the Scrapy package (Scrapy, 2023). It downloads the ATR Airline Review HTML code and extracts the data based on requested paths. Two different crawlers were programmed:

1. The first crawler extracts the URL slug and the number of reviews for each airline listed on the website.

2. The second crawler utilizes information obtained from the first crawler to load the independent URLs for each airline and extract the OCRs present in the specific URL.

After the data extraction process, faulty scraped observations were removed from the dataset. More specifically, observations where the *Review* variable was either an N/A-value or where it was not a string, were removed from the dataset. Additionally, observations of the *Airline Name* variable that did not match one of the airline names from the first crawler were removed. The finished dataset consisted of 128,631 reviews from 547 airlines. This dataset was exported as a comma-separated value (.csv) file, and the final crawling procedure was conducted on May 9th, 2023.

3.2 Pre-processing

Conducting analysis of free-form text can pose considerable challenges, attributable to a number of factors. Tirunilla and Tellis (2014) highlights specifically three reasons to this; (1) There is a lack of structure in free-flowing text, especially so for UGCs and OCRs. The reason for this is that reviewers tend to be irregular and casual in their usage of grammar and choice of words. (2) Non-informative words relating to the product or service must be removed to extract the context of the OCRs. (3) To enable textual analysis and manipulation, text must be converted to a numeric format. Additionally, Guo, Barnes and Jia (2017) further declared that consumers have a tendency to use words relating to their personal experience, leading to only reviewing particular dimensions concerned to them. Consequently, pre-processing and cleaning of textual data is a necessity to remove redundant and unrelated information pertaining to the message of an OCR.

The pre-processing carried out in this study is much similar to ones carried out for LDA analysis in previous studies of OCRs (Guo et al., 2017; Lucini et al., 2020; Tirunilla & Tellis, 2014; Wang, Shu, Hsu, Lin & Tseng, 2011). Firstly, non-English characters present in the OCRs were converted to ASCII code. For example, Zürich was converted to Zurich, Köln to Koln, etc. Secondly, each individual review was converted to a single string of text and was made all lower case by removing special characters, such as non-space characters, symbols, punctuation, exclamation marks, etc. Thirdly, in order to reduce redundant and non-informative words, stop words were removed and words describing named entities were replaced by a common term using spaCy's (https://spacy.io) transformer-powered Named Entity Recognition (NER) model. For example, terms such as "kilogram", "30 minutes",

"dollars", "Stockholm" were identified and replaced with common words "weight", "time", "price", "city". Since the NER model provided by spaCy demonstrates levels of accuracy for entity recognition at 85% precision (Honnibal & Montani, 2017), their model was deemed the most prominent option for computational efficiency and accuracy. However, as a precautionary measure, the fore mentioned OpenFlights data on airports was used to address any instances where the NER model may have missed to identify an airport or a route. In such cases, the identified entities were replaced with generic terms "airport" and "route".

Lastly, the deployment of spaCy's NER model further allowed the usage of Part-of-Speech (POS) tagging as well as lemmatization (Honnibal & Montani, 2017). POS is a method to identify the syntax of each word in a sentence. For instance, the sentence "Plane was super clean" results in "Plane [noun] was [verb] super [adverb] clean [adjective]", and to pertain context from the reviews, only nouns and adjectives were kept (Lucini et al., 2020). Additionally, lemmatization reduces words to their base dictionary form. For example, the words "flew" or "flying" were reduced to the dictionary form "fly".

With these pre-processing steps, the three challenges of free-form text highlighted by Tirunilla and Tellis (2014) have been effectively resolved. An original review which read:

"A one hour flight with an old ATR42 turboprop. Seats surprisingly comfortable and we were offered a free drink plus candy onboard. Nothing special but we arrived on time."

became:

"['hour', 'flight', 'old', 'turboprop', 'seat', 'comfortable', 'free', 'drink', 'candy', 'special', 'time']"

The resulting tokenized text (running text converted to individual words) became the corpus required to be used for the topic detection and sentiment analysis in this study (Blei, Ng, Jordan & Lafferty, 2003). In the study conducted by Guo et al. (2017), the data underwent further processing, where low frequency words were excluded from the OCRs used from TripAdvisor. However, unlike this study, Guo et al. (2017) did not eliminate stop words, resulting in longer text strings with a lower diversity of unique words. Consequently, their implementation of removing low frequency words was justified, whereas in the present study it would carry the potential risk of eliminating a significant number of words, and consequently valuable data. The text pre-processing was implemented with the use of spaCy and Natural Language Toolkit (www.nltk.org) in the Python programming environment.

3.3 Topic Detection with LDA

The goal of this study is to identify key dimensions from OCRs which affect customer recommendation of airlines. To accomplish this, Latent Dirichlet Allocation (LDA) was employed as an unsupervised method for topic detection. LDA is a form of generative probabilistic model for identifying and extracting themes which are found in a collection of documents, also known as a corpus. In this research, the pre-processed OCRs consists of individual words referred to as tokens. The tokens make up the individual documents, which are collectively referred to as the corpus, and the LDA is used analyse all documents and identify the main themes (or topics) of the corpus (Blei et al., 2003). In essence, the algorithm guesses randomly which tokens belong to which topic and updates its guesses until it finds the best fitting ones. With each epoch, it discovers the most probable topics based on how often certain words appear together in the documents, and once these topics have been found, it can extract the probability of words belonging to a topic (Blei et al., 2003).

The model assumes that words in each document are represented as a probabilistic distribution over the latent topics (in this case the dimensions of customer satisfaction), each latent topic is represented as a probabilistic distribution over words, and the distribution of topics and the distribution of words share a common Dirichlet prior (Blei et al., 2003). As a result, similar topics may be found across multiple documents, while every document has its own mixing proportion of topics. The procedure follows a generative process for each document inside the corpus as detailed in Table 3-3 (Blei et al., 2003; Lucini et al., 2020).

Table 3-3: Notation and generative process for LDA.

N	Number of words in each document
ξ	Parameter of the Poisson distribution showing length of each document
θ	The topic distribution of each document
α	The parameter of Dirichlet prior of topic distribution of each document
z_n	The topic of he n-th word
$\mathbf{z}$	A set of topics
$p(w_n \mid z_n, \beta)$	A multinomial probability conditioned on topic z_n
β	Parameter of Dirichlet prior of word distribution per topic

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dirichlet(\alpha)$
3. For each of the N words w_n, where $\boldsymbol{w} = \{w_1, \boldsymbol{w}_2, \dots, w_N\}$:
   - 3.1. Choose $z_n \sim Multinomial(\theta)$
   - 3.2. Choose w_n from $p(w_n|z_n, \beta)$

The inferential problem, which needs to be solved in order to use LDA, is to compute the posterior distribution of the hidden variables $\theta, \boldsymbol{z} \mid \boldsymbol{w}$ for each document, given by Equation 1 (Lucini et al., 2020). Since this distribution is intractable to compute, an approximate algorithm, such as Gibbs Sampling or Variational Bayes (VB), must be used. In this study, a version of VB was used with the LDA model, available in the genism package (Řehuřek & Sojka, 2010).

Equation 1: Posterior distribution of hidden variables.

$$p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}| \alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)}$$

Zeng, Cheung and Liu (2013) compared the accuracy and efficiency of VB, Belief Propagation (BP), and Gibbs Sampling (GS) and found BP to be a faster and more accurate algorithm than VB in topic modelling for LDA. However, in contrast, Hoffman, Blei and Bach (2010) developed the Online Variational Bayes, which demonstrated to be more robust and computationally feasible with larger corpora as it does not incur large memory costs, unlike traditional options such Markov Chain Monte Carlo (MC) algorithms. Since the corpus used in this study may be ruled as large, consisted of 128,631 documents, the usage of VB with the genism package was justified by the researcher.

Another option for researchers is to determine the number of topics they would like the LDA to discover. This can be a challenge for large corpora as "the number of individual dimensions of satisfaction is not known *a priori*" (Lucini et al., 2020), or prior to running the model. Whereas an insufficient number of topics could result in a model that captures inaccurate dimensions, a larger number of topics may result in a complex model which makes "interpretation and subjective validation difficult" (Lucini et al., 2020). To find the optimal number of topics, various LDA models were fitted with different number of topics and evaluated using coherence scores, a popular method used in LDA topic analysis (Ponay, 2022; Khan & Chua, 2021; Wang, Feng & Dai, 2018), and also part of the genism package

(Řehuřek & Sojka, 2010). In essence, the coherence score attempts to measure the human interpretability of words found in a topic generated by LDA, which is useful in present study when identifying the customer dimensions of the latent topics. The topic coherence score selected for this study, denoted *C_v* in genism or *any-any coherence* by Rosner, Hinneburg, Röder, Nettling and Both (2014), measures the degree of similarity between high-scoring words in a single topic and generates an aggregated score for each model by averaging the score for each topic (Ponay, 2022). There are various methods to calculate the coherence score (Rosner et al., 2014). The first evaluation of the different measures conducted by Rosner et al. (2014) showed that the *any-any coherence* approach resulted among the highest scores for human interpretability. Other options include (1) *u_mass*, which calculates the degree of similarity between topics based on the distribution of words, (2) *c_uci* which measures the semantic similarity between words based on a rolling window, and (3) *c_npmi* which is a pointwise mutual information measure taking into account both the within-topic probability as well as the overall probability of word occurrence (Řehuřek & Sojka, 2010).

Previous research into LDA on OCRs have fitted models with $t = 2, \dots, 100$ topics, however, a narrower range for labelling and evaluation of topics have been limited to $t = 10, \dots, 30$ (Guo et al., 2017; Wang et al., 2018). In the light of this, the present study justifiably opted to fit LDA models ranging from $t = 2, \dots, 19$ topics to prioritize interpretability and computational efficiency. To fit a robust LDA model for each topic t, the corpus was divided into k-batches of similar sizes to enable k-fold cross-validation (CV). This procedure allows for repetitive training by repeatedly fitting the LDA models to various subsets of the data, and validating them on the data not included in the training (Lindholm, Wahlström, Lindsten & Schön, 2022). With each loop l over $l = 1, 2, \dots, k$, one batch l was held out as validation data and the LDA was fitted on the remainder $k - 1$ batches of training data. For each iteration of the CV-loop, the coherence score for k was calculated and averaged over the l iterations (Lindholm et al., 2022). For each topic t, the coherence score was averaged over the k iterations to attain a reliable measure of topic coherence. In present study, k was set to 5.

Upon completing the LDA training, one last step remains to extract the dimensions of customer satisfaction; label the latent topics with dimensions based on the key words established by the model. From each latent topic discovered by the LDA, 10 words and their respective probability of occurring within the topic was extracted (Řehuřek & Sojka, 2010), see Figure 3-1 for visual illustration of the procedure. The procedure of assigning labels to

the topics was conducted by the researcher in collaboration with an independent party familiar with the work. Whereas the independent party was only provided with the words associated with each topic, and was not given access to any detailed information of the model itself, the researcher made logical connection between words by interpreting the most frequent words as well as their relative weight (Guo et al., 2017; Lucini et al., 2020). The absence of external influences allowed for an objective representation of the underlying themes in the topics. A candidate topic term was established for each topic independently by the researcher and the individual party. Once the candidate terms were identified, they were further evaluated by logical connection through discussion. By leveraging the expertise of prior knowledge of the researcher and incorporating the unbiased opinion of the third party ensured that topic names were determined through logical connection between words in the topics and their broader contexts.

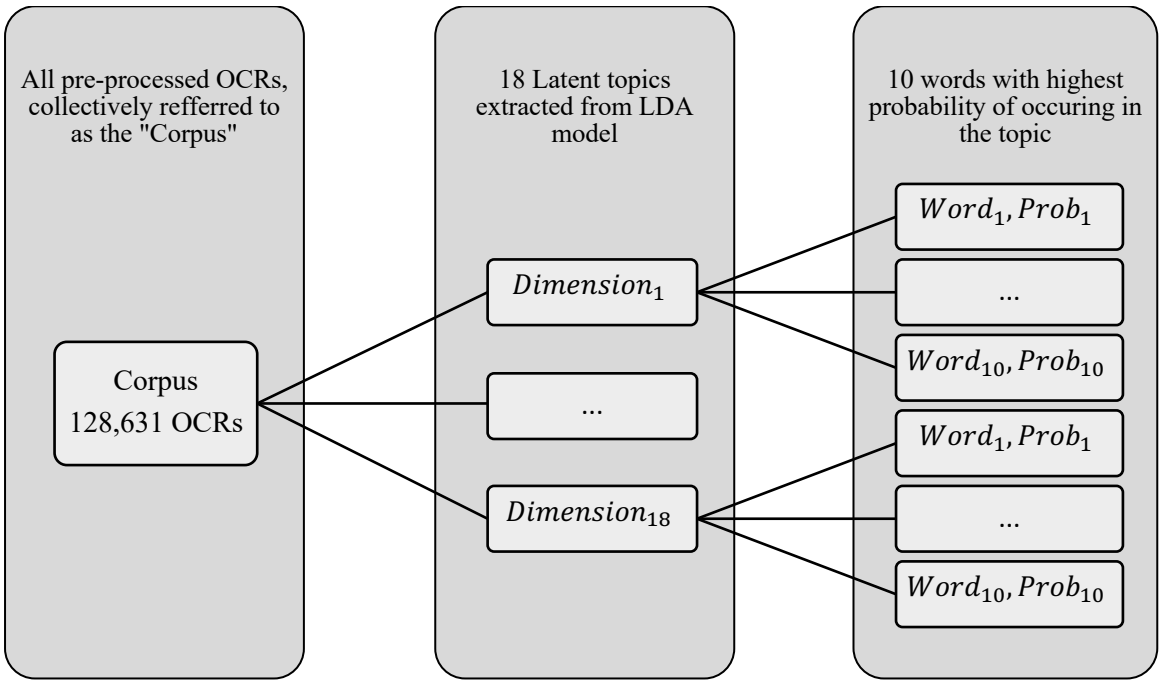

Figure 3-1: Visual illustration of words and probabilities extracted for each latent topic identified by LDA from the corpus.

3.4 Customer Satisfaction

Identifying the dimensions of customer satisfaction of the latent topics enabled evaluation of the topics' influence on the full corpus. More specifically, dimensions could be compared based on different groups of data, which in present study pertaining to the year period of published review, recommendation, and overall score. To accomplish this, the topic distributions were extracted from each OCR individually, resulting in 128,631 separate

probabilities for each dimension. Next, following the research of Lucini et al. (2020), individual dimension probabilities were summarized and divided by the total sum of all dimension probabilities. In doing so, the average probability of a dimension occurring in the corpus could be established and compared between data groups. This enables a simple way to explore the proportion of each dimension's distribution over the entire data set. See Figure 3-2 for visual representation of the procedure.

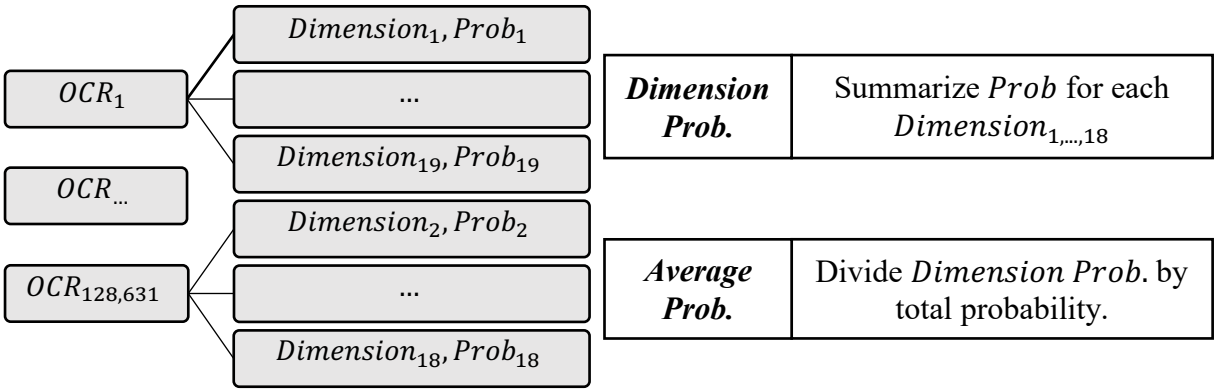

Figure 3-2: Visual illustration on attaining average probabilities of each dimension on the full corpus.

Next, to evaluate the change in dimension probabilities over time, and how the pandemic may have caused a shift in travellers' customer satisfaction, the *DatePublished* variable was used to extract the year of each published OCR. The airline industry has been identified as different eras through its history defined by fundamental changes to either technology or governance. For example, Heiets, La, Zhou, Xu, Wang and Xu (2022) defined the time period following the 1980's deregulation of the industry as a digital transformation for airlines, whereas the extensive disruption of the pandemic on the industry due to the World Health Organisation's (WHO) classification of COVID-19 as a pandemic in January 2020 (WHO, 2020) can be established as the mark of a new era. As such, the years of reviews were categorised into three separate eras: 2002-2015, 2016-2019, and 2020-2023. To allow for comparable measures, OCRs for these time periods were labelled as 'Digitalization era', 'Pre-pandemic era', and 'Post-pandemic era'. Following the grouping of time periods, box and whisker plots were generated using ggplot2 visualization package in R programming environment (Wickham, 2016). This allows for effective comparison between the dimension probabilities across the defined eras through interquartile ranges (IQR), which measures the spread of the middle 50% of the dimension probabilities.

3.5 Predicting Recommendation & Rating

One of the objectives of this research is to identify what dimensions of customer satisfaction contribute to airline recommendation and overall score. To accomplish this, a subset of the pre-processed dataset was created with target variables *Recommended* and *OverallScore*. In addition to the two variables, the dimension occurrence probabilities previously extracted for each OCR (see Customer Satisfaction) populated the new data set. Each dimension became a feature, and each observation represented the probability of said topic being present in the review. See Table 3-4 for an overview of the data set. During modelling, the complete data table was used, excluding the respective target variable of *OverallScore* when modelling *Recommended*, and vice versa.

Table 3-4: Table overview for new data set deployed with logistic regression and classification tree.

OCR Row ID	**Recommended**	**OverallScore**	**Dimension_1**	**Dimension…**	**Dimension_18**
1	Yes/no	1-10	[0-1]	[0-1]	[0-1]
…	Yes/no	1-10	[0-1]	[0-1]	[0-1]
128,631	Yes/no	1-10	[0-1]	[0-1]	[0-1]

To predict these target variables, two machine learning methods were employed: logistic regression and classification trees. Given the categorical nature of these target variables, logistic regression was primarily considered the appropriate method for training a model to predict the occurrence of airline recommendation based on topic probabilities. However, to evaluate the performance of logistic regression, classification trees were also modelled. Logistic regression is a parametric method to model conditional class probabilities by calculating the probability of a predicted label given inputs from the data. Additionally, a decision boundary is set for the probabilities to classify the predicted label (Lindholm et al., 2022). In contrast, classification trees rely on a set of rules to split the input features into multiple disjoint regions, where each region contains a constant value for predicting the label (Lindholm et al., 2022). To assess the performance and compare the predictions of the models, accuracy-, recall-, and precision scores of the predicted labels were established. This was achieved by creating confusion matrices, which tabulates the number of correctly predicted label versus incorrectly labels for each class (Lindholm et al., 2022).

Revisiting Table 3-1, it is evident that there was a class imbalance problem in the data. Specifically, *Recommendation* had approximately 60% observations categorised as ‘no’ and

40% categorised as 'yes'. Likewise, the *OverallScore* displayed and even more evident class imbalance problem. As a result of this, *OverallScore* was binned into three groups considered by the researcher as 'Low score', 'Neutral', and 'High score'. The groups consisted of scores between 1-3, 4-6, and 7-10 respectively. In addition, to optimize training the available data was split up into five equal parts to enable the previously mentioned *k*-fold cross validation to fine-tune the parameters of the models and minimize misclassification. This was performed with the goal of minimising the effect of overfitting, that is, a model which predicts training data well, but is unable to generalise its learning to new, unseen data (Lindholm et al., 2022).

(1) For logistic regression, *k*-fold cross validation was used to determine the optimal regularisation parameter λ. Regularisation is seen as useful methods for avoiding overfitting as they add a penalty term to the loss function, in this case the cross-entropy loss (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot & Duchesnay, 2011). Three models were trained, two with regularisation methods L1 (Lasso regularisation) and L2 (Ridge regularisation) respectively, and one model with no regularisation. L1 regularisation adds a penalty proportional to the absolute values of the features' coefficients, which effectively shrinks less important features to zero and removes them from the model. In contrast, the penalty of the L2 regularisation is the features' coefficients squared, effectively shrinking unimportant features towards zero, but never removes them completely from the model (Lindholm et al., 2022). These penalties are multiplied by a regularization parameter λ, which was optimised through cross-validation. The tested values of λ ranged from 0.5-1.5 in increments of 0.1. This range was determined after first running the model on three values: 0.01, 1, 10, where 1 resulted in lowest loss, consequently leading to the final range of 0.5-1.5. The coefficients of the third model with no regularisation naturally was not subject to a penalty term.

(2) For classification tress, *k*-fold cross validation was used to determine three parameters: the maximum depth of the tree, the minimum number of samples per split, and the minimum number of samples per leaf. In contrast to the cross-entropy in logistic regression, the parameters for the classification trees were optimised through accuracy scores. Additionally, the values tested for optimisation were 1, 2, and 3 for minimum number of samples per leaf, 2, 5, and 10 for minimum samples per split, and none, 5, and 10 for maximum depth.

Using the tuned parameters from cross validation, a final logistic regression model for each penalisation technique as well as for the classification tree was performed on the full data set using hold-out validation, where the data is split into training and testing set. An 80/20 split was utilized (Lindholm et al., 2022). The model was fit on the training set and the final comparison between the models' performance was assessed through precision, recall, and accuracy of predicted labels on the test set. Precision counts the number of correctly classified positive classes divided by the number of observations predicted as positive, whereas recall counts the number of correctly classified classes divided by the number of positive classes in the data (Sokolova & Lapalme, 2009). In addition to these, accuracy measures the number of correctly predicted classes divided by the number of observations, providing a measure of overall effectiveness (Sokolova & Lapalme, 2009). The consideration of the three scores allows for evaluation of imbalance problems. For the multiclass problem of *OverallScore*, the calculation of precision and recall was conducted based on the methods used by Sokolova and Lapalme (2009) in their research into performance measures of classification tasks, where the measures are calculated for each group individually and averaged. The models deployed in this study were carried out through Scikit-learn in the Python programming environment, and *k* was set to 5 (Pedregosa et al., 2011). Furthermore, a method to ensure reliable evaluation of the scores is to compare the accuracy measures to the probabilities of classes occurring respectively, that is, randomly guess a label (Sokolova & Lapalme, 2009). Taking into consideration the imbalance in the data, the evaluation metrics of accuracy scores in this study is presented in Table 3-5.

Table 3-5: Evaluation metrics of accuracy scores.

	Recommended		**Overall Score**		
Category	Yes	No	Low Score	Neutral	High Score
Count	51,622	77,009	63,527	18,872	46,232
Probability of "random guess" (%)	**40.13**	**59.87**	**49.39**	**14.67**	**35.94**

4. Results

4.1 Topic Identification

The optimal LDA model was obtained through a process of comparing various modifications of the model made possible by adjusting the different number of topics. This is

a vital step since deploying an LDA model to identify too few or too many topics in the corpus may lead to too general or overlapping latent topics. As a result, coherence scores were used as a measure for comparison between different LDA models as their range between [0,1] provides a clear and concise way for comparison, where the higher value presents a higher human-interpretability of the latent topic's words. The optimal model was found by (1) for a given number of topics, the data was partitioned into five equally sized parts; four parts for fitting the model with *t* topics, and one part for calculating the coherence scores on unseen data. (2) The partitioning, model fitting and coherence calculation was repeated five times for each $t = 2, \dots, 19$. (3) The average coherence score for each *t* was calculated and compared. Figure 4-1 illustrates the average coherence scores calculated for the range of different number of topics. They can be compared to the levels calculated by Rosner et al. (2014), whom achieved *any-any coherence* scores ranging from 0.52 – 0.55 for two different experiments on English Wikipedia articles. The final LDA model deployed in the remainder of this research yielded a coherence score of 0.5 across 18 topics.

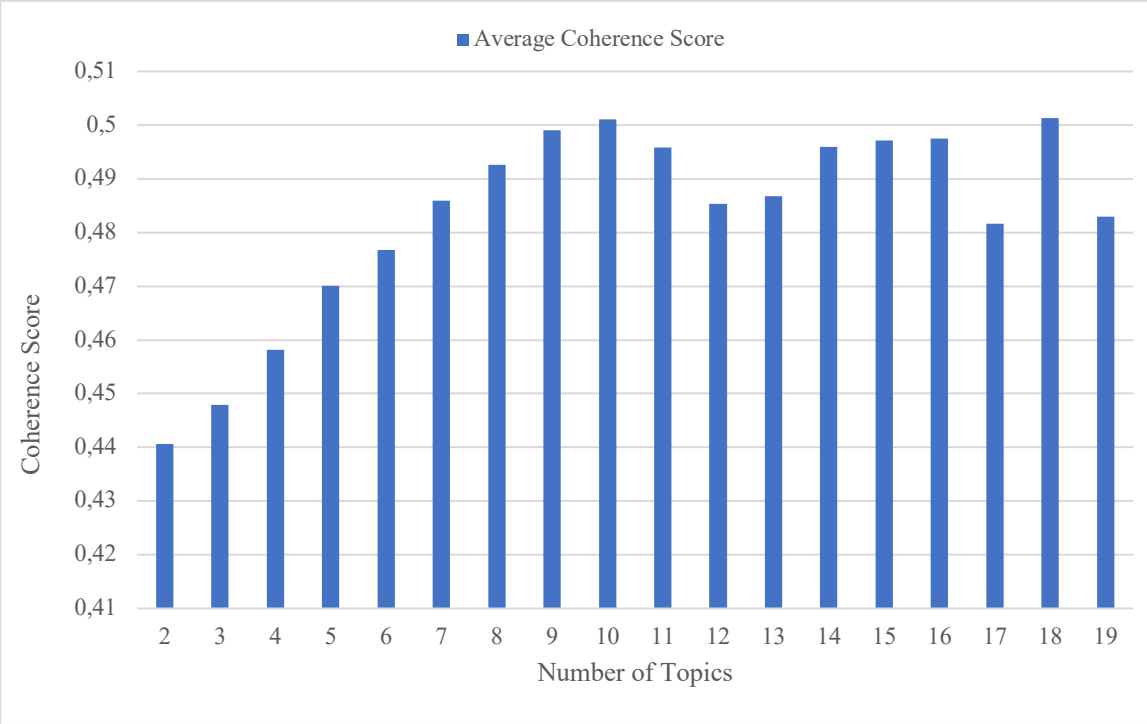

Figure 4-1: Average coherence scores for trained LDA models on different number of topics.

Next, each of the 18 latent topics underwent a labelling process to identify the dimensions of customer satisfaction. The words and their probability of occurring within the topic were logically connected to a term which would encompass their contextual meaning. Table 4-1 illustrates an example for two latent topics generated by the LDA model, with the

top-10 related words as well as their probability of occurring in the topic. The dimensions identified specifically for these two topics were “Inflight Experience” and “Luggage Handling”. Full raw data of identified dimensions can be viewed in Table 0-1 in Appendix D.

Table 4-1: Top 10 attributes for identified topic dimensions "Inflight Experience" and "Check-in bags" and their scores.

“Inflight Experience”	
Word	**Probability of occurrence (%)**
entertainment	6.5
meal	4.9
food	4.3
flight	3.7
inflight	3.4
system	2.7
screen	2.6
crew	2.3
poor	2.3
movie	2.2

“Luggage Handling”	
Word	**Probability of occurrence (%)**
luggage	12.5
bag	11.3
baggage	5.0
route	3.0
hand	1.8
extra	1.7
suitcase	1.6
organization	1.3
weight	1.3
fee	1.3

4.2 Distribution of Customer Satisfaction Dimensions

To find an answer to this study’s first objective of identifying key customer satisfaction dimension which travellers tend to express in OCRs, the next step was to aggregate the dimension distributions across the full corpus, i.e., all OCRs. This was accomplished by (1) generating each dimension’s probability of occurring in individual OCRs. (2) Aggregating the probabilities per dimension, and normalising by dividing with the total probability of all topics. Figure 4-2 illustrates the 18 identified dimensions and their probability of occurring in the full corpus. At a first glance, it is evident that the number one dimension travellers tend to write about is Travel Delays, which has an approximate probability of occurrence of 14%. Following this, Good Service is the second most prevalent topic, accounting for approximately 10% of occurrences. Efficiency, Seat Comfort, Airline Experience, and Bad Customer Service are the subsequent dimensions, occurring approximately 9%, 7.5%, 7.5%, and 6% respectively. These 6 topics can be identified as the predominant dimensions addressed by travellers since they collectively encompass a majority.

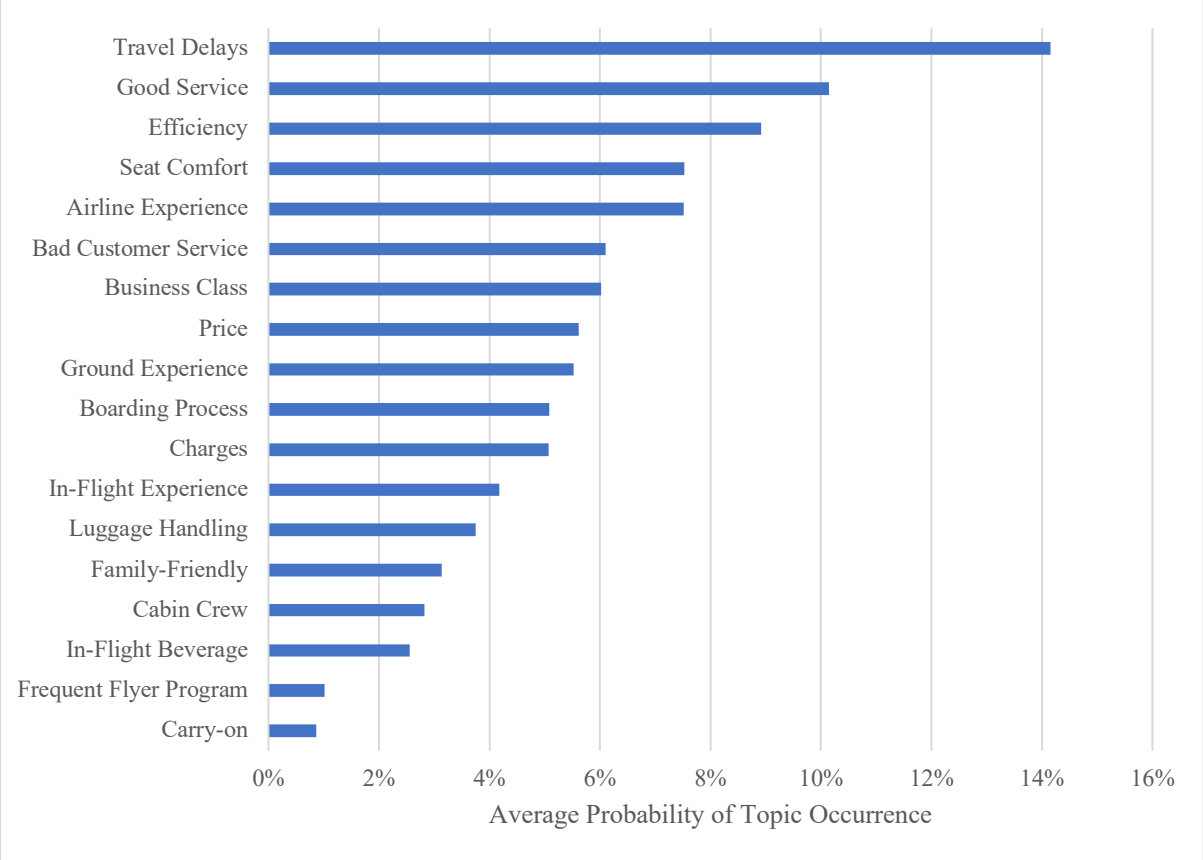

Figure 4-2: Satisfaction dimensions and their average probability of occurring in reviews. Format inspired by Lucini et al., (2020).

The next objective of this study was to evaluate what effect the pandemic had on travellers' reviews. To accomplish this, the dimensions, and their probabilities of occurring in the corpus, were grouped by published year. More specifically, to gain context of the historical development of the topics, the years were binned into three eras: 2002-2015, 2016-2019, and 2020-2023. These were considered as 'Digital era', 'Pre-pandemic era', and 'Post-pandemic era'. The box and whisker plots are illustrated in Figure 4-3. Comparing the interquartile ranges between the various dimension probabilities illustrates the varying spread of the probabilities over time. Firstly, dimensions which IQR spread have consistently exhibited an increase over the time periods are Airline Experience, Charges, Bad Customer Service, Luggage Handling, and Travel Delays. Notably, Charges and Bad Customer Service have shown a particular increase in the Post-pandemic era, whereas more prominent dimensions in the Pre-pandemic- and Digital eras include Business Class, Efficiency, and Good Service. Additionally, dimensions which has exhibited a consistent decreased IQR spread are Business Class, Cabin Crew, Efficiency, Good Service, Ground Experience, In-Flight Experience, Price, and Seat Comfort. The remainder of the dimensions have not had a noticeable change over the analysed time periods.

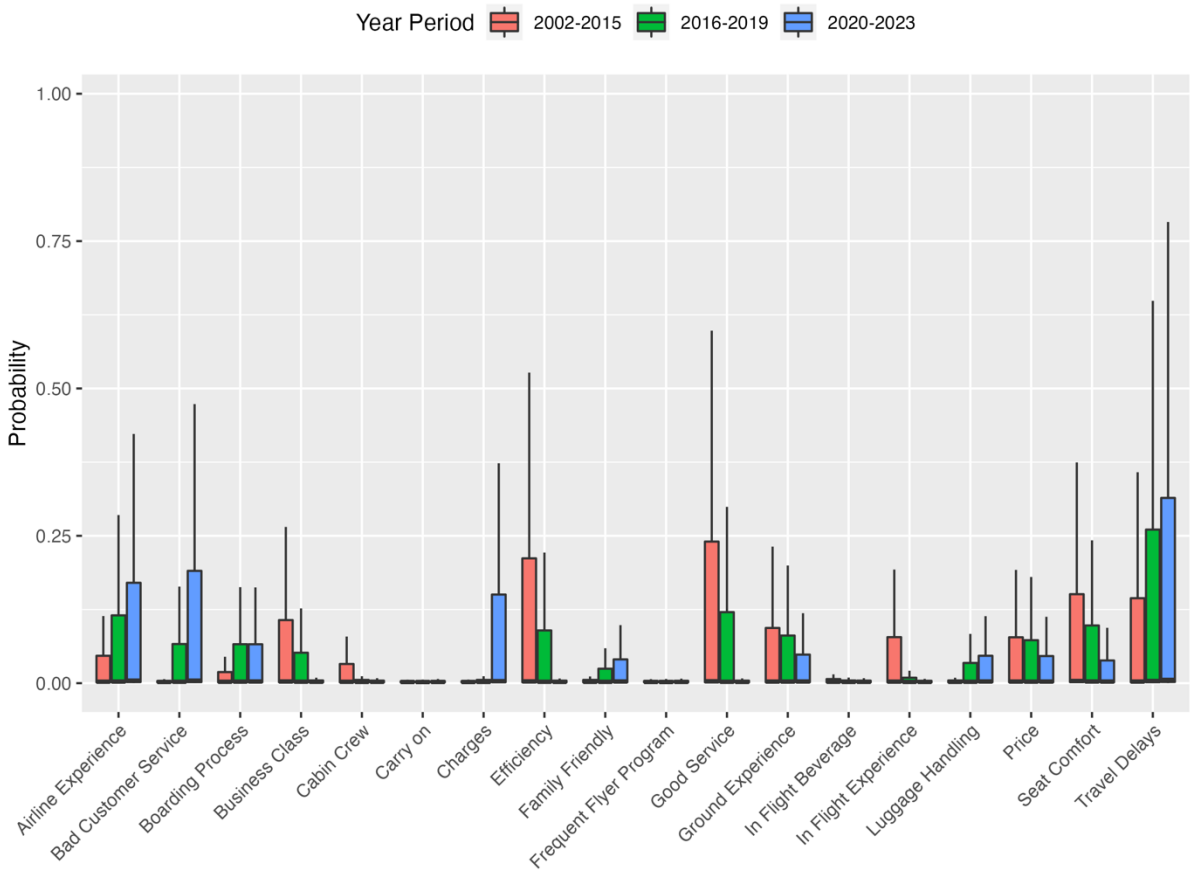

Figure 4-3: Distribution of dimension probabilities over different time periods, excluding outliers.

Provided that consumers' sentiments are shaped by their subjective perception of an experience (Roy, 2023), and that consumers express such experiences through OCRs (Mauri & Minazzi, 2013), it can be argued that the change in dimension probabilities over the three time periods suggests that there has been a shift in travellers' airline expectations and attitude towards flying. In contrast to the customer dimensions emphasized on in the Digital Era, Good Service and Efficiency, a notable shift of these can be seen in later years. In particular, greater emphasis has been put on Bad Customer Service and Travel Delays. Moreover, specifically Bad Customer Service and Charges have emerged as prominent customer dimensions in the Post-pandemic era, indicating their significance amongst travellers and suggesting that these areas warrant attention form airlines. The findings further suggest that Price has become a dimension which is slightly less of a concern, but on the other hand, Charges have increased in its place. Factors which may be attributed to this phenomenon include the introduction of new ticket categories at major airlines. In an attempt to compete with emerging low-cost carriers through the 2010's, major carriers introduced restricted low-fare tickets, which incorporated additional fees for check-in luggage and ticket changes (American Airlines, 2023; Air France, 2023; Lufthansa, 2023).

4.3 Predicting Recommendation

In exploring the predictability of customer dimension's effect on travellers' recommendation and overall rating of an airline, two machine learning methods were utilized with various modifications. Three different logistic regression models were used with varying regularisation methods. Firstly, L1 (Lasso regularisation) which adds a penalty term proportionate to the absolute values of the dimension coefficients and can lead to unimportant dimensions being completely disregarded by the model. Secondly, L2 (Ridge regularisation) was used, which adds a penalty term proportionate to the squared dimension coefficients. This can lead to dimensions having less of an influence in prediction, but it never disregards dimensions completely. Thirdly, a logistic regression model with no penalisation technique was also used. The strength of the penalty terms in L1 and L2 are controlled by parameter λ. The optimal value for λ was found through cross validation, where each model underwent training and testing of different subsets of the dataset, and cross-entropy loss was used to evaluate performance. The optimal value for λ was found to be 1 for L1 and 1.4 for L2 when predicting the binary target variable *Recommended*.

Similar to the parameter optimisation for the logistic regression models, classification trees also underwent parameter optimisation. In the context of classification trees, determining the size of the trees may present a challenge. As a result, parameters relating to maximum depth, minimum number of samples per split, and minimum number of samples per leaf were optimised through cross validation to determine an ideal size of the trees. These parameters were optimised to a max depth of 5, at least 1 sample per leaf, and at least 2 samples per split in predicting *Recommendation*.

Table 4-2 presents the precision, recall and accuracy scores for the different classification models employed in predicting travellers' recommendation of airlines using the dimension probabilities. Accuracy scores were attained well-above the probabilities of random sampling as previously established in Table 3-5. Although similar results across all methods, classification tree exhibited lowest performance. Conversely, logistic regression with L1 regularisation was deemed the best performing model.

Table 4-2: Precision, recall, and accuracy for four classification models of binary target variable Recommended.

		Precision (%)	Recall (%)	Accuracy (%)
Logistic Regression	L1 regularisation	**83.50**	**81.52**	**85.77**
	L2 regularisation	83.50	81.50	85.76
	No regularisation	83.50	81.50	85.76

Classification Tree	82.82	80.78	85.34

Reviewing the coefficients established by the L1 model indicates which dimensions influence the prediction of the classifications. The most relevant dimensions for predicting airline recommendation are presented in Table 4-3. Notably, the top three dimensions were Good Service (7.6), Efficiency (6.55) and Cabin Crew (2.92). In contrast, the bottom three dimensions were Bad Customer Service (-5.05), Travel Delays (-2.93), promptly followed by Charges (-2.84). The dimensions deemed by L1 as unimportant were Price and Frequent Flyer Program, both with coefficients of 0.0. These findings suggest that for an airline to be recommended, travellers put emphasis especially on Good Service and Efficiency. In contrast, Bad Customer Service in particular is a dimension which may lead to a negative recommendation.

Table 4-3: Coefficients of dimensions from modelling target variable Recommendation, generated by L1 regularisation.

Customer Dimension	Coefficient	Customer Dimension	Coefficient
Good Service	7.60	Bad Customer Service	-5.05
Efficiency	6.55	Travel Delays	-2.93
Cabin Crew	2.92	Charges	-2.84
Business Class	1.81	Luggage Handling	-2.68
Carry-on	1.46	Airline Experience	-2.61
Family-Friendly	0.64	In-Flight Beverage	-1.87
Ground Experience	0.51	In-Flight Experience	-1.79
Price	0.00	Seat Comfort	-1.64
Frequent Flyer Program	0.00	Boarding Process	-0.73

4.4 Predicting Overall Score

Training the multiclass classification models followed the same routine as utilized in the binary problem. Through cross validation, optimal value for penalising parameter λ was found to be 0.7 and 0.9 for L1 and L2 respectively. Additionally, the optimised splitting criterions for the decision tree were similar as the previous model: max depth of 5, at least 3 sample per leaf, and minimum 2 samples per split. The corresponding precision, recall, and accuracy measures are found in Table 4-4, where the precision and recall were calculated for each group individually and averaged. Although L2 regularisation displayed satisfactory precision and recall scores, the superior model yet again proved to be logistic regression with L1 regularisation with a considerably higher accuracy score.

Table 4-4: Mean precision & recall, and accuracy for four classification models of multiclass target variable OverallScore.

		Precision (%)	Recall (%)	Accuracy (%)
Logistic Regression	L1 regularisation	61.90	62.41	**74.76**
	L2 regularisation	**63.70**	**64.45**	69.78
	No regularisation	63.60	63.51	72.97
Classification Tree		62.00	61.90	70.50

Tabulating a confusion matrix in Table 4-5 with the predicted- and true labels reveals the performance of the model on individual classes. Recall shows what percentage of each class were correctly predicted. When comparing these to the evaluation matrices in Table 3-5, it can be observed that the model is unable to capture the relationship between topic probabilities and High Scores as the accuracy of 29.37% is unsatisfactory. Low Score and Neutral displayed satisfactory accuracies (77.89% and 79.95% respectively) when compared with the evaluation metrics.

Table 4-5: Performance of L1 regularisation on OverallScore.

	Low Score	Neutral	High Score
Precision	81.72%	87.22%	16.76%
Recall/Accuracy	77.89%	79.95%	29.37%

Plotting the coefficients generated by the L1 logistic regression for the multiclass *OverallScore* in a bar chart illustrates the vast difference between the importance of customer dimensions. In line with the coefficients of *Recommendation*, Good Service (-8.87), Efficiency (-7.79), and Cabin Crew (-3.74) are deemed as less important in predicting Low Scores. Conversely, these are the dimensions deemed as more important in Neutral Scores, with coefficients of 6.35, 5.25, and 1.8 respectively. Additionally, Bad Customer Service (4.20), Charges (2.63) and Luggage Handling (2.24) are given more influence in predicting a Lower Score, inversely of a Neutral Score. Since the accuracy of a High Score is unsatisfactory, these will not be reflected on.

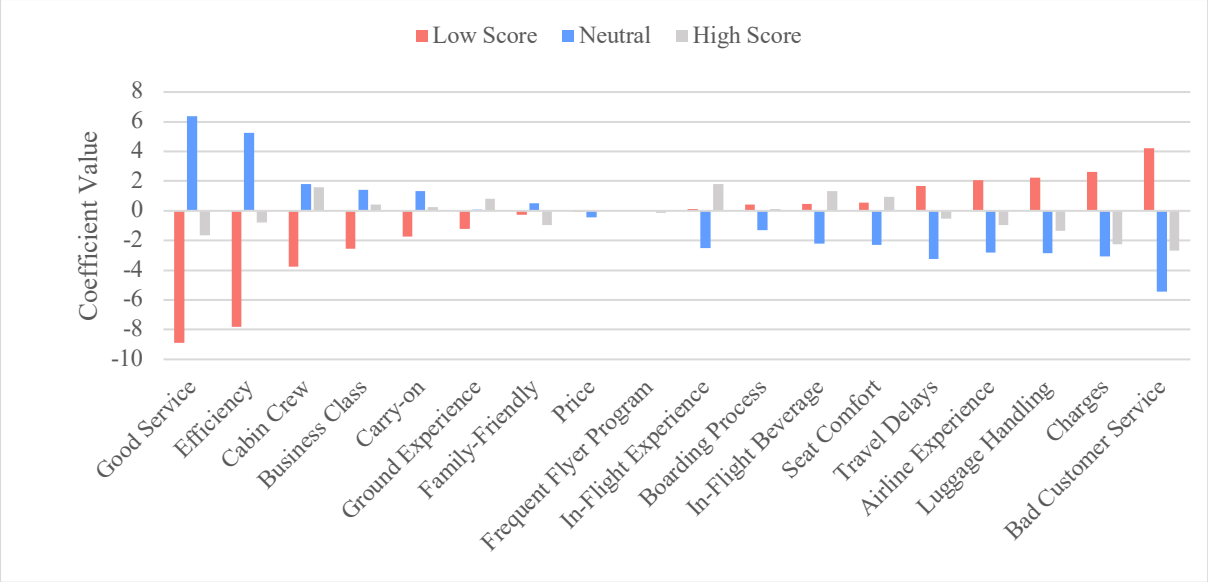

Figure 4-4: Coefficients of dimensions from modelling OverallScore, generated by L1 regularisation.

5. Discussion

Electronic Customer Reviews (OCRs) and electronic Word-of-Mouth (eWOM) presents a unique opportunity for airlines to identify key dimensions of customer satisfaction. Not only does eWOM create an ever-lasting impression for potential customers (for example, the oldest review in our data set was from 2002), but they also offer a source of information for consumers which is different from costly marketing campaigns (Mauri & Minazzi, 2013). This information can further be leveraged by Airlines to capture feedback that may not be available in their own customer feedback systems and in predicting recommendation. Thus, identifying customer satisfaction dimensions is a key for companies to evaluate how their consumers perceive their products and services.

This paper contributes to the literature of OCRs as a predictor for airline recommendation. Firstly, 18 distinct dimensions of customer satisfaction was successfully extracted from a large corpus of 128,631 reviews, where probabilities of occurrence illustrated meaningful insight for airlines. Specifically, the findings suggest the following:

(1) Travel Delays, Good Service, and Efficiency are the most wildly occurring customer satisfaction dimensions in OCRs. This is partly in line with what the topic analysis conducted by Lucini et al. (2020) found, whose study identified Customer Service, Flight Description, and Food & Drinks as the top proportions of satisfaction dimensions. However, contradicting proportions of satisfaction dimensions can also be identified. For example, Lucini et al. (2020) found that In-Flight Experiences (such as Onboard entertainment, onboard service, food & drinks) were among the highest probabilities of occurring, whereas

this study found the opposite of this. Additionally, it was found that these types of topics have declined in the post-pandemic era, suggesting that travellers are less meticulous about these experiences. This is also suggested by our next finding.

(2) Previous studies in predictive modelling of recommendation have suggested that airline ratings established prior to the pandemic had weaker implications during the pandemic (Lee & Leung, 2022). These indications are supported by this study, as Bad Customer Service and Charges were two dimensions which had emerged from the post-pandemic era, and dimensions such as Business Class and Efficiency showed signs of decline. However, other dimensions (e.g., Airline Experience and Boarding Process) contradicts the predictive models of Lee and Leung (2022) as these either showed an upward trend between the three eras established, or had a neutral difference between the years leading up to the pandemic and the years post-pandemic. Additionally, the best performing model to predict recommendation in this study had an accuracy score of 85.77%, in contrast to the study conducted by Lucini et al. (2020) whom attained an accuracy score of 79.95%.

(3) In line with common belief, this study identified Good Service, Efficiency and Cabin Crew as key determinants in positive airline recommendation, and Bad Customer Service, Travel Delays, and Charges as determinants in negative recommendation. Although this could not be confirmed in the analysis of overall ratings for High Scores, the Low Score category showed similar customer dimensions as drivers for a negative score. However, in contrast to common beliefs, both models predicting recommendation and overall score deemed price as an unimportant factor. Khudhair, Jusoha, Mardani and Nor (2019) found that the success of a pricing model depends on customers' level of sensitivity to changes in price. In other words, if consumers can justify higher prices for a higher level of service, they are unlikely to switch airlines, and vice versa (Khudhair et al., 2019). Therefore, this study suggests that in conjunction of the emergence of low-cost carriers, as previously discussed, consumers have widespread options and are less inclined to write about this dimension in reviews. Consumers possess a clear understanding of the value they are receiving in exchange for the amount they are willing to pay. In contrast, the results imply that when unforeseen charges arise during a trip, travellers are more inclined to express this in reviews.

5.1 Delimitations

This study presents a method for topic modelling of OCRs, and there are some delimitations which may restrict generalization of the results. They are presented as follows:

The sole method employed for topic modelling was chosen to be LDA, a method introduced in 2003 (Blei et al., 2003). The past few years have seen the advancement of new methods emerge, such as deep learning methods including Transformers (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, 2017). Such methods could have been applied and compared with LDA, which offers an opportunity for future research.

Pre-processing of free-written text can be a challenge, even with the usage of advanced named entity recognition, part-of-speech tagging, and replacement of common words. Given the large size of the corpora utilized in this study, it was impossible to ensure complete accuracy in pre-processing reviews. For example, words such as 'on-time' could be replaced by simply 'on-time', whereas the words 'on time' could be replaced by 'time', which may have implications on the performance of the LDA. This challenge can be optimized by using more rigorous methods for pre-processing text.

The research in this study was restricted to OCRs. As such, no relation to other conventional data gathering methods can be made, such as interviews, surveys, or questionnaires. Additionally, opinions gathered was limited to the English language, and as such, only opinions expressed in English was captured in the data set. Future studies may consider expanding the scope of data collection and the languages used.

Finally, restrictions of LDA naturally have its implications in this study as well. For example, LDA assumes that each topic is based on the latent topics discovered. Therefore, every topic has a joint probability of occurring in an OCR equal to 1. However, it would be misleading to state that all 128,631 OCRs are restricted to solely 18 topics. Additional dimensions that are not identified could be underrepresented in the analysis. Moreover, the labelling process of the latent topics may be seen as a flawed method, where some topics could potentially be overlapping, or perceived differently by different researchers. Therefore, future studies can attempt to not restrict the number of latent topics to be discovered and utilize more rigorous methods for identifying a common term for the latent topics.

6. Concluding summary

In this study, 128,631 Online Customer Reviews (OCRs) relating to airlines were scraped from a website specialized in the field called Air Travel Reviews (ATR). Through the employment of Latent Dirichlet Allocation (LDA), 18 latent topics were identified and labelled with various customer satisfaction dimensions relating to the words found in the topics. Probabilities of topic occurrence in each review were used to predict the

recommendation and overall scores of individual reviews using different modifications of logistic regression and classification trees. The findings suggests that there are multiple areas where airline executive have an opportunity to make managerial decisions and investment to improve the recommendation and reputation of their brands in web-based reviews.

References

Airlines, A. (2023). *Conditions of Carriage* [Online]. Available online: https://www.aa.com/i18n/customer-service/support/conditions-of-carriage.jsp?anchorEvent=false&from=footer? [Accessed 18 May 2023].

Blei, D. M., Ng, A. Y., Jordan, M. I. & Lafferty, J. (2003).Latent Dirichlet Allocation, *Journal of Machine Learning Research,* vol. 3, no. 4/5**,** pp 993-1022

Boubker, O. & Naoui, K. (2022).Factors Affecting Airline Brand Love, Passengers' Loyalty, and Positive Word-of-Mouth. A Case Study of Royal Air Maroc, *Case Studies on Transport Policy,* vol. 10, no. 2**,** pp 1388-1400

Calisir, N., Basak, E. & Calisir, F. (2016).Key Drivers of Passenger Loyalty: A Case of Frankfurt–Istanbul Flights, *Journal of Air Transport Management,* vol. 53, no. 211-217

Cosmina Laura, R., Maria, M. & Cristina, T. (2022).The Role of Service Quality in Ensuring Customer Satisfaction in the Airline Industry, *Ovidius University Annals: Economic Sciences Series,* vol. XXII, no. 1**,** pp 708-715

Emidi, C. & Galan, S. (2022). Prospectus Content as Predictor of Ipo Outcome: A Topic Model Approach. M.S.c Data Analytics & Business Economics Master Thesis, Lund University Available online: http://lup.lub.lu.se/student-papers/record/9083567 [Accessed 20 May 2023 Access 2022].

France, A. (2023). *Economy Class New Fare Options* [Online]. Available online: https://wwws.airfrance.co.jp/information/meilleures-offres/fare-option [Accessed 18 May 2023].

Gao, Y. & Koo, T. T. R. (2014).Flying Australia–Europe Via China: A Qualitative Analysis of the Factors Affecting Travelers' Choice of Chinese Carriers Using Online Comments Data, *Journal of Air Transport Management,* vol. 39, no. 23-29

Guo, Y., Barnes, S. J. & Jia, Q. (2017).Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation, *Tourism Management,* vol. 59, no. 467-483

Han, S. & Anderson, C. K. (2022).The Dynamic Customer Engagement Behaviors in the Customer Satisfaction Survey, *Decision Support Systems,* vol. 154, no.

Heiets, I., La, J., Zhou, W., Xu, S., Wang, X. & Xu, Y. (2022).Digital Transformation of Airline Industry, *Research in Transportation Economics,* vol. 92, no.

Hoffman, M., Blei, D. & Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. in: J. Lafferty, C. W., J. Shawe-Taylor, R. Zemel, A. Culotta (ed.) *Advances in Neural Information Processing Systems.* Curran Associates, Inc. pp

Honnibal, M. & Montani, I. (2017). Spacy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing.

IATA. (2021). *Worldwide Revenue with Passengers in Air Traffic from 2005 to 2022 (in Billion U.S. Dollars).* [Online]. Available online: https://www-statista-com.ludwig.lub.lu.se/statistics/263042/worldwide-revenue-with-passengers-in-air-traffic/ [Accessed May 13 2023].

IATA. (2022). *Number of Scheduled Passengers Boarded by the Global Airline Industry from 2004 to 2022 (in Millions).* [Online]. Available online: https://www-statista-com.ludwig.lub.lu.se/statistics/564717/airline-industry-passenger-traffic-globally/ [Accessed May 13 2023].

Khan, Q. & Chua, H. N. (2021). Comparing Topic Modeling Techniques for Identifying Informative and Uninformative Content: A Case Study on Covid-19 Tweets. IEEE.

Khudhair, H. Y., Jusoha, A., Mardani, A. & Nor, K. M. (2019).A Conceptual Model of Customer Satisfaction: Moderating Effects of Price Sensitivity and Quality Seekers in the Airline Industry, *Contemporary Economics,* vol. 13, no. 3**,** pp 283-292-292

Kiraci, K., Tanriverdi, G. & Akan, E. (2023). Analysis of Factors Affecting the Sustainable Success of Airlines During the Covid-19 Pandemic: SAGE Publications Ltd.

Lakshmanarao, A., Gupta, C. & Kiran, T. S. R. (2022). Airline Twitter Sentiment Classification Using Deep Learning Fusion. IEEE.

Lee, C. K. H. & Leung, E. K. H. (2022).Designing Predictive Models for Customer Recommendations During Covid-19 in the Airline Industry, *IEEE Transactions on Engineering Management, Engineering Management, IEEE Transactions on, IEEE Trans. Eng. Manage.,* vol. PP, no. 99**,** pp 1-11

Lindholm, A., Wahlström, N., Lindsten, F. & Schön, T. B. P. (2022). Machine Learning: A First Course for Engineers and Scientists. Cambridge, United Kingdom: Cambridge University Press.

Lucini, F. R., Tonetto, L. M., Fogliatto, F. S. & Anzanello, M. J. (2020).Text Mining Approach to Explore Dimensions of Airline Customer Satisfaction Using Online Customer Reviews, *Journal of Air Transport Management,* vol. 83, no.

Lufthansa. (2023). *Economy Light* [Online]. Available online: https://www.lufthansa.com/us/en/economy-light-fare [Accessed 18 May 2023].

Mauri, A. G. & Minazzi, R. (2013).Web Reviews Influence on Expectations and Purchasing Intentions of Hotel Potential Customers, *International Journal of Hospitality Management,* vol. 34, no. 99-107

Namukasa, J. (2013).The Influence of Airline Service Quality on Passenger Satisfaction and Loyalty : The Case of Uganda Airline Industry, *The TQM Journal,* vol. 25, no. 5**,** pp 520-532

Noh, Y.-G., Jeon, J. & Hong, J.-H. (2023).Understanding of Customer Decision-Making Behaviors Depending on Online Reviews, *APPLIED SCIENCES-BASEL,* vol. 13, no. 6**,** pp 3949

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011).Scikit-Learn: Machine Learning in Python, *Journal of Machine Learning Research,* vol. 12, no. 2825-2830

Ponay, C. S. (2022). Topic Modeling on Customer Feedback from an Online Ticketing System Using Latent Dirichlet Allocation and Bertopic. IEEE.

Poushneh, A. & Rajabi, R. (2022).Can Reviews Predict Reviewers' Numerical Ratings? The Underlying Mechanisms of Customers' Decisions to Rate Products Using Latent Dirichlet Allocation (Lda), *Journal of Consumer Marketing,* vol. 39, no. 2**,** pp 230-241

Řehuřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks.* ELRA pp 45-50.

Rita, P., Moro, S. & Cavalcanti, G. (2022).The Impact of Covid-19 on Tourism: Analysis of Online Reviews in the Airlines Sector, *Journal of Air Transport Management,* vol. 104, no.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M. & Both, A. (2014). Evaluating Topic Coherence Measures.

Roy, G. (2023).Travelers' Online Review on Hotel Performance – Analyzing Facts with the Theory of Lodging and Sentiment Analysis, *International Journal of Hospitality Management,* vol. 111, no.

Scrapy. (2023). *Scrapy 2.9 Documentation* [Online]. Available online: https://docs.scrapy.org/en/latest/ [Accessed April 15 2023].

Sokolova, M. & Lapalme, G. (2009).A Systematic Analysis of Performance Measures for Classification Tasks, *Information Processing and Management,* vol. 45, no. 4**,** pp 427-437

Tirunilla, S. & Tellis, G. J. (2014).Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation, *Journal of Marketing Research,* vol. 51, no. 4**,** pp 463-479

Uthirapathy, S. E. & Sandanam, D. (2023).Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using Lda and Bert Model, *Procedia Computer Science,* vol. 218, no. 908-917

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need.

Wan, Y. & Gao, Q. (2015). An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. IEEE.

Wang, R., Shu, L., Hsu, Lin, Y. H. & Tseng, M.-L. (2011).Evaluation of Customer Perceptions on Airline Service Quality in Uncertainty, *Procedia - Social and Behavioral Sciences,* vol. 25, no. 419-437

Wang, W., Feng, Y. & Dai, W. (2018).Topic Analysis of Online Reviews for Two Competitive Products Using Latent Dirichlet Allocation, *Electronic Commerce Research and Applications,* vol. 29, no. 142-156

WHO. (2020). Who Director-General's Opening Remarks at the Media Briefing on Covid-19 - 11 March 2020. 11 March 2020. in*:* Organization, W. H. (ed.). who.int: World Health Organization.

Wickham, H. (2016). Ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag.

Yao, B., Yuan, H., Qian, Y. & Li, L. (Year) Published. On Exploring Airline Service Features from Massive Online Review. 2015 12th International Conference on Service Systems and Service Management (ICSSSM), 2015/01/01/ 2015 Place of Publication: Piscataway, NJ, USA; Guangzhou, China. Country of Publication: USA.: IEEE.

Zeng, J., Cheung, W. K. & Liu, J. (2013).Learning Topic Models by Belief Propagation, *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. 35, no. 5**,** pp 1121-1134

APPENDIX A.

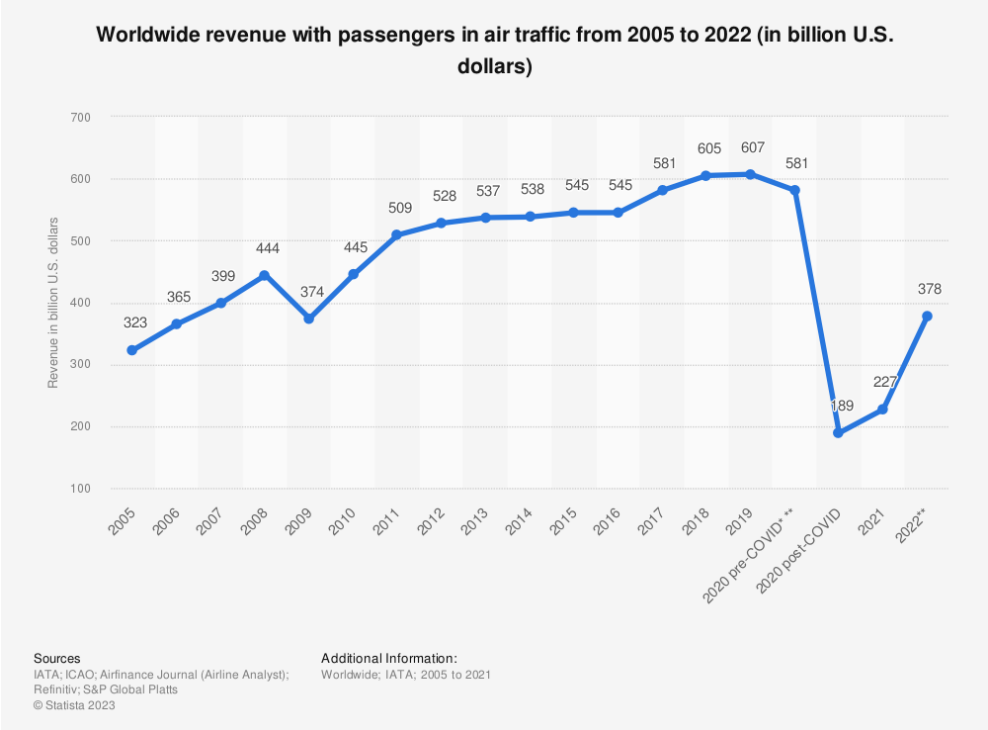

Figure 0-1: Worldwide revenue with passengers in air traffic from 2005 to 2022 (in billion U.S. dollars) (IATA 2021).

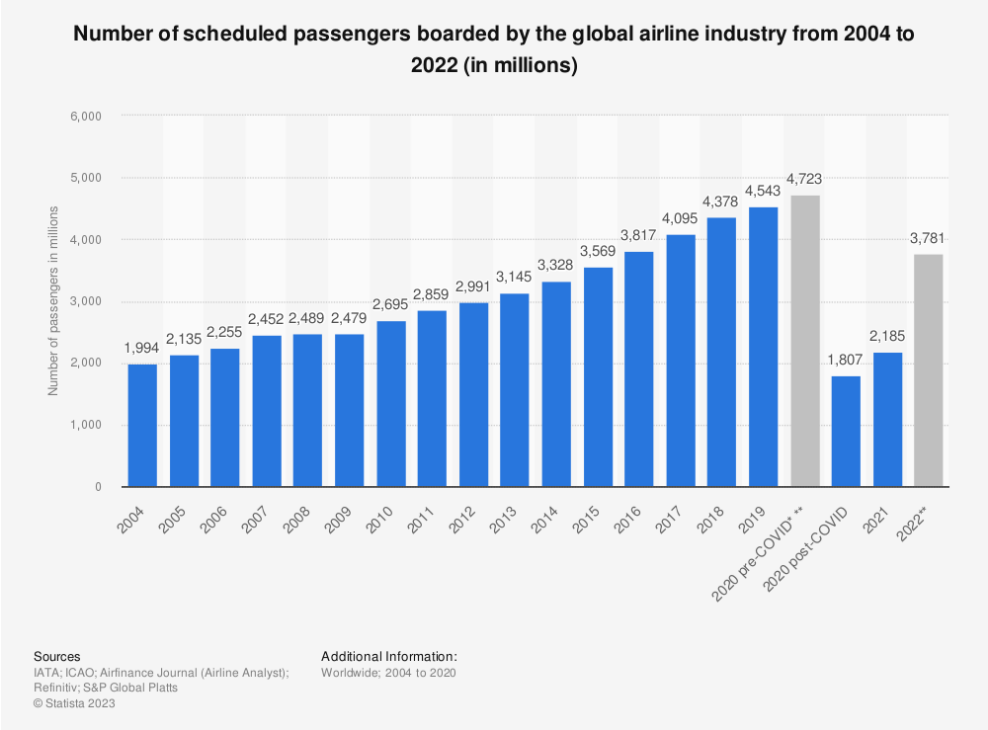

Figure 0-2: Number of scheduled passengers boarded by the global airline industry from 2004 to 2022 (in millions) (IATA 2022).

APPENDIX B.

Table 0-1: Explanatory list of all available variables at ATR.

Variabel Name	Description	Data type
AirlineName	Name of the airline	String
Title	Title of review	String
OriginCountry	Origin country of reviewer	String
DatePub	Date review was published	Date
DateFlown	Date of flight(s)	Date
Review	Customer review	String
Aircraft	Aircraft equipment flown on	String
TravelType	Type of travel	Categorical (Solo Leisure/Couple Leisure/Family Leisure/Business)
CabinType	Cabin type	Categorical (Economy/Premium Economy/Business/First)
Route	Route flown	String (example: Stockholm to Miami via Copenhagen)
OverallScore	Rating for overall experience	Integer, 1-10
SeatComfortRating	Rating of seat comfort	Integer, 1-5
ServiceRating	Rating of on-board service	Integer, 1-5
FoodRating	Rating of on-board food and beverages	Integer, 1-5
EntertainmentRating	Rating of in-flight entertainment (IFE)	Integer, 1-5
GroundServiceRating	Rating of ground service	Integer, 1-5
WifiRating	Rating of wifi connectivity	Integer, 1-5
ValueRating	Rating of value for money	Integer, 1-5
Recommended?	Recommend the airline	Binary, Yes/No
TripVerified	If the trip has been verified by ATR or not	Binary, Yes/No
unique_id	A unique string-code appended to each review for easy identification	String

APPENDIX C.

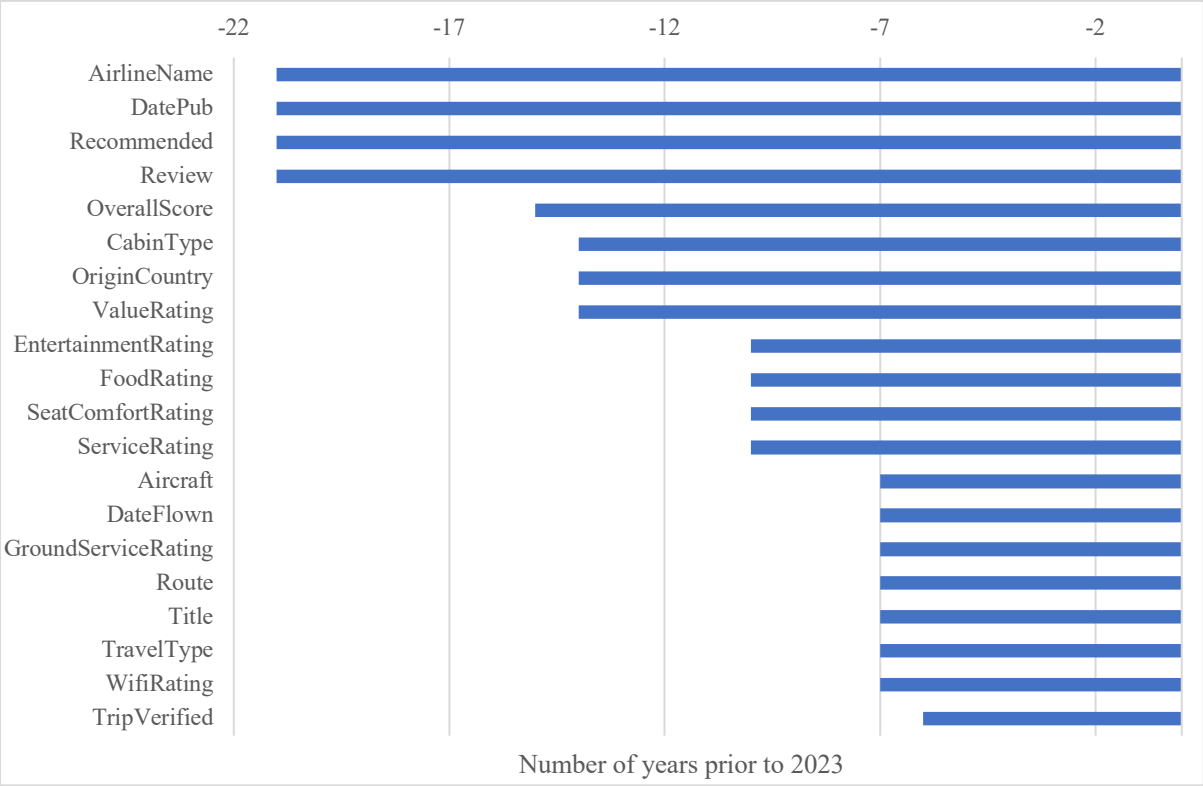

Figure 0-1: Number of years prior to 2023 which individual variables have been a feature of ATR.

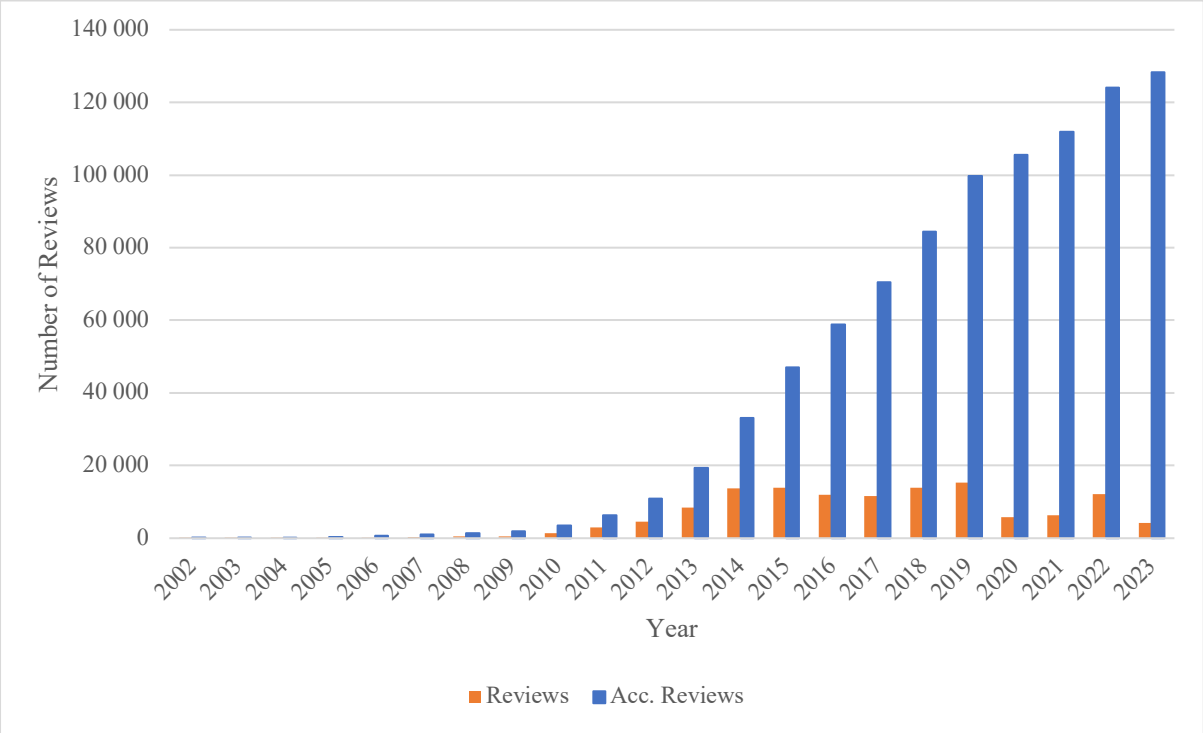

Figure 0-2: Number of new OCRs uploaded to ATR (yearly) and accumulation of OCRs over time.

APPENDIX D.

Table 0-1: Identified dimensions and top 10 words with highest probability of occurring in the topic.

Identified Dimension	Words & Probability of Occurring in Topic
Travel Delays	'0.134*"flight" + 0.067*"route" + 0.064*"hour" + 0.031*"time" + 0.029*"plane" + 0.024*"airport" + 0.022*"day" + 0.020*"delay" + 0.016*"next" + 0.014*"hotel"'
Luggage Handling	'0.125*"luggage" + 0.113*"bag" + 0.050*"baggage" + 0.030*"route" + 0.018*"hand" + 0.017*"extra" + 0.016*"suitcase" + 0.013*"organization" + 0.013*"kg" + 0.013*"fee"'
Good Service	'0.067*"good" + 0.044*"route" + 0.040*"flight" + 0.033*"food" + 0.031*"service" + 0.031*"great" + 0.028*"seat" + 0.025*"crew" + 0.025*"comfortable" + 0.025*"organization"'
Efficiency	'0.107*"flight" + 0.069*"route" + 0.041*"time" + 0.032*"good" + 0.028*"friendly" + 0.018*"crew" + 0.017*"service" + 0.016*"nice" + 0.015*"short" + 0.015*"staff"'
Price	'0.077*"airline" + 0.032*"price" + 0.030*"flight" + 0.029*"cost" + 0.028*"cheap" + 0.025*"low" + 0.023*"extra" + 0.023*"organization" + 0.020*"time" + 0.014*"money"'
In-Flight Beverage	'0.087*"water" + 0.052*"drink" + 0.034*"flight" + 0.028*"coffee" + 0.025*"hour" + 0.024*"snack" + 0.023*"meal" + 0.022*"bottle" + 0.021*"food" + 0.021*"hot"'
Bad Customer Service	'0.083*"customer" + 0.070*"service" + 0.048*"flight" + 0.037*"airline" + 0.024*"phone" + 0.023*"bad" + 0.023*"email" + 0.023*"day" + 0.023*"time" + 0.019*"call"'
Carry-on	'0.135*"free" + 0.073*"charge" + 0.055*"mile" + 0.049*"rule" + 0.041*"allowance" + 0.040*"baggage" + 0.036*"kg" + 0.032*"duty" + 0.028*"carryon" + 0.020*"bin"'
Frequent Flyer Program	'0.047*"frequent" + 0.043*"app" + 0.040*"flyer" + 0.028*"device" + 0.019*"power" + 0.019*"status" + 0.018*"wifi" + 0.016*"tablet" + 0.016*"program" + 0.016*"code"'
Charges	'0.075*"ticket" + 0.069*"flight" + 0.035*"refund" + 0.022*"day" + 0.021*"money" + 0.021*"organization" + 0.020*"change" + 0.019*"credit" + 0.018*"fee" + 0.014*"month"'
In-Flight Experience	'0.065*"entertainment" + 0.049*"meal" + 0.043*"food" + 0.037*"flight" + 0.034*"inflight" + 0.027*"system" + 0.026*"screen" + 0.023*"crew" + 0.023*"poor" + 0.022*"movie"'
Cabin Crew	'0.120*"crew" + 0.091*"cabin" + 0.039*"passenger" + 0.036*"aircraft" + 0.034*"flight" + 0.021*"outbound" + 0.017*"announcement" + 0.016*"time" + 0.013*"safety" + 0.012*"return"'
Seat Comfort	'0.155*"seat" + 0.033*"flight" + 0.026*"plane" + 0.022*"row" + 0.019*"leg" + 0.015*"front" + 0.015*"room" + 0.013*"extra" + 0.013*"economy" + 0.013*"uncomfortable"'
Family-Friendly	'0.035*"attendant" + 0.034*"family" + 0.030*"child" + 0.027*"flight" + 0.026*"year" + 0.026*"old" + 0.020*"staff" + 0.019*"son" + 0.018*"kid" + 0.018*"daughter"'
Business Class	'0.071*"class" + 0.060*"business" + 0.048*"route" + 0.028*"service" + 0.026*"economy" + 0.025*"lounge" + 0.023*"food" + 0.020*"flight" + 0.020*"good" + 0.013*"first"'
Boarding Process	'0.059*"boarding" + 0.042*"gate" + 0.036*"check" + 0.032*"bag" + 0.026*"line" + 0.024*"minute" + 0.023*"airport" + 0.023*"pass" + 0.020*"checkin" + 0.019*"staff"'
Airline Experience	'0.066*"flight" + 0.051*"organization" + 0.033*"airline" + 0.033*"route" + 0.025*"staff" + 0.025*"service" + 0.021*"customer" + 0.019*"time" + 0.017*"experience" + 0.017*"bad"'
Ground Experience	'0.288*"route" + 0.043*"flight" + 0.026*"airport" + 0.021*"bus" + 0.018*"hour" + 0.015*"return" + 0.014*"trip" + 0.014*"staff" + 0.014*"organization" + 0.011*"connection"'