Topic Modelling of Online Reviews for Airports

MSc Research Project
Data Analytics

Dona Elizabeth John
Student ID: x21228531

School of Computing
National College of Ireland

Supervisor: Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Dona Elizabeth John
Student ID:	x21228531
Programme:	Data Analytics
Year:	2022-2023
Module:	MSc Research Project
Supervisor:	Catherine Mulwa
Submission Due Date:	14/08/2023
Project Title:	Topic Modelling of Online Reviews for Airports
Word Count:	9789
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	DONA ELIZABETH JOHN
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Topic Modelling of Online Reviews for Airports

Dona Elizabeth John
x21228531

Abstract

This study used topic modeling approaches to successfully categorize travelers' reviews of major airports around the world by department. The goal was to distribute these categories to appropriate department managers in order to gain a better knowledge of the traveler experience and suggest areas for improvement within airport departments. The existing research highlights the significance of understanding the traveler experience and how airports influence it. Prior to this effort, however, there has been no study focused on reviewing traveler opinions to determine specific areas for improvement in various airport departments. This work bridges that gap by employing topic modeling, particularly by utilizing qualitative and quantitative methodologies such as statistical data analysis and topic modeling techniques such as Latent Dirichlet Allocation (LDA), CountVectorization, and TF-IDF. Topic modeling was used to discover major themes significant to different airport departments in accordance with traveler reviews. The identified topics were: "Parking desk", "Travel desk" and "Outlet desk". The coherence score resulted for the LDA model using CountVectorization was 0.553. The study gained important insights into particular areas for development throughout various airport divisions by taking this method. These information are expected to help airport management handle traveler problems and improve the overall experience for travelers. The study's findings not only help to improve services and raise airport standards, but they also establish the groundwork for developing a comprehensive system for airport administration to efficiently solve traveler difficulties. Airport management can proactively improve their services and prioritize areas that demand attention by taking into account the identified concerns and employing the results of topic modeling.

1 Introduction

The global economy's travel and tourism sectors are substantially reliant on the aviation sector. Airports have been critical components of the aviation sector. Airports have controlled and facilitated passenger traffic, commodities transfer, and aircraft operations. Airport management is under increasing pressure to provide good services in order to meet rising demand as more people fly throughout the world.

However, many travelers have discovered that traversing the airport's multiple departments which includes security, immigration, baggage claim, lounges, etc. is time-consuming and complicated, leaving them annoyed and disappointed. Airports identified areas for improvement and took appropriate measures to improve the overall quality of their facilities and the customer experience.

This study attempted to address the question of how to measure traveler reviews of airports around the world in order to identify areas that required growth across various airport departments. The method used was to use topic modeling to aggregate reviews of specific departments inside each airport and then give them in bulk to those departments' management. The airport management has profited from the insightful information provided by this strategy, allowing them to improve both the overall standard of the airport and overall satisfaction for travelers.

1.1 Motivation and Background

The existing research emphasizes the necessity of understanding the traveler experience and how airports shape it. Previous studies are primarily focused on how airport factors such as layout, cleanliness, and accessibility affect travelers' experiences. However, there has been a scarcity of researches that focuses specifically on measuring traveler attitudes in order to highlight exact areas for development in various airport departments. As a result, the purpose of this study was to bridge that gap by using topic modeling to identify relevant issues linked with various airport departments based on reviews from travelers.

1.2 Project Requirement Specifications

1.2.1 Research Question

The research question tackled in this study was:

RQ: "How can traveler reviews of airports around the world be analyzed in order to determine the topics concerning various airport departments?"

Sub-RQ1: "What are the different topics that are evident from traveller reviews of various airports all over the world?"

Sub-RQ2: "How can these topics be clustered to corresponding departments within the airport?"

1.2.2 Objectives of the Research

The main objective of this project is to analyze and examine traveler reviews based on airports utilizing text mining techniques, specifically using LDA topic modeling using CountVectorization. Based on the knowledge gained from online airport reviews, the study helps to pinpoint the main sources of passenger happiness and disappointment. The project focuses on the following particular goals.

Objective A was to select a dataset specifically for airport reviews. The dataset was then preprocessed and transformed in order to make it ready for the modelling purpose. **Objective B** was to perform TF-IDF Vectorization. **Objective C** was to create LDA model using the TF-IDF vectors. The model was evaluated using various methods. **Objective D** was to implement CountVectorization. **Objective E** was LDA model creation using CountVectorization vectors. Similar to TF-IDF, various evaluation methods were performed for this model as well. **Objective F** was to combine the similar generated topics and **Objective G** was to categorize the generated final topics and the reviews that come under those topics were grouped in order to send to corresponding airport departments.

1.2.3 Contributions of the Research

The proposed solution was to use a mixed-methods approach that included qualitative and quantitative data analysis to identify specific areas in various airport departments that needed improvement and to create a framework for efficient communication between travelers and airport management.

This project report describes the implementation and findings of the research project focusing on using the KDD technique to turn data into actionable insights for influencing business choices and improving outcomes. The project's specific goal is to assess airport reviews using topic modeling, exploiting the wealth of data available in multiple areas.

The first phase in the study was to choose the data to be studied, taking into account the research objectives and available resources. Airport reviews were recognized as a useful data source for topic modeling. Preprocessing techniques were used to eliminate insignificant information from the dataset, such as dates, names, and metadata, in order to prepare it for analysis.

Creating a CountVectorization model necessitated a number of preprocessing processes. Word-level tokenization was used for tokenization, which includes translating text into numerical tokens. This approach divides the text into separate words and assigns a unique numerical ID to each word, allowing for additional analysis and modeling.

The transformation stage was critical in improving the overall quality and precision of the final model. Adequate preparation and modification of the text data were required for the topic modeling method to discover significant and relevant themes that properly reflected the corpus's content. The CountVectorization method was used by the project to convert the text data into a numerical representation appropriate for topic modeling. Each word or token found in that document is counted by CountVectorizer to determine its frequency. These counts are then kept in the respective feature matrix rows. The weights for the words in the feature matrix are based on the counts. The entry in the matrix that corresponds to a word that appears more than once in a document will have a larger count value.

The proposed solution has helped in two ways. First, the study identified specific areas for improvement across a variety of airport departments in order to improve the traveler experience and elevate airport standards. Second, the study has contributed to the development of a method that airport management may use to engage with and successfully handle the difficulties of travelers. Airport officials can use the research findings to identify problem areas and improve the quality of services offered to travelers

This report will go over the literature review, the research techniques used, the implementation process, the model evaluation,the results involved, discussion and the future work.

2 Literature Review

A growing number of people are taking an interest in using online traveler feedback to assess the quality of services provided by various sectors of transportation services, such as travel agencies, lodging places, and airports. This section looks at previous studies on the use of online travelers reviews to evaluate the quality of airport services, identify areas for development, and improve the overall traveler experience.

2.1 Similar topic modelling researches related to Airline Industry

A strategy for categorizing airport service reviews based on their topics and polarities was put out by Mizufune and Katsumata (2018). The objective of the study was to identify the areas of airport services that have an impact on both passenger happiness and dissatisfaction. In order to categorize a dataset of 18,000 airport service reviews in Japan into topics and polarity, the study employed a combined classification model. Airport services, employee services, security, and transportation were among the topics covered. There were positive, negative, and neutral polarities. The data was preprocessed by the researchers using stop-word removal, feature selection, and stemming techniques in natural language processing. They then used a logistic regression model to classify the reviews based on their polarities and topics. According to the findings of the study, airport amenities and hospitality were the two factors most strongly influencing traveler happiness. The two main variables that had the biggest effects on traveler dissatisfaction were security and transportation. The study also discovered that particular issues like lengthy waits at security checkpoints and crowded airport terminals were causing irritation among travelers. A way for analyzing customer feedback on airport services is provided by the study, which could help to raise the level of service. Airport management can use the findings to pinpoint areas that need to be changed in order to increase traveler happiness. With an accuracy percentage of 80.5%, the suggested combined classification model was successful in correctly classifying the reviews. Topic and polarity classification had strong precision and recall scores. The classification of topics and polarities performed well according to the F1 score, which is a measure of overall performance.The result of the research indicate that this combined classification model can be used to improve airport services by gathering insightful information from customer reviews.

A study similar to this by Hasib et al. (2021) used topic modeling and sentiment analysis to examine internet reviews of Bangladesh Airlines. 1,000 reviews were gathered by the researchers, who then utilized LDA to categorize them into eight primary topics: service quality, in-flight and ground services, ticketing, safety procedures, staff behavior, and flight delays. With regard to flight delays, staff behavior, and in-flight facilities, sentiment analysis using SVM revealed that the majority of ratings were unfavorable or negative. The findings from the research can be applied to various industries and can aid Bangladesh Airlines in increasing customer satisfaction. The accuracy of the suggested model for sentiment analysis was 83.3%, with precision and recall scores for positive, negative, and neutral feelings of 0.86, 0.79, and 0.78, respectively. The F1 score for sentiment analysis was 0.85 for positive sentiment, 0.80 for negative sentiment, and 0.83 for neutral sentiment.

Topic modeling and sentiment analysis of online evaluations from the Skytrax platform (airlinequality.com) are the main focuses of the work by Kwon et al. (2021). More than 14,000 online comments from 27 Asian airlines were gathered for the study. The research tries to identify important topics and sentiments stated in these reviews in order to comprehend customer preferences, demands, and satisfaction levels. According to the study, customers prioritize in-flight meals, entertainment, seat class, seat comfort, and staff service when choosing an airline . Positive feelings like good, comfortable, great, friendly, excellent, nice, clean, helpful, and pleasant are more common in reviews than negative emotions like return, delayed, late, didn't, poor, and bad as per the study. In order to increase consumer satisfaction and loyalty, the study advises airlines to concentrate on these areas.

Using text mining techniques, a study by Lucini et al. (2020) provided a unique model for measuring passenger satisfaction in the airline sector. More than 55,000 online reviews from 400 airlines and travelers from 170 countries were reviewed for the study. The study found 882 adjectives to characterize the 27 dimensions of satisfaction. With an accuracy of 79.95%, the study's LDA model was employed to forecast airline recommendation for customers. To forecast airline recommendations, a logistic regression classifier was also utilized. According to the study, cabin crew, onboard service, and value for money were the factors that affected airline recommendation prediction the most. The study also discovered that the type of passenger and the type of cabin flown influenced various aspects of customer satisfaction. The results of the study show that in order to increase satisfaction among customers, airline service providers should concentrate on comfort, value for money, and customer service.

The service quality views of full service carriers (FSCs) and low cost airlines (LCCs) in the airline industry were compared using LDA model in a study by Lim and Lee (2020). According to the study, passengers felt that FSCs provided much greater levels of in-flight, ground, food and beverage, seat comfort, and on-time performance than did LCCs. Passengers' opinions of value for money were, nevertheless, substantially higher for LCCs than for FSCs. The study concludes by saying FSCs should prioritize raising customer views of service quality by enhancing their in-flight, ground, food and beverage, seat comfort, and on-time performance. LCCs should concentrate on increasing value for money in order to draw in and keep customers.

The above literature review includes a number of research that investigate the issue of examining customer feedback in the airline sector to comprehend different aspects of passenger satisfaction and disappointment. These studies use topic modeling and sentiment analysis to categorize reviews, pinpoint the most important satisfaction factors, and provide suggestions for enhancing airline services. To analyze customer evaluations and extract insightful information from unstructured text data, all examined research make use of topic modeling and sentiment analysis approaches. The main goals of the researches are to pinpoint the major issues or factors that affect passengers' satisfaction and dissatisfaction, as well as sentiments connected to those issues. By addressing certain issues raised in the evaluations, the studies offer recommendations for airlines to improve customer satisfaction. All studies analyze customer reviews, however they vary in dataset size, source, and the specific airlines or geographic areas they cover. There are variances in the precise topics recognized as essential based on the study's setting and dataset, even though some topics like in-flight services, personnel behavior, and safety protocols emerge consistently across the research. Together, these studies provided insightful information that was very helpful to the research on topic modeling of online airport reviews. They offered a strong framework and comprehension of the techniques used for categorizing reviews according to sentiments and topics. Additionally, the studies provide advice on the particular factors that affect customers' satisfaction and dissatisfaction in the aviation sector.

2.2 Topic Modeling Researches using NLP in LDA Model

Researchers Choi and Joo (2019) applied topic modeling usind LDA to assess online book reviews. They discovered seven major topics: plot, characters, writing style, setting, narration, genre, and theme. These topics give light on the aspects of works that critics usually emphasize. The researchers did not provide any detailed metrics. Instead, they used a qualitative evaluation approach to assess the coherence and interpretability of the highlighted topics. They manually checked the reviews as well as the associated topics to ensure that each was logical and understandable. The researchers came to the conclusion that the topic modeling method was effective in extracting meaningful topics from book reviews. They noted that the found topics were rational and clear, and that the data revealed a bias toward positive sentiment in the reviews.

In contrast, Saranya and Geetha (2020) offers a method for creating word clouds from apparel reviews in their study. They identified topics that would be of interest in apparel reviews using the Latent Dirichlet Allocation (LDA) topic modeling technique. The LDA model was trained using a dataset of online apparel reviews. Among the reviews, researchers identified six major topics: fit, quality, fabric, cost, design, and delivery. They then produced word clouds for every topic, highlighting the most frequently used words in the reviews. The word clouds are a simple and fast way to summarise the important ideas and viewpoints expressed in the reviews. The authors compared the produced word clouds to manual review summaries to assess their quality. The study found that the word clouds were effective at capturing the major concepts and viewpoints expressed in the reviews. The proposed approach can assist apparel retailers and manufacturers in learning more regarding consumer preferences and improving the standard of their products and services.

The researchers Alam et al. (2020) in their paper presents an approach for trend observation in Bangla news using topic modeling. To find topics in a corpus of Bangla articles, the authors apply LDA. The evolution of news trends through time is then tracked using these topics. The researchers use a corpus of 100,000 Bangla articles to evaluate their methodology. They demonstrate how well their algorithm can locate topics in the corpus. They also demonstrate how their approach may be used to track how topics change over time. This method might be used to monitor the news coverage of particular events, track the evolution of public opinion, and identify upcoming news topics. The topic modeling discipline has benefited from the paper's numerous contributions. It presents a technique for topic modeling in Bangla, a language that hasn't received much attention in this context. It also demonstrates how LDA may be utilized to track the evolution of topics through time. In addition, it offers a framework for spotting trends in news coverage.

Another research by Ashok et al. (2022) offer a novel deep learning model for predicting the toxicity of social media comments. They claim that current machine learning models are ineffective in detecting harmful comments, and they suggest a novel strategy based on a neuro-nlp architecture that integrates neural networks and natural language processing (NLP) approaches. The model was trained using a dataset of comments categorized as harmful or non-toxic. The results proved that the model predicted toxicity with an accuracy of 88%.The proposed approach is made up of three major parts: an attention mechanism, a neural network layer and a word embedding layer. The attention mechanism serves to bring attention to the most relevant portions of the input data, hence improving the model's accuracy. The neural network layer consists of numerous convolutional and recurrent layers which study the properties of the input data and identify the presence of toxicity. The word embedding layer turns the textual data to a numerical format which the neural network can interpret. The researchers tested their approach on a publicly available dataset of harmful comments. Comparing their model to numerous cutting-edge models, they discovered that it performed better than the other models in terms of accuracy. They also performed a sensitivity analysis to establish the most relevant elements in determining toxicity, and they discovered that features like profanity and personal insults were given a lot of weight by the algorithm. The results of this research could have a big impact on the way social media sites are monitored and bad content is kept from spreading.

The research work by Devi and Sharmila (2022)suggests an innovative method for identifying false information on social media. NLP tools were used to preprocess the data. The text was vectorized using text embedding methods like bag-of-words and TF-IDF. The LDA topic model and the graph neural network were subsequently trained using the vectorized text. The method groups the data into topics using LDA topic modeling, and then detects the outlier topics—those that are not similarly popular as the others. A graph neural network is then used to classify these anomalous topics as false data. The technique was tested on a dataset of tweets regarding the COVID-19 epidemic, and the results demonstrated that the technique was highly accurate at detecting misleading information.

These five studies use sentiment analysis and topic modeling from natural language processing (NLP) to analyze customer evaluations. Businesses can use the study's findings to better understand consumer sentiment and modify the caliber of their services as a result. The studies significantly advances the fields of NLP and data mining. Even while previous research has revealed the crucial factors affecting visitors' satisfaction with airport services, there is always room for improvement in terms of using travelers' opinions to drive targeted changes in various airport departments. This study aggregated evaluations from different airport departments to close this gap and provide feedback to the respective management teams to support targeted changes. This study then designed a plan for management to collaborate and handle traveler issues more efficiently, providing the airport industry with useful information into specific areas that need improvement across many divisions.There is always space for improvement when it comes to using travelers' opinions to encourage targeted changes in various airport departments, even though previous research have highlighted the elements most strongly connected with travelers' satisfaction with airport services. By accumulating evaluations of several airport departments and offering input to the corresponding management teams in order to support focused improvements, the study aims to close this gap. The study's findings can be utilized to create a management strategy for better interacting with and addressing traveler complaints, giving the airport sector valuable knowledge about the particular areas that require development across several divisions.

Table 1 shows the summary of the literature review.

Table1: Summary Table of the Literature Review

Literature Review Summary Table			
Author	Objective of Study	Method	Findings
Mizufune and Katsumata (2018)	To create a joint classification model of topic and polarity to identify elements that contribute to customer happiness and dissatisfaction in reviews of airport services.	Support Vector Machine (SVM) classifier	Satisfaction factors: cleanliness, friendly staff, and on-time flights. Dissatisfaction factors: long wait times, lost luggage, and delayed flights.
Hasib et al. (2021)	To determine the main topics and sentiment of internet reviews for Bangladesh airlines using topic modeling and sentiment analysis.	Topic modeling: LDA. Sentiment analysis: lexicon-based approach	Common topics in the reviews: airline, service, crew, and food
Kwon et al. (2021)	To perform topic modeling and sentiment analysis on postings from Skytrax website that hosts numerous airline reviews online.	Topic modeling: LDA. Sentiment analysis: lexicon-based approach	Common topics in the reviews: airline service, food, and aircraft.
Lucini et al. (2020)	To find airline customer satisfaction metrics from internet reviews using text mining.	LDA was utilized to create customer satisfaction dimensions and a logistic regression classifier was employed to predict airline recommendation.	Most important dimensions of airline customer satisfaction: cabin staff, onboard service, and value for money.

Lim and Lee (2020)	To assess how full service carriers (FSCs) and low cost carriers (LCCs) in the airline industry are perceived in terms of service quality.	LDA	The most significant dimensions for FSCs and LCCs were tangibles and reliability.
Choi and Joo (2019)	To perform topic detection on online book reviews.	LDA	Common topics in the reviews were plot, characters, writing style, and overall rating.
Saranya and Geetha (2020)	To use topic modeling to generate word clouds on clothing reviews.	LDA	Most common topics in the reviews were fit, fabric, and color.
Ashok et al. (2022)	To develop a deep learning model for toxicity prediction in comments.	Bidirectional Encoder Represent-ations from Transformers (BERT)	Most common words associated with toxicity were ”hate,” ”abuse,” and ”threat.”
Devi and Sharmila (2022)	To use topic modeling to detect infodemic on Twitter during pandemic situations.	LDA	Most common topics in the tweets were related to vaccination, lockdown, and conspiracy theories
Alam et al. (2020)	To use topic modeling to observe the trend of Bangla news.	LDA	Most common topics in the news articles were politics, crime, and entertainment

3 Methodology

3.1 Introduction

The research methodology applied the Knowledge Discovery in Databases (KDD) approach to evaluate airport reviews and derive insightful information using topic modeling. KDD offers a logical and structured framework for transforming raw data into information that can be applied to enhance business outcomes. It was aimed to identify hidden themes, trends, and sentiments expressed by travelers in their reviews using the KDD approach, which can provide insight into a number of areas of airport services, amenities, and traveler experiences. The goal of using this study methodology was to learn more about the attitudes and preferences of travelers toward airports. Businesspeople can use the information from this study to inform their actions, which can enhance airport services and increase overall consumer satisfaction.

3.2 Modified KDD Methodology

In this study, the modified Knowledge Discovery and Data Mining (KDD) approach has been successfully applied to evaluate airport reviews and produce meaningful topic models. The quality and contextuality of the results have been improved by incorporating particular data mining techniques, particularly Latent Dirichlet Allocation (LDA), into the traditional KDD process, producing more useful and relevant data. Fig.1 shows the modified KDD Metthodology implemented in this study.

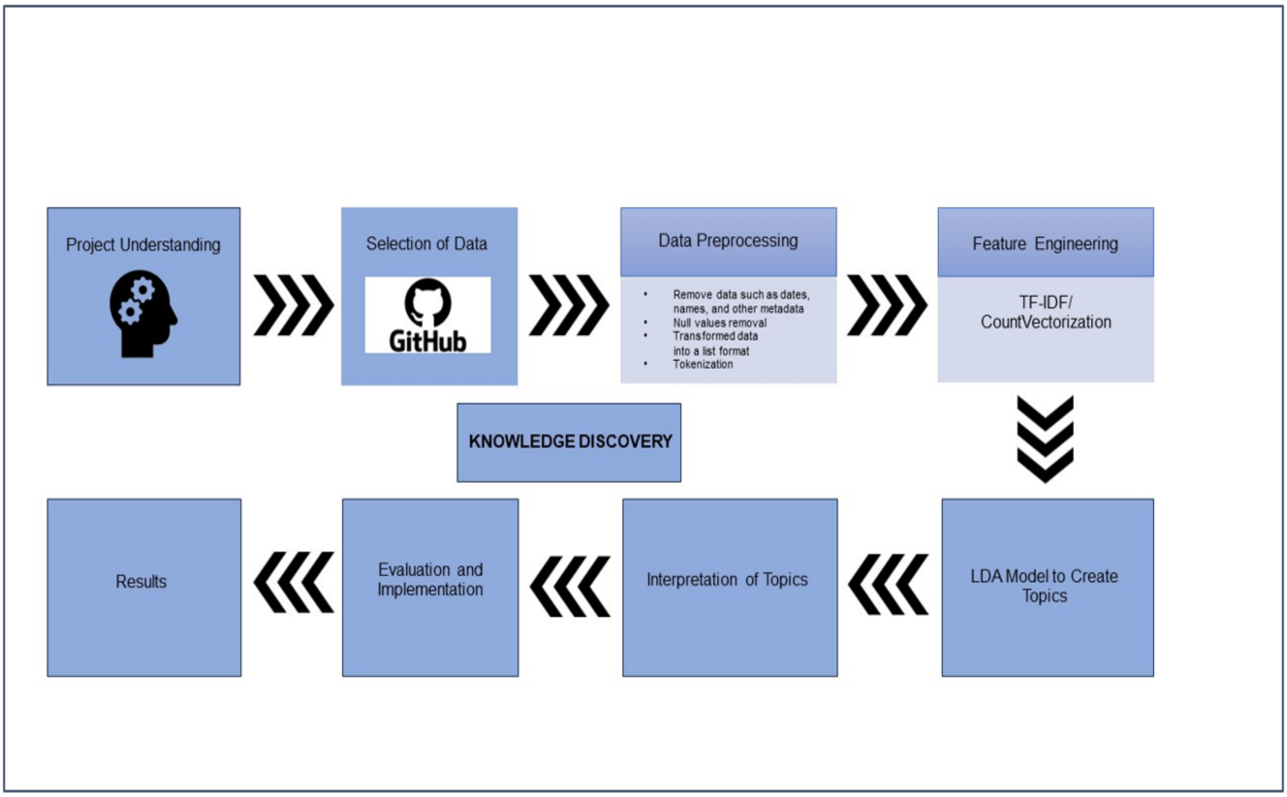

Project Understanding
Selection of Data
GitHub
Data Preprocessing
Remove data such as dates, names, and other metadata
Null values removal
Transformed data into a list format
Tokenization
Feature Engineering
TF-IDF/ CountVectorization
KNOWLEDGE DISCOVERY
LDA Model to Create Topics
Interpretation of Topics
Evaluation and Implementation
Results

Figure 1: Modified KDD Methodology

3.2.1 Project Understanding

The foundation for efficient project planning, implementation, and monitoring is a thorough grasp of the project. It enables one to take well-informed judgments, foresee difficulties, and adjust to changes that might occur throughout the project. The project's objectives and the precise results it plans to produce were made known during this phase. On the basis of user reviews, this study uses topic modeling to identify significant issues related to several airport departments.

3.2.2 Selection of Data

17,747 airport reviews scraped from the Skytrax website made up the dataset used in this study, which was obtained from GitHub. These reviews offer useful information on a variety of airports throughout the world. To guarantee that the insights obtained from the analysis appropriately represent the topic of interest, thorough data selection was necessary. By selecting a dataset specifically for airport reviews, this project's analysis is founded in the actual experiences of travelers, allowing for informed decision-making and practical recommendations for airport administrations and companies looking to enhance the quality of their services.

3.2.3 Data Preprocessing

The dataset was preprocessed to eliminate extraneous data such as dates, names, and other metadata from the reviews before topic modeling approaches were applied. Additionally, all the numeric values and null values were taken out of the dataset. The information was then transformed into a list format. The text was subsequently transformed into numerical tokens using the tokenization process. Word-level tokenization was used in this study to tokenize the text. Every word received its own unique numerical ID and was treated as a separate entity. By efficiently breaking up the text into individual words, this method built an organized framework for further study. Stop-word elimination was used to get rid of overused, meaningless terms. These preprocessing methods transform the text data into a topic modeling-friendly format.

3.2.4 Feature Engineering

A detailed text data transformation and feature engineering process was carried out in order to prepare the raw airport review data for topic modeling. The goal of these procedures was to transform unstructured text input into a structured numerical format appropriate for efficient topic modeling analysis. To convert the textual data into numerical representations for topic modeling, Term Frequency-Inverse Document Frequency (TF-IDF) and CountVectorizer were used as the different approaches.

3.2.5 LDA Model to Create Topics

The LDA model, a popular topic modeling method, was developed to find hidden topics in text review data. The LDA model was able to precisely identify underlying topics within the airport reviews by taking into account the word co-occurrence patterns within the dataset. This greater level of comprehension made topic modeling more precise and provided insights that were more in line with the aims and feelings of the travelers.

3.2.6 Interpretation of Topics

The probability distributions of the LDA model made it possible to identify the key topics that emerged from the airport reviews. The main themes that frequently appeared in the evaluations was better understood by looking at the top words connected to each topic. These top words served as markers, providing perceptions into the particular areas of traveler experiences at airports that were commonly highlighted.

3.2.7 Evaluation and Implementation

The evaluation procedure used in this study aims to thoroughly evaluate the LDA model's performance in the context of airport reviews. This methodical evaluation process aimed to learn more about the caliber and interpretability of the topics that the model produced. It was aimed to assess the effectiveness of the LDA model and its suitability for deriving meaningful insights from the dataset by carefully assessing several elements of the generated topics and their relevance.

3.2.8 Results

The evaluation's findings offer strong proof of how well the LDA model extracts pertinent and understandable issues from airport reviews. Airport management may make intelligent decisions for service improvement because to the clear and meaningful topics' insightful descriptions of traveler experiences and preferences.

3.3 Conclusion

In conclusion, it has been proved that this modified KDD methodology works well for converting raw airport review data into actionable knowledge. Travelers and airport administrations will be able to make data-driven decisions that will enhance airport services, make the greatest use of available resources, and raise general consumer happiness with the help of the insights gathered from this study. The findings of this study can be used in a range of domains where KDD and data mining techniques can be adjusted to extract meaningful insights from a variety of datasets, resulting in improved user experiences and beneficial business outcomes.

4 Implementation of Topic Modeling

4.1 Introduction

This chapter demonstrates how topic modeling is implemented to analyze airport review data and get insightful conclusions. The candidate has created a thorough evaluation strategy that takes into account four important criteria: Coherence Score, Topic Distribution, Keyword Analysis, and Model Stability.

4.2 Data Extraction and Preprocessing

The data was gathered from GitHub and is a scraped dataset derived from the passenger reviews that has been published on the Skytrax website. It includes 17747 airport reviews that have been submitted by diverse users regarding various airports throughout the world. The reviews that were published between the years of 2002 and 2015 were used to build the dataset. As it is shown in the Fig.2, the majority of the data comes from the years 2012 to 2014.

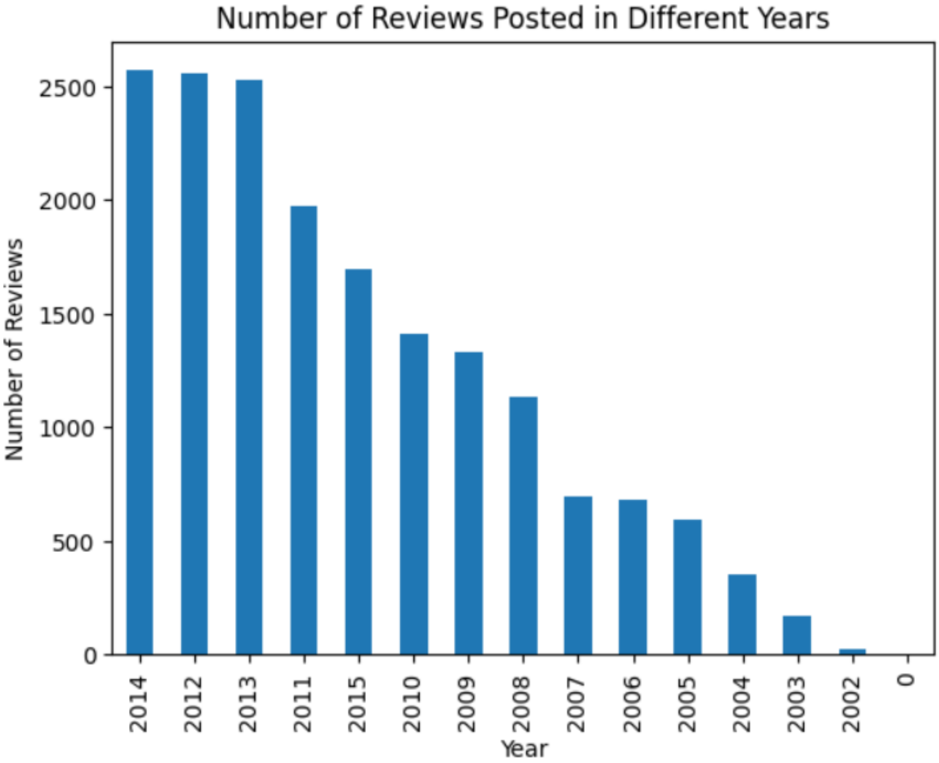


Figure 2: Number of Reviews Posted in Different years

As the Fig.3 depicts, the London Heathrow Airport received the most reviews, 520, which is the most of any airport.

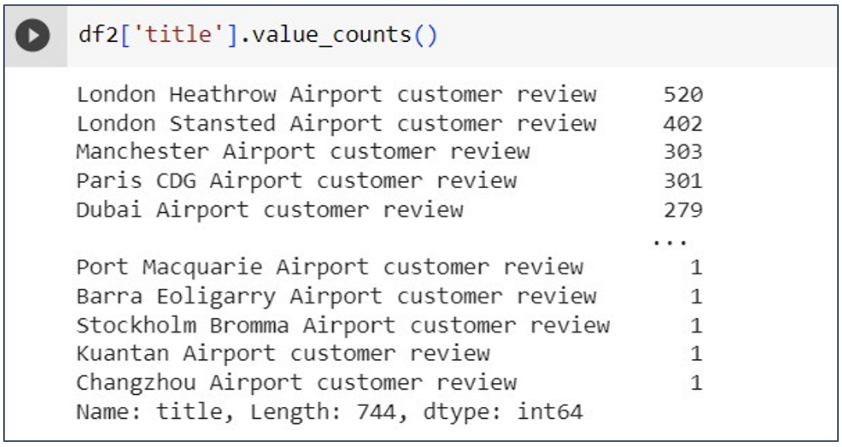
Prior to implementing topic modeling, the acquired dataset underwent preparation to assure its quality and relevance. To convert the raw data into a format suited for efficient topic modeling, many crucial procedures were taken during the preprocessing phase.

4.2.1 Elimination of Extraneous Data

Dates, names, and other unnecessary metadata were deleted from the dataset in order to maintain the primary focus of the evaluation. This stage intended to identify the significant of traveler comments without being affected by irrelevant information. The study might be more accurate and illuminating by removing this unnecessary data.

4.2.2 Handling Numeric and Null Values

The dataset was cleaned up to improve the quality of the data by eliminating all numbers and null entries. This was done to make sure that the dataset only contained the text reviews. Numbers were removed to avoid any influence from unrelated data and null values were removed to lessen the effect of missing or incomplete data on the analysis.



```
df2['title'].value_counts()

London Heathrow Airport customer review     520
London Stansted Airport customer review     402
Manchester Airport customer review          303
Paris CDG Airport customer review           301
Dubai Airport customer review               279
                                           ...
Port Macquarie Airport customer review        1
Barra Eoligarry Airport customer review       1
Stockholm Bromma Airport customer review      1
Kuantan Airport customer review               1
Changzhou Airport customer review             1
Name: title, Length: 744, dtype: int64
```

Figure 3: Airport Names and Corresponding Number of Reviews

4.2.3 Transformation into List Format

The data obtained from the reviews was converted into a structured list format to aid in additional processing and analysis. This stage consistently arranged the data to make it easier to manipulate and set the stage for succeeding steps.

4.2.4 Tokenization

Tokenization, a key step in natural language processing, separates text into individual components, usually words or tokens. The text was divided into words in this study using word-level tokenization, and each word is given a distinct numerical ID.The candidate defines a tokenizer by utilizing the Python RegexpTokenizer function from the nltk.tokenize package. The review texts are divided into individual words by this tokenizer using a regular expression pattern. As a parameter to the RegexpTokenizer, the regular expression pattern r'w+' is used. While ignoring punctuation and whitespace, the pattern r'w+' will match one or more consecutive alphanumeric letters and underscores, essentially breaking the text into distinct words.

4.2.5 Stop-Word Elimination

Stop-word removal was used to further improve the quality of data and boost topic modeling's effectiveness. Stop words—such as 'the,' 'and,' and 'is'—are frequently used words that have little significance when employed in a particular context. The emphasis was switched to words that made more significant contributions to the topic modeling process by eliminating these uninformative keywords.

4.3 Topic Modeling using TF-IDF and CountVectorizer

The textual data were converted into numerical representations for topic modeling using TF-IDF and CountVectorizer. Words were given weights by TF-IDF according to their frequency in the review and their inverse frequency over the entire dataset. Each distinct word in the reviews was considered as a feature by CountVectorizer, and the frequency of each word was assigned as the corresponding value for each feature. The goal of both techniques was to provide organized matrices for analysis.

4.3.1 Topic Modeling using TF-IDF

With the use of the Term Frequency-Inverse Document Frequency (TF-IDF) approach, the text data is converted into a numerical format appropriate for topic modeling. Words are given weights by TF-IDF based on how frequently they occur in a document and their inverse frequency across all documents. The generated TF-IDF matrix provides the dataset in the form appropriate for topic modeling. Various parameters used in the TfidfVectorizer function are: lowercase, stop_words, ngram_range and tokenizer. All review texts are transformed to lowercase before processing in order to execute the TF-IDF technique. It helps to guarantee that words with different cases are treated equally, decreasing the dimensionality of the feature matrix. Stopwords used frequently in English are eliminated from the reviews. Stopwords are words that are frequently used but don't convey a lot of meaning such as "the," "and," and "in". Stopwords are eliminated, allowing the feature matrix to concentrate more on words that convey meaning. The range of n-grams to consider is specified by the ngram_range option. It is set to (1, 1) in this implementation, meaning that the feature matrix only takes into account single words or unigrams. Here, the candidate pass the custom tokenizer (tokenizer.tokenize) that separates the text into words according to the regular expression pattern, r'w+'.

After vectorizing the reviews using TF-IDF, the fit_Transform function is used to both fit the vectorizer to the data (creating the vocabulary, determining IDF values, etc.) as well as convert the data into a TF-IDF weighted feature matrix.

4.3.2 Topic Modeling using CountVectorization

In this study, the collection of airport reviews was transformed into a numerical representation using CountVectorizer, a key text preprocessing method in natural language processing (NLP). Each distinct word in the reviews is treated as a feature by CountVectorizer, and its frequency in each review serves as the corresponding feature value. This process creates a "bag of words" representation. Tokenization, which separates the text into individual words, is followed by the production of a vocabulary made up entirely of the unique terms contained in the evaluations. A distinct numerical ID is given to every word in the vocabulary. In the following step, CountVectorizer counts the instances of each word in each review. This produces a sparse matrix where each row denotes a review and each column denotes a word from the vocabulary. The cell values of the matrix show how frequently each word appears in each review. This numerical representation gives our machine learning algorithms useful input data that was used to do topic modeling and, in turn, get a deeper understanding of traveler experiences and feedback at airports.

The parameters ngram_range, min_df, max_df are passed to the function CountVectorizer. The ngram_range is set as (range_min, range_max) where range_min and range_max are the minimum and maximum n-gram lengths, respectively. The minimal frequency necessary for a word to be included in the feature matrix is specified by min_df. It is an integer or float that indicates how many instances or what proportion of review texts the term must appear in. Below this threshold, words were not considered. The highest frequency threshold for a word to be included in the feature matrix is specified by the max_df option. Above this threshold, words were not considered.Next, the review text data from the a data frame is converted into a numerical feature matrix based on word counts using the fit_transform method. A sparse matrix reflecting the number of words (or n-grams) in the data is the resulting variable.The sparse matrix was then transformed into a dense representation with the help of the toarray() function, and a pandas data frame was produced from it. The function get_feature_names(), which returns the names of the features (words or n-grams) obtained by the CountVectorizer, is used to set the column names of the data frame. This creates a data frame with each row representing the data, each column representing a word or n-gram from the corpus, and each column including the word or n-gram's matching count in each document. Further analysis was performed using this data frame.

The CountVectorizer was used in this project to convert the collection of unprocessed review text data into a numerical feature matrix based on word counts. The candidate did this by creating an instance of the CountVectorizer object with particular settings catered to the needs of the project. To automatically remove frequent English stopwords, the stop_words parameter was set to 'english'. This was advantageous because stopwords have little meaning and may affect our data. In addition, by utilizing the max_df parameter, the inclusion of words was managed according to their word frequency. The terms that appeared in more than 10% of the papers were guaranteed to be deleted by setting max_df=0.1.

The max_features parameter was utilized to set a word count restriction of 5000 in order to better regulate the size of the feature matrix. It was possible to create a more concise representation without sacrificing important information by choosing the top 5000 most frequently occurring terms.

The fit_transform method was called on the review text data following the configuration of the CountVectorizer. By creating the vocabulary and applying preprocessing settings, the vectorizer was fitted to the data, and the text was changed into a sparse matrix that represented the word counts in each document.

The generated numerical feature matrix was then used for the further tasks. The CountVectorizer enables the nlp model to efficiently learn from and make meaningful predictions on the text data by transforming unstructured textual data into an organized and quantitative manner.

4.3.3 Selection of CountVectorizer for Further Analysis

A comparison study was done after TF-IDF and CountVectorizer were both applied to find the best strategy for further steps. It was discovered that CountVectorizer performed better than TF-IDF in terms of producing useful insights and representing the textual material. Themain points and underlying themes of the reviews were more accurately captured using CountVectorizer. CountVectorizer was chosen for the further stages.

4.4 LDA Model Training

It was discovered that CountVectorization outperformed TF-IDF approach in terms of results. Therefore, the CountVectorization matrix was applied to the following processes. After the text input was transformed into a sparse matrix using CountVectorizer, it was then used to train the Latent Dirichlet Allocation (LDA) model. LDA will use the sparse matrix, which holds the word frequencies of the data, to identify the underlying topics and their distributions throughout the data. The LDA model can capture deeper linkages between words and their meanings because to its greater understanding of word usage context and relationships, which leads to more accurate topic modeling.

The LDA model's implementation started with deciding how many topics were intended to be discovered; this crucial choice affects the level of detail and precision of the topics revealed. This was set to fifteen. It was found that 15 topics provide a meaningful and thorough representation of the underlying themes in the data after experimenting with various numbers of topics and analyzing coherence scores. This range of themes, which included a wide range of airport-related experiences, enabled for a comprehensive understanding of traveler opinion. Additionally, it achieved a balance between giving in-depth analyses of certain airport improvement areas and preserving interpretability for stakeholders.

Using the scikit-learn library, an LDA object was subsequently created. The specified number of topics were configured for this object. On the basis of this object, it was feasible to carry out the model fitting and transformation phase on the preprocessed text data. The fit_transform function was used to train the LDA model on the dataset and create a transformed matrix. This matrix—known as the LDA matrix—captures the relationship between each review and the found topics. Each entry in the matrix indicates the degree to which a review is associated with a particular topic. Additionally, the topic-word distributions were extracted which is essential to comprehending the themes supporting the discovered topics.It was able to determine the relevance of individual words inside each topic by using the distributions that were derived from the LDA model's components.

After running LDA on the dataset, topics and their top associated words were printed. The topic's component was iteratively examined one by one within the matrix created by the LDA model. This loop allowed for an extensive examination of the topics the LDA model had suggested. Zipping allowed for the pairing of terms with their associated weights, which made it easier to understand the relationship between words and their importance.

The terms were sorted according to their weights, which effectively put them in descending order of significance within each topic. This procedure of sorting made sure that the most important terms were at the top of the analysis. The essential concept of the topic was then taken from a selection of the top 15 terms and collated for further steps.

The outcome was the printing of topics and the key words that went along with them. This method of presentation systematically emphasized the identified topics and the words that defined them. This illustration provides a brief but informative depiction of the text data's determined thematic dimensions. In the end, this understanding enables a deeper understanding of the project's findings and establishes the framework for additional research and interpretation.

The number of top words that will be displayed for each topic is set at 15 in order to highlight the most significant words for that topic.

4.5 Interpretation of Topics

The LDA model produces probability distributions of the topics and words that are related to them. The key topics that appear in the airport reviews were better understood by looking at the top words for each topic. Further background information was provided by examining the data most directly related to each topic, provides more insights into recurring themes in the reviewers' comments on the airport journey. Some of the topics generated were comparable and fit the same theme. These topics were integrated in order to send the corresponding reviews to the appropriate department.The model commonly produces the following topics: Parking desk, Travel desk, and Outlet desk. This indicates that parking, travel-related queries, and outlet amenities like restaurants, stores, etc. are of greater significance to people. There were 5996 reviews for the matter of parking, 5184 reviews for the area of transport, and 1973 reviews for the topic of outlet amenities. These reviews could be forwarded to the corresponding airport departments to help them operate better and raise the standard of the airport as a whole.

5 Evaluation and Results

5.1 Introduction

The method of evaluation presented here is a methodical approach used in this study to thoroughly evaluate the Latent Dirichlet Allocation (LDA) model's performance in the context of airport reviews. This evaluation approach seeks to offer insights into the quality and interpretability of the topics extracted by the model by carefully assessing several aspects of the generated topics and their relevance. The following are the different evaluation strategies used in this research.

5.2 Coherence Score Evaluation

A crucial criterion for evaluating the LDA model's efficacy is the coherence score. The relevance and coherence of the topics detected are revealed by calculating the coherence score for each created topic. Greater interpretability and relevance of the issues are indicated by a higher coherence score. This evaluation aspect was crucial for determining whether the model can generate interesting and well-organized topics that fit the nature of airport reviews.

There is no direct built-in method to determine coherence scores in Scikit-learn's LDA implementation. The Gensim library is more frequently associated with coherence scores. This led to the installation of the Gensim library in order to determine the coherence score. The objective was to assess the quality and coherence of the topics produced by the LDA model in light of the available data. In order to evaluate the LDA model's performance in the context of the supplied data, this function provides a way to quantitatively examine the coherence and quality of the topics it generates. How well the topics capture the underlying themes in the data is determined by the final coherence score.

The results showed that the topics produced by the LDA model were of high quality and coherence, with a coherence score of 0.558. Higher values denote more coherent and interpretable issues. Coherence ratings normally range from 0 to 1. With a score of 0.558, it is possible that the topics in this case may be made more logical.

5.3 Topic Distribution Analysis

Determining the popularity and acceptability of different topics is mostly dependent on the evaluation of topic distribution across the reviews. We can determine which topics are more common and appropriate for the setting of airport reviews by looking at how frequently various topics appear in the reviews. The things that reviewers frequently talk about in their input are likely to be reflected in high-frequency topics. This feature provides insightful information about the applicability of the extracted topics to actual situations.

In order to use topic distribution as an evaluation metric, the proportion of each topic in the entire data of reviews was calculated initially. Every review in the data is given a distribution of topics by the LDA model. These topic assignments were extracted using the LDA model's .transform() method.

The next step was to aggregate the topic proportions across all the data to determine how frequent each topic is over the full corpus. The normalized proportions, which add up to 1, were obtained by dividing the aggregated topic proportions by the total quantity of data. Interpreting and comparing the relative popularity of various topics is made simpler by this step.

The resulting graph in the Fig.4 displays the frequency of each topic over the whole corpus. It is evident from the graph that Topic 12 is the most prevalent topics while Topic 9 is the least significant one. Topic 12 is all about the airport lounges and it is clear that people are more concerned about the lounge services while they are in an airport.

There where topics which produced similar key words. Such topics where combined and insignificant words where removed from the topics. The final topics generated where Parking desk, Travel desk and Outlet desk. The travel desk topic was created by combining topic 6, topic 7 and topic 0. parking desk topic was the combination of topic 11, topic 12 and topic 13. Outlet desk was topic 9.

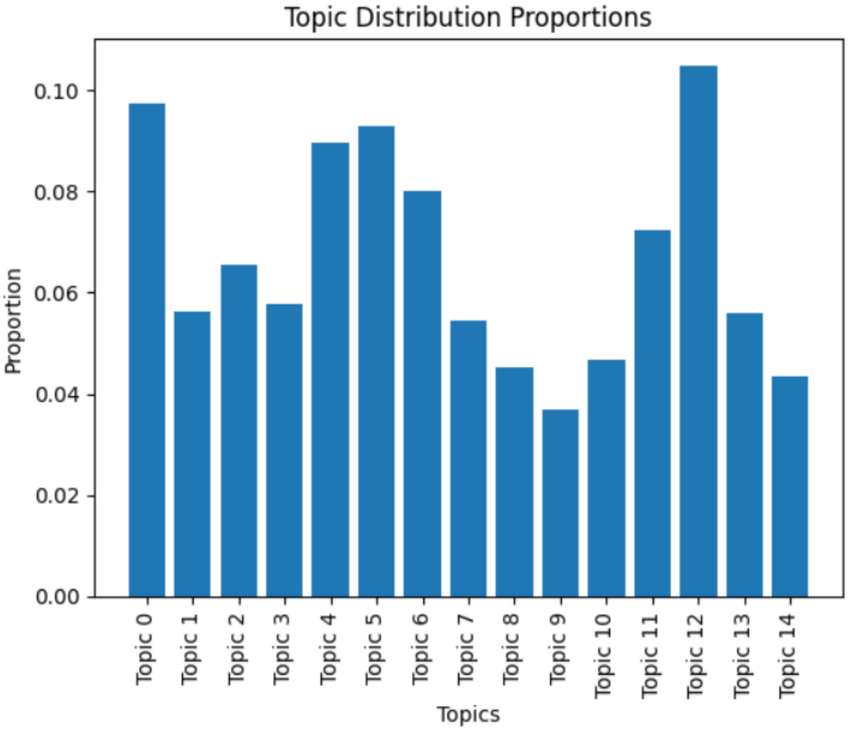


Figure 4: Proportion of Each Topic in the Dataset

5.4 Keyword Analysis for Relevance

Analyzing keywords in depth makes it easier to evaluate the words connected to each topic. We acquire a better grasp of the consistency and application of the words used to characterize the topics by evaluating the importance of these terms within the context of airport reviews. This study provides a qualitative viewpoint on how well the discovered topics correspond with the targeted area of interest.

The process of performing a keyword analysis included looking at the words that are most closely related to each topic. The main areas and issues covered in airport reviews are reflected in pertinent keywords. It was determined whether the keywords related to each topic are appropriate for the airport reviews category. The keywords include words like "baggage," "handling," "lost," etc. if the topic is relevant to "travel desk." If a keyword doesn't fit the domain, the model could not be capturing the appropriate topics.

Assigning meaningful labels to topics was made easier by keyword analysis. When the keywords in a topic are consistent and closely related, it becomes simpler to label the topic in a way that accurately conveys its content. The top keywords for each topic were manually reviewed as part of keyword analysis. This qualitative analysis made it possible to determine whether the model's identified topics match domain knowledge and expectations.

The topic model is refined using keyword analysis as well. The model parameters, preprocessing procedures, and number of topics were adjusted when topics with inconsistent or irrelevant keywords were discovered in order to improve topic quality.

5.5 Model Stability Assessment

The evaluation strategy includes a technique for evaluating the stability of the model by contrasting topics derived from randomly chosen subsets of the data with those obtained from the complete dataset. The mean stability score obtained for the TF-IDF model was 0.253 and for CountVectorization model was 0.0113. This method enabled to assess the LDA model's susceptibility to overfitting or underfitting. It was able to determine the model's resilience and consistency in capturing hidden themes across various data segments by assessing its performance on various subsets.

An in-depth understanding of the dependability of the generated topics in the LDA model was provided by the model stability assessment. There were numerous crucial steps in this examination. The dataset was first split into several subgroups at random to ensure diversity and representation. The LDA model was trained and topics were created for each subset. Then, using metrics of similarity or consistency, these topics were compared across various subgroups. As a quantitative measure of how consistently topics are created across different data divisions, the comparison findings were used to calculate a stability score. It was feasible to pinpoint topics that hold up well and stay cohesive regardless of the particular subset utilized for training by looking at the stability ratings.

This evaluation metric aids in improving model parameters, preprocessing procedures, and topic extraction methods by shedding light on the model's sensitivity to data variations and enabling a deeper understanding of the model's capacity to generalize topics across various dataset segments.

5.6 Conclusion

These evaluation techniques combine to provide a thorough understanding of the performance of the LDA model in the context of airport evaluations. The model has proven its effectiveness at analyzing the review data to extract meaningful and pertinent topics by quantifying coherence, examining topic distributions, estimating keyword relevance, and evaluating model stability. These understandings directed how the project interprets the findings and how to apply the created issues to the actual situation of airport assessments.

6 Discussion

6.1 Introduction

The outcome of this study has produced insightful information that will be useful to major airports throughout the world when analyzing traveler comments. The discussion section tries to explore this project's main findings, implications, limitations and future possibilities.

6.2 Main Findings

The project successfully implemented topic modelling on traveler reviews by department, revealing information on many aspects of airport experiences, using techniques such as LDA, CountVectorizer, and TF-IDF. Notably, the selection of 15 themes for the LDA model offered a compromise between granularity and interpretability.

The topics chosen, such as Parking desk, Travel desk, and Outlet desk, provide useful information on the areas that passengers value most. The project's capacity to identify particular development opportunities among several airport divisions increases the likelihood of focused improvements.

6.3 Quantitative Evaluation of Number of Topics

The selection of the 15 subjects for this project was made after a quantitative analysis. Coherence scores was the quantitative criteria used to assess the number of subjects. It was intended to determine the point at which coherence scores were maximum by methodically altering the number of topics and computing coherence ratings for each configuration. With fifteen subjects, it was possible to analyze traveler comments in more detail. Similar key words were developed by some topics. These themes were integrated, and unnecessary words were eliminated. The final subjects produced were Travel desk, Outlet desk, and Parking desk. Topics 6, 7, and 0 were combined to create the travel desk topic. Topics 11, 12, and 13 were combined to form the topic for the parking desk. Topic 9 was the outlet desk.

6.4 Conclusion

The results have important significance for airport administrations, decision-makers, and service providers. Airports may more efficiently deploy resources, enhance service options, and address particular pain spots by developing a thorough awareness of traveler worries and satisfaction levels. This not only improves the entire travel experience but also encourages customer loyalty.

7 Conclusion and Future Work

Airport reviews have benefited greatly from the effective application of the KDD methodology, combined with modified methodologies for data preprocessing and topic modeling.

Through the use of textual review data, this study set out on an objective to investigate and examine the opinions, preferences, and experiences of travelers. It was able to accomplish the study goals and make valuable contributions to the field of improving airport services by using a methodical approach to data gathering, preprocessing, and topic modeling.

The modified KDD methodology used in this study was successful in converting the raw airport review data into useful information. The combination of LDA for topic modeling and CountVectorizer for text preprocessing turned out to be a reliable and effective method for uncovering hidden topics, trends, and attitudes conveyed by travelers in their reviews.

By carefully examining the model, it was able to confirm its coherence, topic distribution, keyword relevance, and stability, proving the model's reliability in identifying hidden patterns and generating relevant topics. The coherance score obtained by the model was 0.553. The topic modeling implementation was a particularly successful undertaking since it enabled to identify significant and sensible topics that offered deeper insights into various aspects of airport situations. The topics identified are Parking desk, Travel desk and Outlet desk. There are 5996 reviews under parking desk topic, 5184 reviews under travel desk topic and 1973 reviews under outlet desk topic. These reviews can be send to corresponding airport departments.

The findings of this study have important implications for decision-makers, service suppliers, and airport authorities. The issues that have been taken out provide a thorough insight of worries, wants, and areas of satisfaction and displeasure of travelers. Airports may improve overall customer happiness and loyalty by effectively allocating resources, improving service offerings, and addressing particular problem areas.

Recognizing the shortcomings of the study is crucial, though. The correctness and authenticity of the given data are key factors in the performance of the model. Future developments would contain a number of significant enhancements. Advanced text preprocessing techniques, such as word embeddings like Word2Vec or GloVe, would be utilized in order to capture deeper semantic relationships between words and improve the quality of features used in later topic modelling. A detailed analysis of various topic modelling methods like Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA) could be undertaken to reveal numerous features of the data that might not be as clear with LDA alone. Systematic experimentation using techniques like grid search or cross-validation could be used to ascertain the best arrangement, especially with regard to the number of topics. In order to comprehend the sentiments underlying the topics that have been addressed, the project's scope could also be expanded to incorporate sentiment analysis, which would integrate topic modelling with sentiment analysis. Finally, the adoption of more complex deep learning models, such as LSTM, according to the project's goals, would be taken into account to further increase the breadth and accuracy of sentiment analysis and predictive modeling. The main objective of these improvements is to improve the project's analytical capabilities and supply even more beneficial data for enhancing airports and enhancing traveler experiences.

In summary, this study shows the effectiveness of data-driven approaches in gaining insightful information from textual data. As data science develops, approaches like the one described here—including the effective application of topic modeling—will become increasingly important in converting raw data into useful insights that boost business results and user experiences.

References

Alam, K. M., Hemel, M. T. H., Muhaiminul Islam, S. and Akther, A. (2020). Bangla news trend observation using lda based topic modeling, *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6.

Ashok, K. S., Ashok, K. A. and Naseem, S. M. B. (2022). A neuro-nlp induced deep learning model developed towards comment based toxicity prediction, *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pp. 94–99.

Choi, Y. and Joo, S. (2019). Topic detection of online book reviews: Preliminary results, *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 418–419.

Devi, N. S. and Sharmila, K. (2022). Infodemic detection on twitter during pandemic situation using lda topic modelling, *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, pp. 1253–1257.

Hasib, K. M., Towhid, N. A. and Alam, M. G. R. (2021). Topic modeling and sentiment analysis using online reviews for bangladesh airlines, *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0428–0434.

Kwon, H.-J., Ban, H.-J., Jun, J.-K. and Kim, H.-S. (2021). Topic modeling and sentiment analysis of online review for airlines, *Information* **12**(2).
URL: *https://www.mdpi.com/2078-2489/12/2/78*

Lim, J. and Lee, H. C. (2020). Comparisons of service quality perceptions between full service carriers and low cost carriers in airline travel, *Current Issues in Tourism* **23**(10): 1261–1276.
URL: *https://doi.org/10.1080/13683500.2019.1604638*

Lucini, F. R., Tonetto, L. M., Fogliatto, F. S. and Anzanello, M. J. (2020). Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews, *Journal of Air Transport Management* **83**: 101760.
URL: *https://www.sciencedirect.com/science/article/pii/S0969699719302959*

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

Saranya, M. and Geetha, P. (2020). Word cloud generation on clothing reviews using topic model, *2020 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0177–0180.