VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

PROJECT REPORT

ON

CUSTOMER SENTIMENT ANALYSIS ON AIRLINE REVIEWS

BECE352E- IOT Domain Analyst

SUBMITTED BY:

AVINASH V – 21BEC1676

VIKASH VIJAYAN A – 21BEC1462

GOPI VARSHAN K R- 21BEC1584

GOWTHAM MALLYA S - 21BEC1590

SAI KRISHNA S– 21BEC1804

Table of Contents

1 INTRODUCTION

In an increasingly competitive airline industry, understanding and effectively responding to customer sentiment are critical for maintaining and improving customer satisfaction and loyalty. The ability to analyze and interpret customer reviews provides valuable insights into the strengths and weaknesses of airlines' services and offerings.

This report presents a comprehensive analysis of customer sentiment derived from reviews of five different airlines, gathered from the Skytrax website. The primary objective of this project is to leverage natural language processing (NLP) techniques and sentiment analysis to evaluate customer sentiment expressed in these reviews. By employing Vader sentiment analysis, sentiments within the reviews were quantified, enabling the assignment of star ratings to reflect the polarity of sentiment: positive (5 stars), negative (1 star), and neutral (2.5 stars).

The project encompasses a multi-step approach, beginning with the collection of review data through web scraping techniques, followed by preprocessing and analysis using tools such as MySQL, Power BI, and Python. Moreover, automation of the data collection process has been achieved through the development of a Python script, executed on the IBM Cloud platform, ensuring regular and timely data updates.

Through this endeavor, insights into the nuanced sentiments of airline customers have been unearthed, facilitating a deeper understanding of their preferences and concerns. Such insights are invaluable for airlines seeking to enhance their services, tailor their offerings, and ultimately improve customer satisfaction.

In the subsequent sections of this report, the methodology, analysis, results, and implications of this project will be expounded upon in detail, providing a comprehensive understanding of the endeavor and its significance within the context of the airline industry.

Technology Stack involved:

The successful execution of this project relied on a robust technology stack encompassing a range of tools and platforms tailored to the specific requirements of each stage. The technology stack employed includes:

1. **Web Scraping Tools:** Python libraries such as Beautiful Soup and Scrapy were utilized for web scraping, enabling the systematic extraction of data from the Skytrax website.
2. **Data Storage:** MySQL was chosen as the database management system for storing the collected data. Its relational structure and query capabilities provided a suitable framework for organizing and managing the dataset efficiently.
3. **Natural Language Processing (NLP):** NLP techniques were applied to preprocess and analyze the textual content of the reviews. Libraries such as NLTK (Natural Language Toolkit) and spaCy were instrumental in tasks such as text tokenization, lemmatization, and part-of-speech tagging.
4. **Sentiment Analysis:** The Vader sentiment analysis tool, part of the NLTK library, was employed to quantify the sentiment expressed in the reviews. This tool assigns sentiment scores to text based on a lexicon of predefined terms, facilitating the classification of reviews as positive, negative, or neutral.
5. **Data Visualization:** Power BI (Business Intelligence) was chosen for data visualization and analysis. Its intuitive interface and powerful visualization capabilities enabled the creation of insightful dashboards and reports to communicate findings effectively.
6. **Automation:** A Python script was developed to automate the process of data collection from the Skytrax website. This script, hosted on the IBM Cloud platform, executes daily at a predefined time, ensuring regular updates to the dataset without manual intervention.

2 DATA COLLECTION

The foundation of this analysis lies in the collection of data from the Skytrax website, a reputable source for airline reviews. The data encompasses crucial elements such as the publication date of reviews, the geographical location of users, and the verbatim text of the reviews themselves.

Web scraping techniques were employed to extract this data efficiently and systematically from the Skytrax website. The process involved accessing the HTML structure of the website and programmatically extracting relevant information using Python libraries such as BeautifulSoup and Scrapy.

Once extracted, the data was organized and stored in a MySQL database for further processing and analysis. This step ensured the integrity and accessibility of the data, facilitating subsequent stages of the project seamlessly.

The data collection process adhered to ethical considerations, respecting user privacy and usage policies outlined by Skytrax. Additionally, measures were implemented to handle potential challenges such as data inconsistency or irregularities in the website's structure.

In the following sections, we delve into the intricacies of data preprocessing, analysis, and automation, building upon the foundation laid by the comprehensive data collection process.

3 DATA PREPROCESSING

Data preprocessing serves as a crucial step in ensuring the quality and reliability of the dataset for subsequent analysis. In this section, we outline the methodologies employed to preprocess the raw data collected from the Skytrax website.

Raw Data Processing Techniques:

The raw data obtained from the web scraping process underwent several preprocessing steps to standardize and enhance its quality. This involved tasks such as:

- Handling missing or inconsistent data entries.
- Standardizing date formats to ensure uniformity across the dataset.
- Cleaning text data by removing HTML tags, special characters, and irrelevant information.

Application of NLP Techniques:

Natural Language Processing (NLP) techniques were applied to the textual content of the reviews to extract meaningful insights. Key NLP tasks performed include:

- Tokenization: Breaking down text into individual words or tokens.
- Lemmatization: Reducing words to their base or root form to normalize variations.
- Part-of-speech tagging: Identifying the grammatical components of each word in the text.

Vader Sentiment Analysis:

Vader sentiment analysis was utilized to quantify the sentiment expressed in the reviews. This tool assigns sentiment scores to text based on a lexicon of predefined terms, capturing both polarity and intensity of sentiment. The sentiment scores obtained facilitated the categorization of reviews into positive, negative, or neutral categories.

Star Rating Assignment:
Based on the sentiment scores derived from Vader analysis, star ratings were assigned to each review to further simplify the sentiment categorization process. Reviews were categorized as follows:

- Positive sentiment: Assigned 5 stars.
- Negative sentiment: Assigned 1 star.
- Neutral sentiment: Assigned 2.5 stars.

By implementing these preprocessing techniques, the dataset was refined and standardized, laying the groundwork for meaningful analysis and interpretation of customer sentiment in the subsequent stages of the project.

4 DATA ANALYSIS:

In this section, we delve into the analysis phase of the project, where the preprocessed data is transformed and interpreted to derive actionable insights regarding customer sentiment towards the five airlines under study.

Importing Data into Power BI:

The preprocessed data from the MySQL database was imported into Power BI for analysis. Power BI's intuitive interface and powerful data visualization capabilities facilitated the creation of dynamic and insightful dashboards.

Data Modeling:

The imported data underwent modeling to structure and organize it in a format conducive to analysis. This involved defining relationships between different tables and creating calculated columns or measures to enhance analytical capabilities.

Visualization Techniques Used:

A variety of visualization techniques were employed to represent the data effectively and uncover underlying patterns or trends. Common visualization types utilized include:

- Bar charts and pie charts to visualize distribution of sentiment categories.
- Line charts to track sentiment trends over time.
- Heatmaps to identify geographic variations in sentiment.
- Word clouds to highlight frequently occurring terms in reviews

5 RESULTS AND INSIGHTS:

In this section, we uncover key insights gleaned from the analysis of customer sentiment towards the five chosen airlines, along with demographic insights regarding user countries.

Airlines' Sentiment Ratings:

Air India: 2.46
British Airways: 2.69
Qatar Airways: 3.59
Emirates: 2.74
Etihad: 2.42

Qatar Airways emerges as the leader in the rating scoreboard, with the highest sentiment rating among the selected airlines, indicating predominantly positive customer sentiment. Conversely, Air India and Etihad exhibit lower sentiment ratings, suggesting a less favorable perception among customers.

Geographical Distribution of Reviews:

The analysis reveals insights into the geographical distribution of reviews, with the following top countries contributing to the analysis:

United Kingdom: 3002 reviews
United States: 1167 reviews
Australia: 947 reviews
India: 439 reviews
United Arab Emirates: 324 reviews

These insights shed light on the global reach and popularity of the airlines, as well as the geographic diversity of their customer base. Furthermore, they provide valuable input for targeted marketing strategies and service enhancements tailored to specific regions.

Airline	Negative	Neutral	Positive	Total
Qatar	242	664	956	**1862**
Ethihad	491	1033	236	**1760**
Emirates	516	654	466	**1636**
British Airways	834	1272	720	**2826**
Air India	401	341	225	**967**
Total	**2484**	**3964**	**2603**	**9051**

Airline	Rating out of 5
Ethihad	2.42
Air India	2.46
British Airways	2.69
Emirates	2.74
Qatar	3.59

Comparison with the existing ratings:

The sentiment analysis results, which categorized reviews into positive, negative, or neutral sentiments, were converted into star ratings for further analysis. These star ratings were then compared against the ratings provided on the Skytrax website for each airline under study.

Validation Methodology:

To validate the accuracy of the sentiment-based ratings, a systematic comparison was conducted between our derived ratings and the ratings displayed on the Skytrax website. This involved:

Extracting the official ratings from the Skytrax website for each airline. Matching these ratings with the sentiment-based ratings derived from our analysis. Assessing the degree of alignment or discrepancy between the two sets of ratings. The comparison revealed a high degree of consistency between the sentiment-based ratings derived from our analysis and the ratings available on the Skytrax website. Across the five airlines under study, the majority of sentiment-based ratings closely matched the official ratings published by Skytrax, affirming the accuracy and reliability of our sentiment analysis approach.

Implications:

The validation of sentiment-based ratings against official ratings from Skytrax enhances the credibility and trustworthiness of our analysis results. This alignment underscores the effectiveness of sentiment analysis techniques in accurately capturing customer sentiment and reflects positively on the validity of insights derived from our analysis.

6 Conclusion

In this concluding section, we summarize the key findings of the project and reflect on its significance within the context of analyzing customer sentiment from airline reviews sourced from the Skytrax website.

Summary of Findings:

The analysis revealed varying levels of customer sentiment across the five selected airlines, with Qatar Airways leading the rating scoreboard and Air India and Etihad exhibiting comparatively lower sentiment ratings.

Insights into the geographical distribution of reviews highlighted key markets and regions contributing to the analysis, providing valuable input for targeted marketing strategies and service enhancements.

Implications for Airlines:

High sentiment ratings indicate a strong customer satisfaction level and present opportunities for airlines to further enhance customer experience and build loyalty.

Airlines with lower sentiment ratings may benefit from targeted improvements in service quality, communication, and customer engagement to address areas of concern and improve overall satisfaction.

Significance of the Project:

By leveraging sentiment analysis techniques and advanced data analytics, this project provides airlines with actionable insights into customer sentiment, enabling them to make informed decisions and drive initiatives aimed at delivering exceptional customer experiences.

The project underscores the importance of leveraging data-driven approaches to understand and respond to customer needs, ultimately contributing to the long-term success and competitiveness of airlines in the dynamic aviation industry.

Future Directions:

Future research and analysis could delve deeper into specific factors driving customer sentiment, such as service quality, pricing, and convenience, to

provide a more nuanced understanding of customer preferences and behavior.

Continued monitoring and analysis of customer sentiment over time can help airlines track trends, identify emerging patterns, and adapt strategies accordingly to maintain competitiveness and relevance in the market.

In conclusion, this project has provided valuable insights into customer sentiment towards airlines, demonstrating the power of data analytics in informing strategic decision-making and driving customer-centric initiatives. As airlines navigate the evolving landscape of the aviation industry, leveraging data-driven approaches will be essential in meeting the evolving needs and expectations of customers and maintaining a competitive edge in the market.