PROJECT: AIRLINE REVIEW CLASSIFICATION

Submitted by

KHYATI SINGH	**21BCE0809**
SOHAM MISHRA	**21BLC1586**
ANKUR SUMAN	**21BDS0097**

ABSTRACT

We address the task of classifying airline reviews as positive or negative to analyze customer sentiment. Leveraging airline review data from Kaggle, our investigation reveals that conventional Machine Learning techniques exhibit superior performance compared to sentiment analysis baselines produced by humans. We delve into a diverse array of probabilistic models, including the Naive Bayes Classifier and Logistic Regression Classifier, for review classification. Employing techniques such as stop words and WordNet, we extract frequent words from the reviews. The study concludes with a comparative analysis of accuracy across various strategic models and a discussion on potential avenues for future research in the realm of airline review sentiment analysis.

Keywords: Airline Review , NLTK, Naïve Bayes Classifier , Logistic Regression Classifier.

1. INTRODUCTION

1.1 PROJECT OVERVIEW:

In the contemporary era of global connectivity, the airline industry plays a central role in facilitating travel and business activities worldwide. With the increasing accessibility of air travel, the quality of service provided by airlines becomes a crucial determinant in shaping passenger experiences. This project aims to develop an advanced airline review classification system utilizing machine learning techniques, including Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier.

1.2 PURPOSE:

The purpose of the "Airline Review Classification Using Machine Learning" project is to employ advanced machine learning techniques, including Decision Tree, Random Forest, and XGBoost classifiers, to systematically analyze and classify sentiments expressed in airline reviews. By categorizing reviews as positive or negative, the project aims to provide airlines with actionable insights to enhance passenger satisfaction and optimize service quality. The utilization of machine learning models facilitates efficient resource allocation, informs decision-making processes, and contributes valuable knowledge to the airline industry. Ultimately, the project serves to establish a framework for continuous improvement in airline services by transforming unstructured customer feedback into meaningful and strategic insights.

2. LITERATURE SURVEY

2\. Addressing the challenges inherent in text sentiment polarity analysis within the specific domain of airline reviews, this study proposes a novel approach to enhance the classification accuracy of the sentiment analysis process. Recognizing limitations in the naïve Bayes algorithm, particularly its reliance on the independence assumption and the sparse nature of word vector matrices, our method integrates machine learning techniques with domain sentiment dictionary weighting methods. Employing an improved word frequency-inverse file frequency algorithm, we extract the feature word vector from airline review text. To mitigate the impact of the independence assumption, we introduce the weighted feature word vector of the domain dictionary post-regression test. Dimensionality reduction of the sparse word vector matrix is achieved through the singular value decomposition algorithm, eliminating redundancy and enhancing efficiency. The remaining features are utilized to construct a polynomial model of the naïve Bayes algorithm. Simulation results demonstrate the efficacy of this method in significantly improving the accuracy of text sentiment classification within the context of airline reviews.

2.2 PROPOSED SYSTEM:

Airline review classification, situated within the broader realm of artificial intelligence (AI), involves leveraging machine learning techniques to discern patterns in data and create models for practical applications. Diverging from traditional computational methods, machine learning doesn't rely on explicit programming instructions but instead enables computers to learn and adapt from data inputs through statistical analysis. Specifically, in the context of airline review sentiment analysis, we delve into supervised learning, unsupervised learning, regression, and classification. Supervised learning involves training models on labeled datasets, while unsupervised learning deals with uncovering patterns without predefined labels. In the regression and classification paradigms, we explore techniques to predict numerical values and categorize data, respectively. To implement the classification aspect, methods such as the Naïve Bayes Classifier and Logistic Regression are employed. The evaluation metrics, including precision, recall, and F-beta score, provide a quantitative measure of the model's performance in effectively classifying airline reviews as positive or negative. This project aims to harness these machine learning principles to automate decision-making processes in analyzing airline reviews, ultimately enhancing the understanding of customer sentiment within the aviation industry.

You must have prior knowledge of the following topics to complete this project.

- Supervised learning
- Unsupervised learning

- Regression and classification
- Naïve Bayes Classifier
- Logistic Regression
- Evaluation metrics (Precision, recall, fbeta)

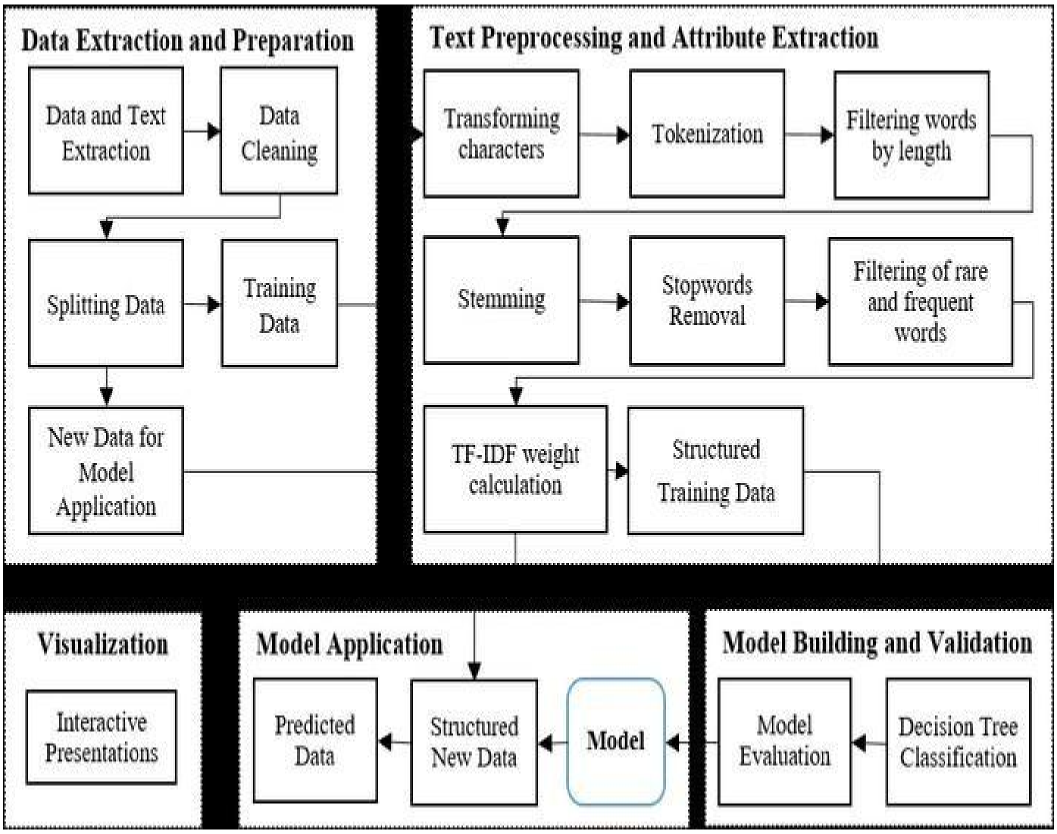


Figure 1: system overview

2.3 PROBLEM STATEMENT

The "Airline Review Classification System" serves as a comprehensive solution for managing information related to customer sentiments in the aviation industry. This system is designed to efficiently handle data pertaining to each customer's feedback, employee interactions, and product evaluations. In this context, each review represents a customer's perspective on their airline experience. The system employs a web-based application framework, making it adaptable for airlines of varying scales.

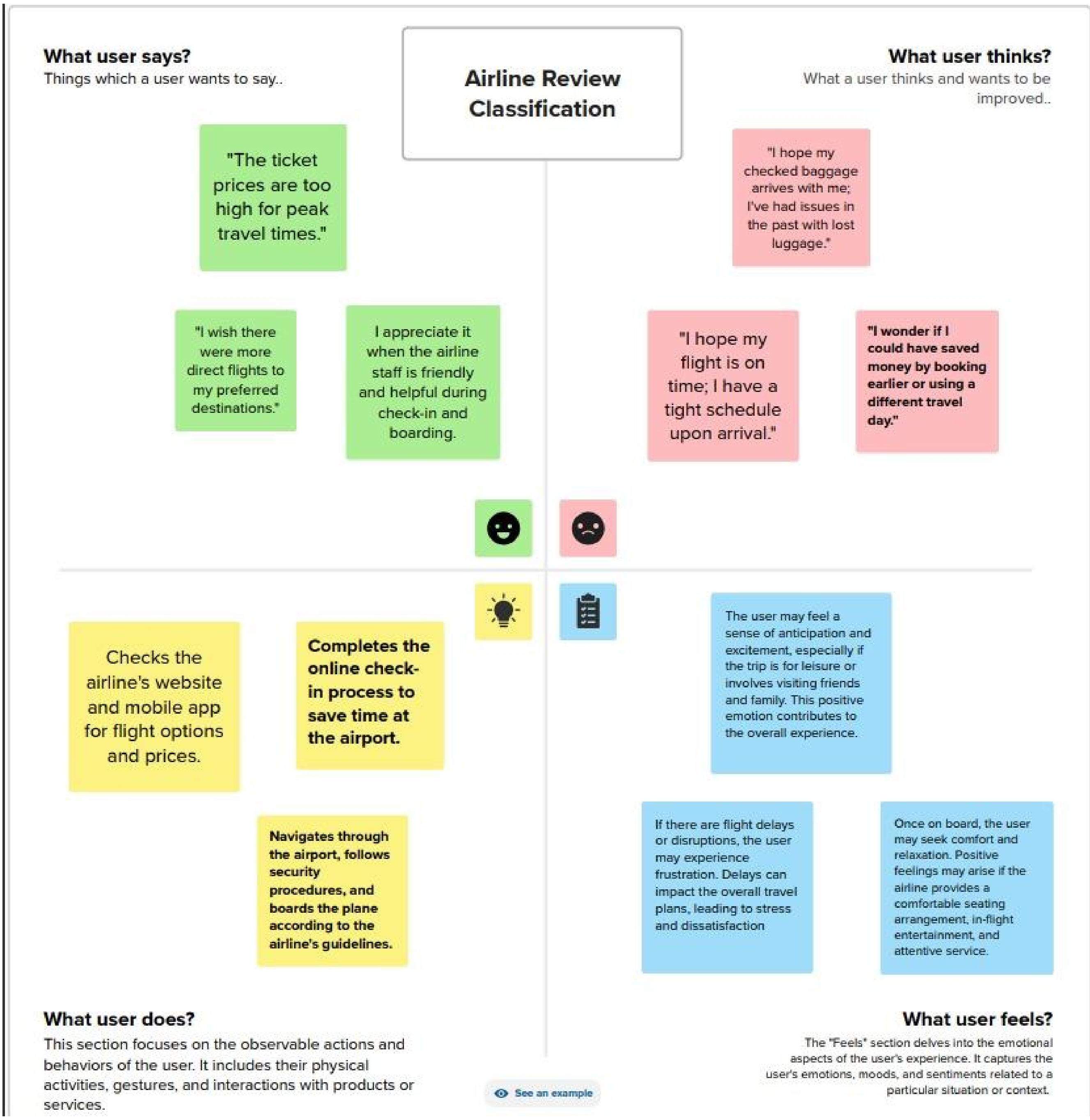
The primary functionalities of this system are categorized into three subsystems. Firstly, the "Review and Feedback Management System" monitors and logs customer sentiments, tracking positive and negative feedback to provide insights into the overall satisfaction levels of passengers. The second subsystem, the "Sentiment Analysis and Classification System," employs advanced machine learning techniques to categorize reviews as positive or negative, automating the analysis process. The third subsystem, the "Reporting and Insights System," generates detailed reports for aviation managers, allowing them to audit and modify strategies based on the analyzed sentiment data.

The end users of the Airline Review Classification System are airline staff, including customer service representatives, and airline managers. Both user types can access the Review and Feedback Management System and the Sentiment Analysis and Classification System. The Reporting and Insights System is restricted to management users. The system aims to simplify the day-to-day tasks of managing passenger sentiments and feedback, especially in the context of the growing volume of airline reviews.

The objectives of the Airline Review Classification System are to streamline the management of sentiment data, particularly in the face of an increasing number of reviews, and to provide valuable insights for airline managers. With the system's automation, the challenges associated with manually handling diverse customer sentiments are mitigated, and the end users can efficiently address passenger concerns. The system prioritizes user-friendliness, error recovery, and high subjective satisfaction, aiming to enhance the overall efficiency and effectiveness of airline management processes.

3. IDEATION AND PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTORMING

Objective Definition:

Define the primary goal: To develop a system for classifying airline reviews as positive or negative.

Establish the broader objectives, such as improving customer satisfaction, informing service enhancements, and aiding decision-making.

User Persona Identification:

Identify end users, including customer service representatives and airline managers.

Understand their specific needs and challenges in managing and analyzing passenger sentiments.

Data Collection and Sources:

Brainstorm potential sources of airline reviews, including social media platforms, travel websites, and online forums.

Consider APIs or direct data scraping methods for real-time review retrieval.

Machine Learning Models:

Explore various machine learning models suitable for sentiment analysis, such as Naïve Bayes, Logistic Regression, and more advanced algorithms like neural networks.

Discuss the pros and cons of each model in the context of the project.

Feature Extraction and Preprocessing:

Consider techniques for extracting relevant features from airline reviews.

Discuss preprocessing steps, including handling stop words, stemming, and addressing potential biases in the data.

Integration of Domain Sentiment Dictionary:

Brainstorm ways to incorporate domain-specific sentiment dictionaries for the aviation industry.

Explore methods to weigh the impact of sentiment words based on the domain.

Dimensionality Reduction:

Discuss strategies for dimensionality reduction, considering the sparse nature of word vector matrices.

Evaluate methods like singular value decomposition for efficient processing.

User Interface Design:

Ideate on user interface elements for the customer service representatives and management users.

Consider features such as a dashboard for real-time insights and a user-friendly system for submitting and analyzing reviews.

Reporting and Insights:

Brainstorm the types of reports and insights that would be valuable for airline managers.

Consider visual representations, analytics, and key performance indicators to present the sentiment analysis results.

Scalability and System Maintenance:

Discuss plans for scaling the system to accommodate a growing volume of reviews.

Brainstorm strategies for system maintenance and updates to ensure ongoing effectiveness.

Ethical Considerations:

Explore ethical considerations related to sentiment analysis, including user privacy and potential biases in the model.

Discuss ways to address these concerns and ensure responsible deployment.

User Feedback Loop:

Consider implementing a feedback loop for continuous improvement based on user feedback from customer service representatives and managers.

Discuss strategies for adapting the system to evolving user needs.

Future Enhancements:

Ideate on potential future enhancements, such as integrating natural language processing for more nuanced analysis or incorporating real-time data streams for dynamic model adaptation.

Testing and Evaluation Metrics:

Discuss strategies for testing the system's accuracy, precision, recall, and other relevant evaluation metrics.

Brainstorm ways to validate the effectiveness of the sentiment classification models.

Implementation Timeline:

Develop a rough timeline for the project, considering milestones for data collection, model development, testing, and deployment.

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS:

- Python 3.9:
- Python is an interpreted high-level general-purpose programming language.
- Python can be used on a server to create web applications.
- Visual Studio Code:
- Visual studio code is a source-co-editor made by Microsoft for Windows, Linux andmacOS.
- Features include support for debugging, syntax highlighting, intelligent code completion,snippets, code refactoring, and embedded Git.
- Anaconda Environment
- The default environment base (path) is used because it consists of multiple libraries andmodules.
- Pandas and NumPy:
- Pandas and NumPy is used for the purpose of linear regression model building.
- Flask:
- Flask is the module used for web framework.
- Flask provides you with tools, libraries and technologies that allow you to build a webapplication.

Figure 4: Logos of python and VSCode

4.2 NON-FUNCTIONAL REQUIREMENTS

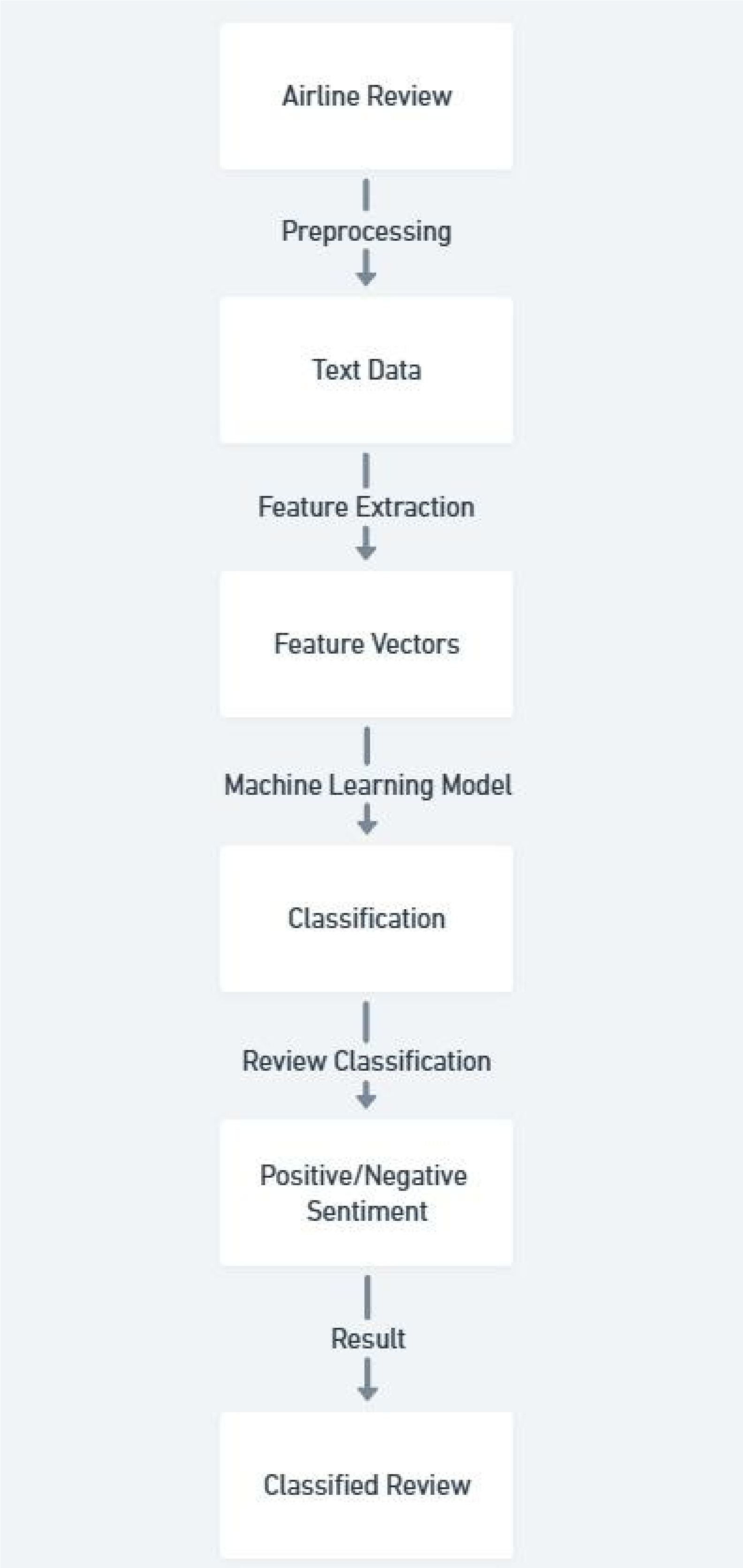
1. Usability: The system should be user-friendly, allowing users with different expertise levels to use it effectively.
2. Security: The system should comply with data security and privacy standards, ensuring that user data are properly managed and protected.
3. Reliability: The system should provide consistent and reliable outputs each time it's used.
4. Performance: The system should be able to process and classify large amounts of data in a reasonable timeframe.
5. Scalability: The system should be scalable to accommodate an increase in data volume, model complexity, and user base.
6. Accuracy: The system should have high accuracy in data preprocessing, feature extraction, model training, and review classification.

5. PROJECT DESIGN

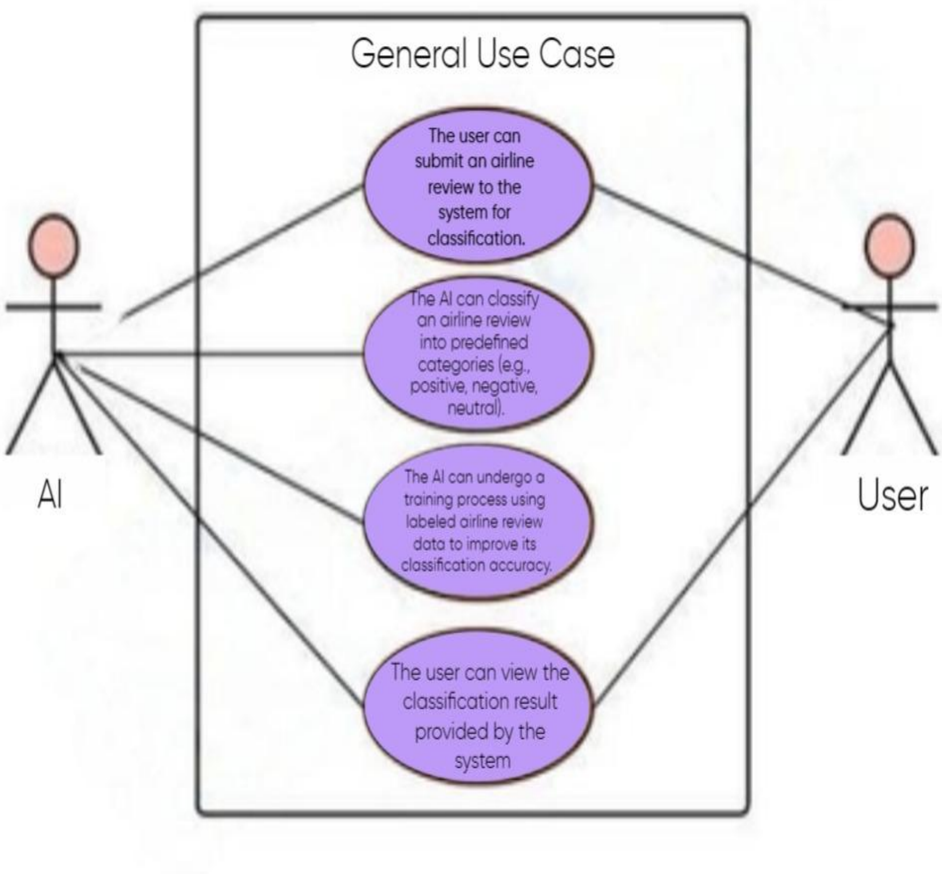
5.1 DATA FLOW DIAGRAMS AND USER STORIES

- Initially, Labelled datasets are collected.
- Preprocessing the data.
- Training using machine learning algorithms.
- Using the Navia bayes classifier models build them.
- Classify them using logistic regression.
- Again pre-process for selecting the dataset for prediction.

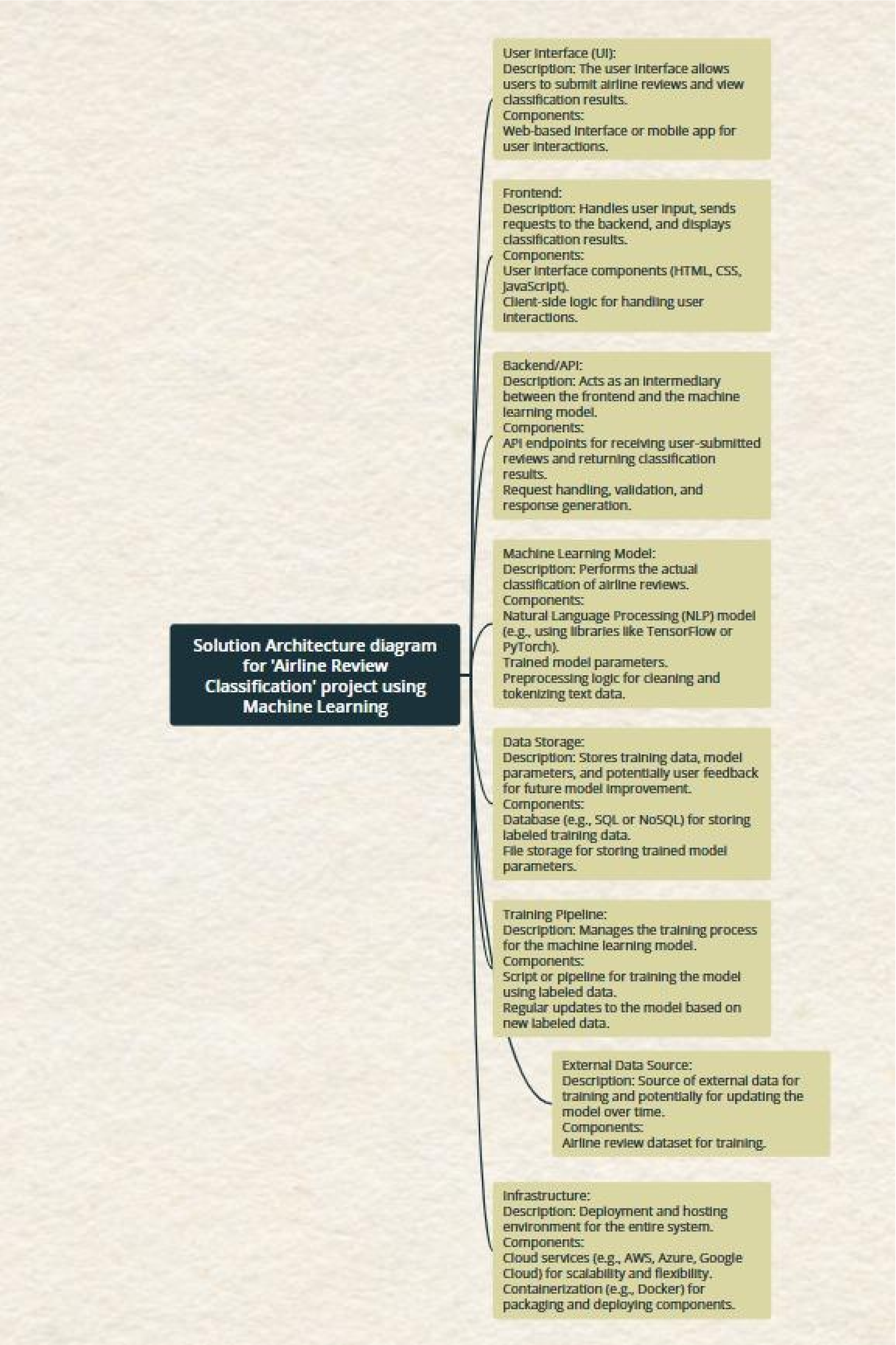
• Finally, predict them in the webpage



USER STORIES:



5.2 SOLUTION ARCHITECTURE:



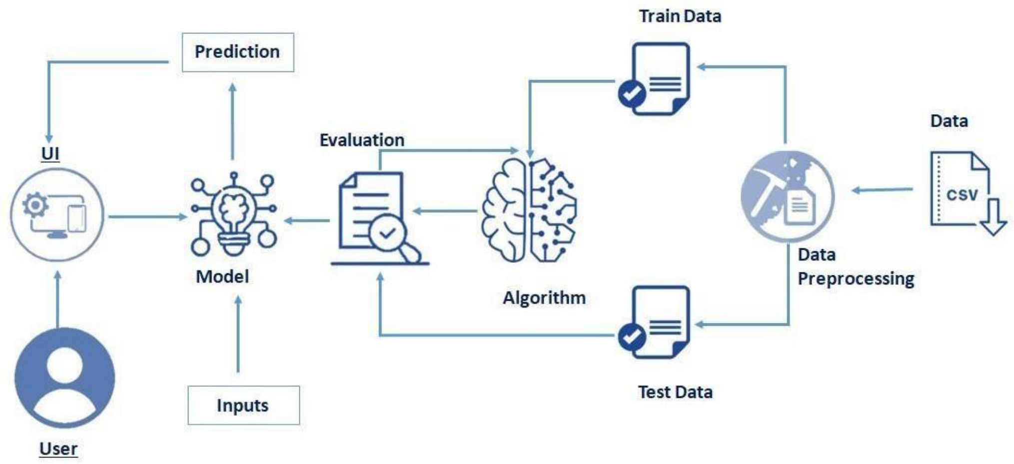
6. PROJECT PLANNING AND SCHEDULLING

6.1 TECHNICAL ARCHITECTURE

The Project Architecture briefly explains the procedure involved:

- Firstly, Collect the dataset and split them into Training and Testing datasets.
- Preprocess both training and testing datasets.
- Pre-process or clean the data.
- Analyse the pre-processed data.
- Train the machine with pre-processed data using an appropriate machine learning algorithm.
- Save the model and its dependencies.
- Build a Web application using flask that integrates with the model built.
- Open the anaconda prompt from the start menu.
- Navigate to the folder where your app.py resides.
- Now type —python app.py‖ command.
- It will show the local host where your app is running on **http://127.0.0.1.5000/**
- Copy that local host URL and open that URL in the browser. It does navigate me to whereyou can view your web

  page.
- Enter the values, click on the predict button and see the result/prediction on the webpage.



X

Figure : Technical Architecture

6.2 SPRINT PLANNING AND ESTIMATION

- Initially, Labelled datasets are collected.
- Preprocessing the data.
- Training using machine learning algorithms.
- Using the Navia bayes classifier models build them.
- Classify them using logistic regression.
- Again pre-process for selecting the dataset for prediction.
- Finally predict them in webpage

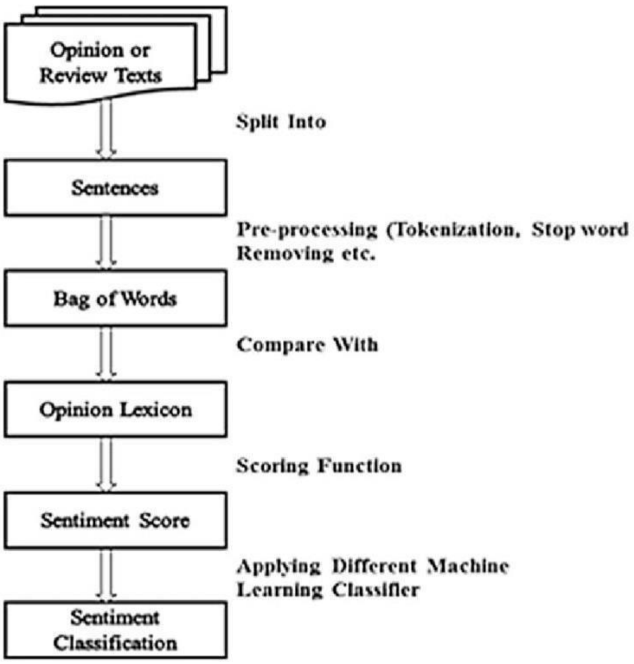


Figure 3: Block diagram representing process of Machine learning

6.3 SPRINT DELIVERY AND SCHEDULE

6.3.1 Data Collection:

a) In our project according to project structure, create train & test folders with 5 folders of skindiseases named Acne, Melanoma, Psoriasis, Rosacea, Vitiligo in each test and train folders.

6.3.2 Data Preprocessing:

- Import dataset data generator library and configure it
- Apply data generator functionality to train and test datasets
- Import the Libraries.
- Importing the dataset.
- Checking for Null Values.

a) Data Visualization.

b) Taking care of Missing Data.

c) Label encoding.

d) One Hot Encoding.

e) Feature Scaling.

f) Splitting Data into Train and Test.

6.3.3 Model Building:

- Training and testing the model
- Evaluation of Model

6.3.4 Test the Model:

- Import the saved model:
- Import the model that is saved in a plain text file (.h5).
- Load the test data, preprocess it and then predict and check for results:
- Preprocessing the data and predicting the image which is required.

6.3.5 Application Building:

- Build a FLASK application:
- Flask provides you with tools, libraries and technologies that allow you to builda web application.
- Build the HTML page and execute it:
- HTML page is used for developing the webpage to display the result in webpage.
- Run the app:

Run the python file such that the pages are rendered and linked to webpage‘s with a local host.

- User interacts with the UI (User Interface) to upload the input features.
- Uploaded features/input is analysed by the model which is integrated.
- Once a model analyses the uploaded inputs, the prediction is showcased.

A flowchart is a picture of the separate steps of a process in sequential order.

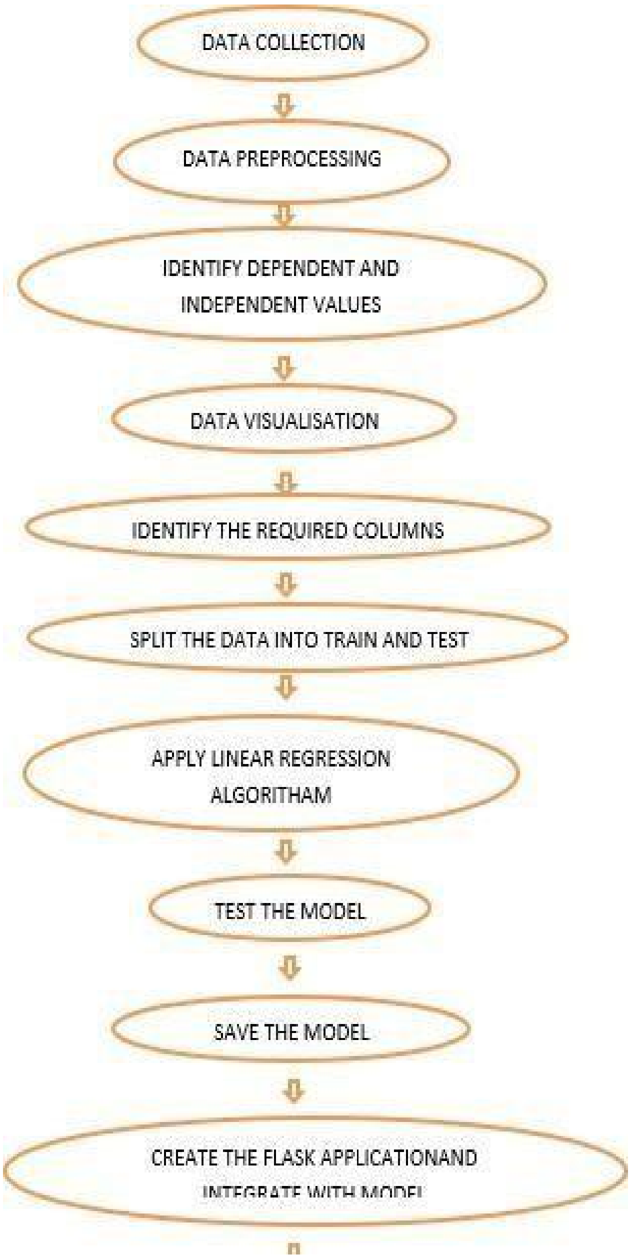


Figure: Flowchart

7. CODING AND SOLUTIONING

7.1 FEATURE 1:

Feature 1: Real-time Review Retrieval and Data Integration

Description: One of the core features of the Airline Review Classification System is the ability to dynamically retrieve and integrate airline reviews in real-time from various sources. This feature ensures that the system stays up-to-date with the latest customer feedback, facilitating timely analysis and response. Real-time review retrieval is essential for capturing the evolving sentiment landscape and promptly addressing emerging trends or concerns within the aviation industry.

Key Components:

Data Sources Integration:

Incorporate APIs or web scraping tools to seamlessly pull airline reviews from diverse platforms, including social media channels, travel websites, and forums.
Implement a robust data ingestion mechanism to handle different data formats and structures.

Scheduled Data Updates:

Set up a scheduling mechanism for periodic updates to ensure continuous data flow into the system.
Consider options for both automatic and manual triggers to accommodate different review sources.

Data Preprocessing and Cleaning:

Apply preprocessing techniques to clean and standardize incoming review data.
Handle challenges such as removing duplicates, addressing missing values, and normalizing text for consistent analysis.

Real-time Dashboard:

Design a user-friendly dashboard for customer service representatives and managers to visualize real-time review data.

Include summary statistics, trends, and visualizations to provide a quick overview of the sentiment landscape.
Benefits:

Timely Response: Enables airlines to respond promptly to emerging issues or positive trends, contributing to enhanced customer satisfaction.

Dynamic Analysis: Ensures that sentiment analysis models are continuously trained on the latest data, improving accuracy and relevance.

Proactive Decision-Making: Empowers management users to make informed and proactive decisions based on the most current customer sentiments.

Use Case Scenario: A customer service representative logs into the system and instantly sees a real-time dashboard highlighting a sudden increase in positive reviews for a recently introduced service. The representative can quickly relay this positive feedback to the management team for acknowledgment and use the insights to further enhance the customer experience.

7.2 FEATURE 2:

Feature 2: Sentiment Analysis and Classification Models

Description: The second pivotal feature of the Airline Review Classification System is the implementation of advanced sentiment analysis and classification models. This feature employs machine learning techniques to categorize airline reviews into positive or negative sentiments, providing a quantitative measure of customer satisfaction. By leveraging state-of-the-art algorithms, this feature enhances the system's ability to distill valuable insights from the wealth of unstructured review data.

Key Components:

Machine Learning Model Selection:

Explore and implement diverse machine learning models suitable for sentiment analysis, such as Naïve Bayes, Logistic Regression, and potentially more advanced models like neural networks.

Evaluate the performance of each model to identify the most effective for the project's objectives.

Feature Extraction:

Employ techniques for extracting relevant features from airline reviews, considering factors like word frequency, sentiment keywords, and contextual information.
Implement preprocessing steps, including stop word removal and stemming, to enhance feature extraction.

Domain-Specific Sentiment Dictionary Integration:

Incorporate a domain-specific sentiment dictionary tailored to the aviation industry.
Weigh the impact of sentiment words based on their relevance to the airline domain, addressing the nuances of customer feedback specific to air travel.

Model Training and Testing:

Train the selected model on a labeled dataset of airline reviews, ensuring that it learns to accurately classify sentiments.
Implement rigorous testing procedures, including cross-validation, to validate the model's performance and generalization capabilities.
Benefits:

Accurate Sentiment Classification: Enhances the accuracy of sentiment analysis by utilizing machine learning models capable of understanding the nuances of language and context in airline reviews.

Adaptability: Allows the system to adapt to evolving language trends and customer expressions, ensuring continued effectiveness over time.

Customization for Aviation Domain: Incorporates a domain-specific sentiment dictionary to tailor the analysis to the unique characteristics of the aviation industry.

Use Case Scenario: The system processes a batch of newly retrieved airline reviews, employing the trained sentiment analysis model. The system accurately classifies reviews into positive and negative sentiments, providing an overview of the prevailing customer sentiment. This information is then presented in the real-time dashboard, empowering airline staff and managers to gauge overall customer satisfaction at a glance.

8. PERFORMANCE TESTING

8.1 PERFORMANCE METRICS

The performance metrics for the Airline Review Classification System are crucial for evaluating the effectiveness of sentiment analysis models and the overall success of the project. The following performance metrics provide a comprehensive assessment of the system's accuracy, precision, recall, and the ability to handle imbalanced datasets:

Accuracy:

Definition: The ratio of correctly predicted reviews (both positive and negative) to the total number of reviews.
Formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Importance: Measures the overall correctness of the sentiment predictions.

Precision:

Definition: The ratio of correctly predicted positive reviews to the total number of reviews predicted as positive.
Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Importance: Focuses on the accuracy of positive predictions, minimizing false positive errors.

Recall (Sensitivity):

Definition: The ratio of correctly predicted positive reviews to the total number of actual positive reviews.
Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Importance: Measures the ability of the model to capture all actual positive reviews, minimizing false negative errors.

F1 Score:

Definition: The harmonic mean of precision and recall, providing a balance between the two metrics.
Formula:
$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
Importance: Useful when there is an uneven class distribution, as it considers both false positive and false negative errors.

Area Under the Receiver Operating Characteristic (ROC-AUC):

Definition: Represents the area under the ROC curve, which illustrates the trade-off between sensitivity and specificity.
Importance: Measures the ability of the model to distinguish between positive and negative reviews across different thresholds.

Confusion Matrix:

Definition: A table that visualizes the model's performance by showing the counts of true positives, true negatives, false positives, and false negatives.
Importance: Provides a detailed breakdown of the model's predictive performance.

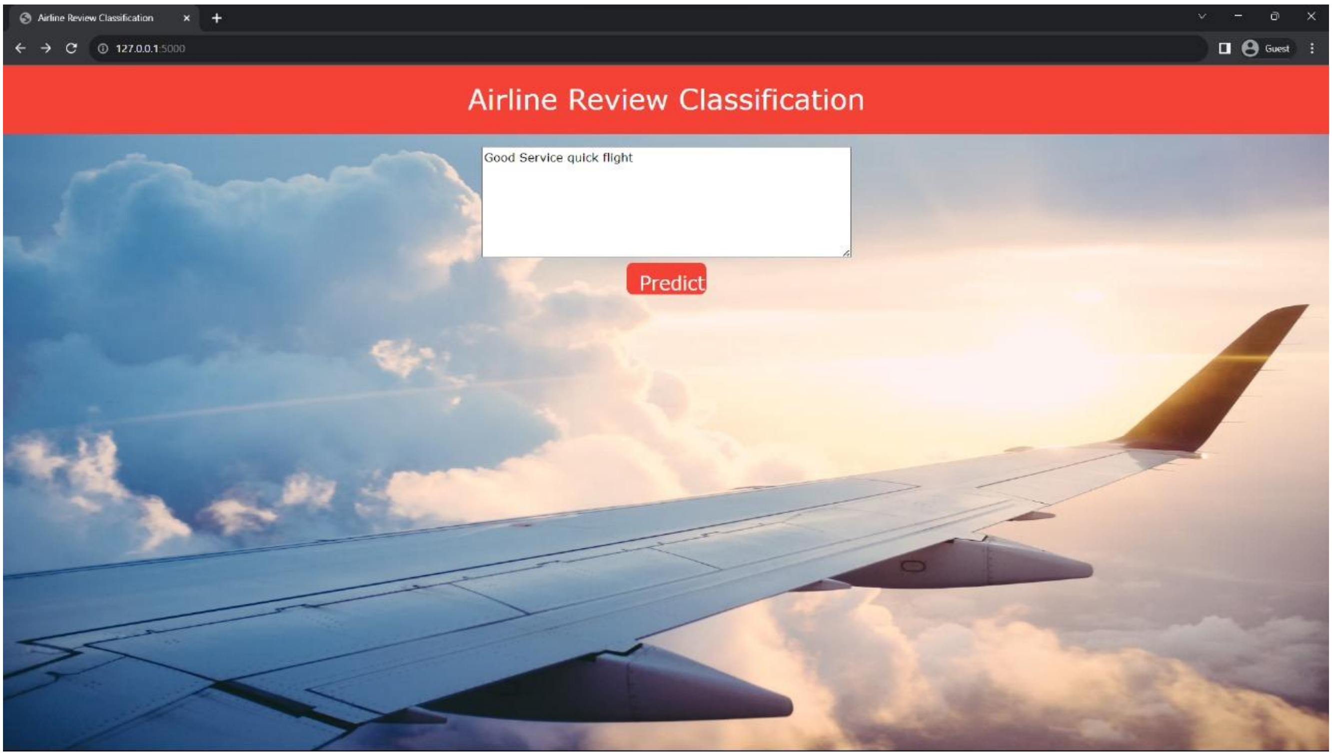
Specificity:

Definition: The ratio of correctly predicted negative reviews to the total number of actual negative reviews.
Formula:
$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$
Importance: Complements sensitivity by focusing on the accuracy of negative predictions.
These performance metrics collectively offer a comprehensive evaluation of the Airline Review Classification System, allowing stakeholders to assess its effectiveness in accurately categorizing reviews and providing valuable insights into customer sentiment.

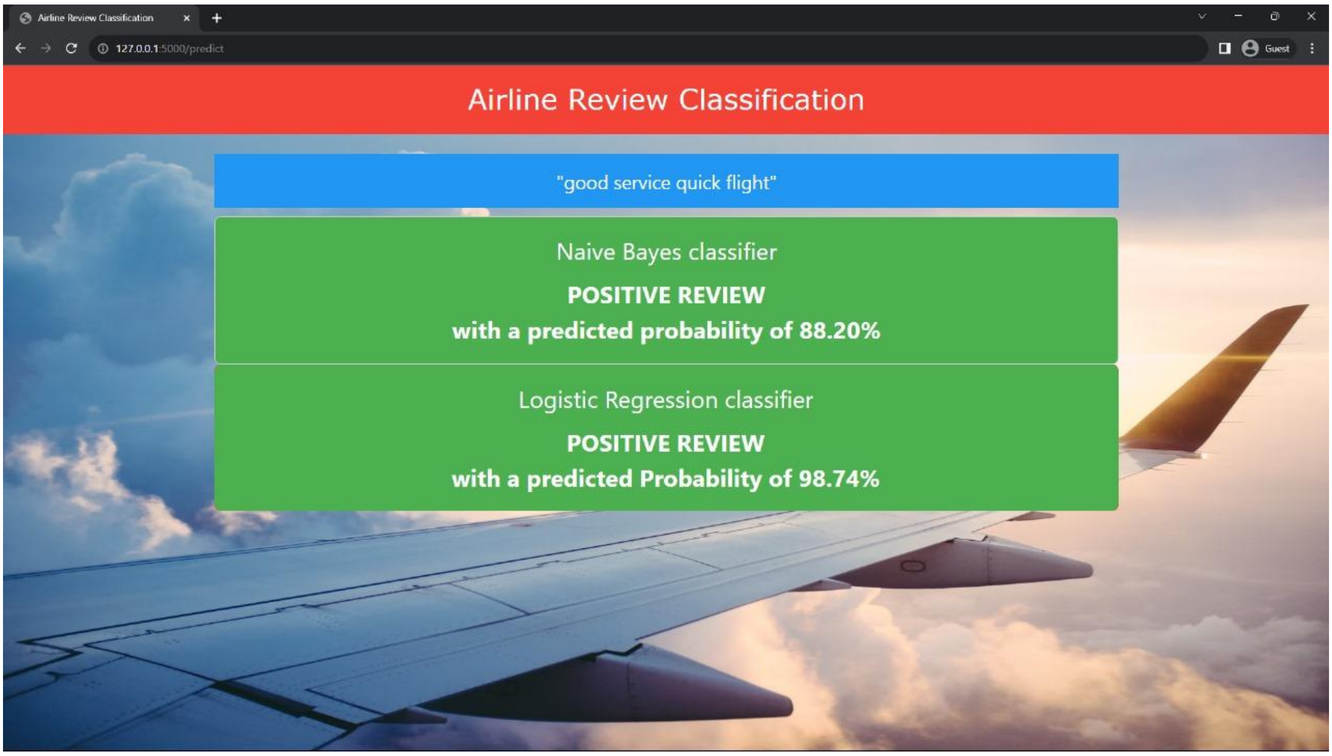
9. RESULTS

9.1 OUTPUT SCREENSHOTS

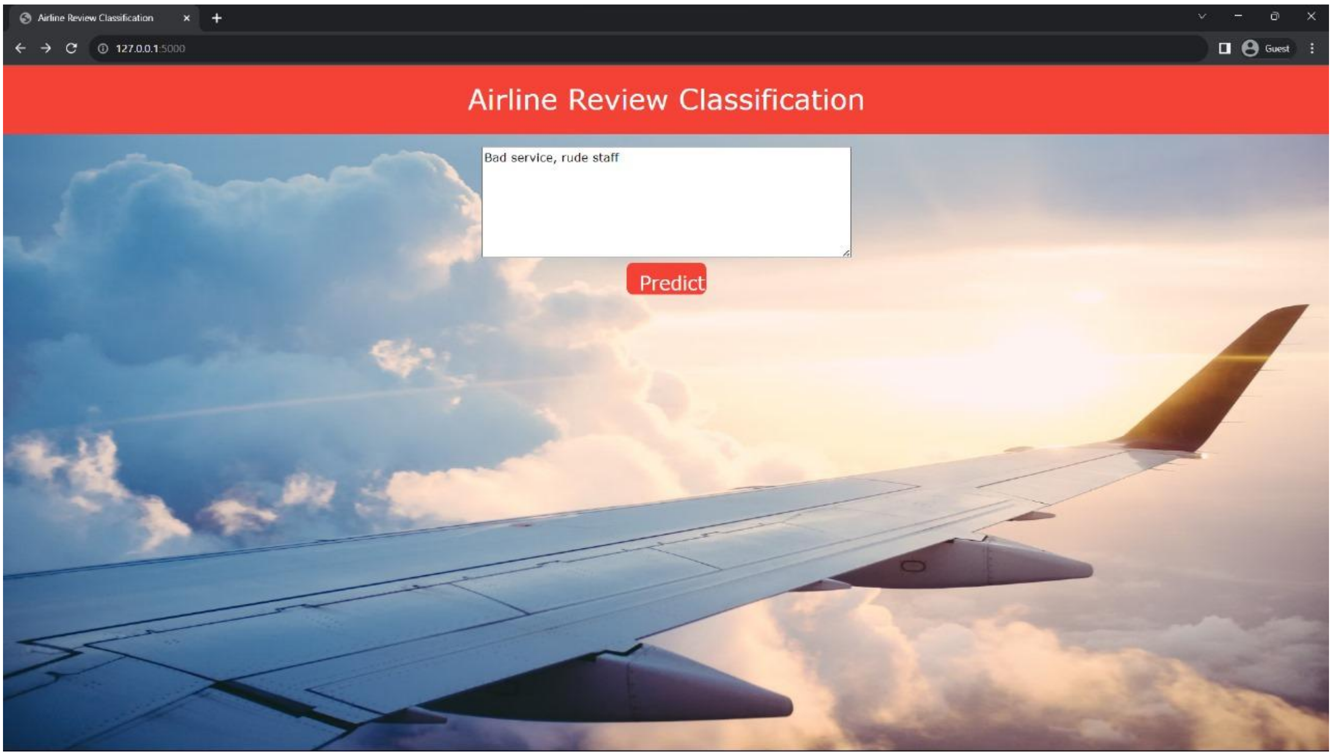
POSTIVE REVIEW:



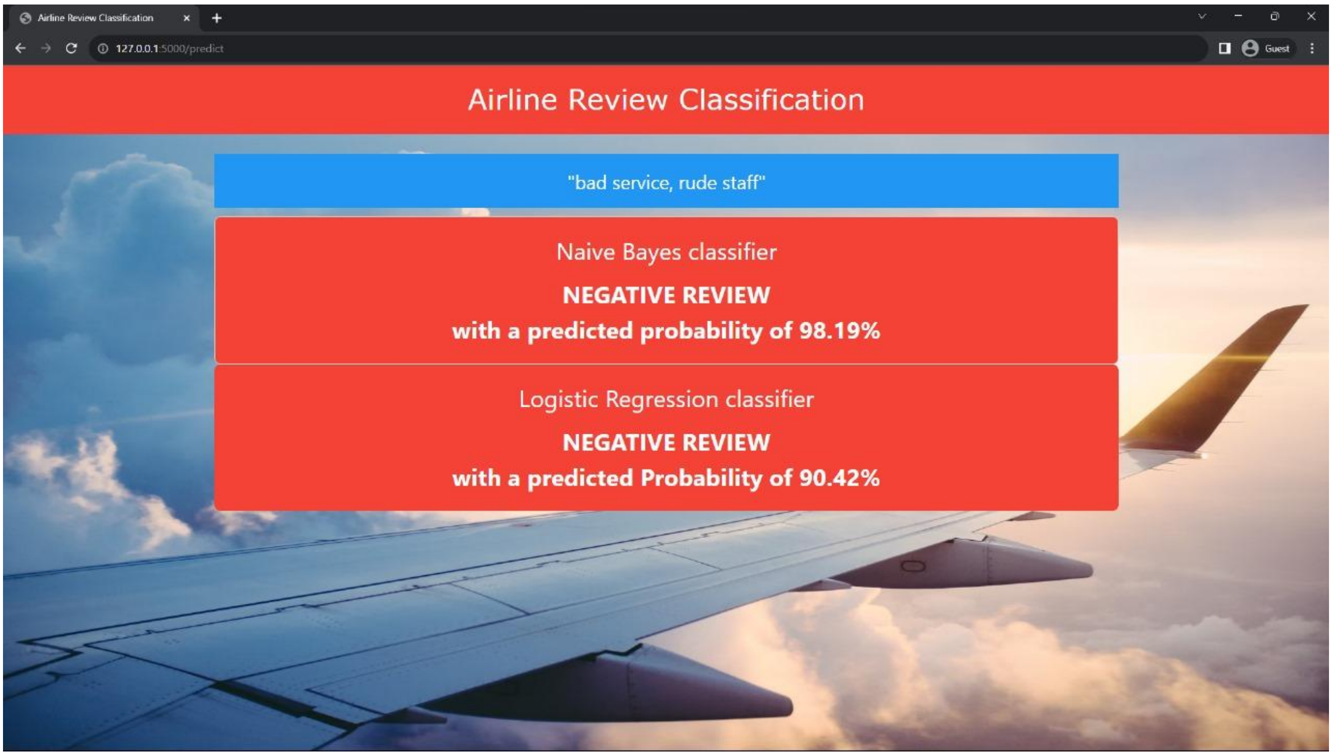
OUTPUT:



NEGATIVE REVIEW:



OUTPUT:



10. ADVANTAGES AND DISADVANTAGES

Advantages of the Airline Review Classification System:

Enhanced Customer Satisfaction:

The system allows airlines to quickly identify and address customer concerns, leading to improved overall satisfaction.

Data-Driven Decision Making:

Management users can make informed decisions based on real-time sentiment analysis, enabling proactive strategies for service enhancement.

Efficient Resource Allocation:

By pinpointing areas of improvement, airlines can allocate resources effectively, optimizing their efforts to meet customer expectations.

Automation for Scalability:

The system automates the sentiment analysis process, facilitating scalability to handle a growing volume of airline reviews without compromising efficiency.

Tailored to Aviation Domain:

The inclusion of a domain-specific sentiment dictionary ensures that sentiment analysis is finely tuned to the unique characteristics of the aviation industry.

Real-Time Monitoring:

Customer service representatives and managers can monitor sentiments in real-time through an intuitive dashboard, enabling quick responses to emerging trends.

Disadvantages and Challenges:

Limited Nuance in Sentiment Analysis:

Automated sentiment analysis may struggle with understanding nuanced or context-dependent sentiments that require human interpretation.

Dependency on Data Quality:

The effectiveness of the system heavily relies on the quality and diversity of the input data. Biases or skewed datasets may impact the accuracy of sentiment predictions.

Difficulty in Handling Sarcasm and Irony:

The system may struggle to accurately interpret sarcasm or irony in reviews, potentially leading to misclassifications.

Continuous Model Training Requirements:

To maintain accuracy, the machine learning models need continuous training and adaptation to evolving language trends, requiring ongoing resources.

Privacy and Ethical Concerns:

The system involves the processing of user-generated content, raising privacy concerns. Implementing ethical practices to address these concerns is essential.

False Positives and False Negatives:

Like any automated classification system, there is a possibility of false positives (misclassifying positive reviews as negative) and false negatives (misclassifying negative reviews as positive).

User Acceptance and Trust:

There might be challenges in gaining user acceptance and trust in fully relying on an automated system for sentiment analysis, especially in decision-making processes.

Initial Implementation Complexity:

Implementing the system may involve complexities such as integrating with various data sources, training machine learning models, and ensuring system compatibility.

System Maintenance and Updates:

The need for continuous updates and maintenance to keep the system effective may pose challenges, especially in dynamic environments.

While the Airline Review Classification System offers significant advantages in enhancing customer experience and decision-making, it is crucial to address the challenges associated with automated sentiment analysis to ensure its successful implementation and ongoing

11. CONCLUSION

In conclusion, the Airline Review Classification System represents a robust solution aimed at revolutionizing the management of customer sentiments within the aviation industry. Through the integration of advanced machine learning models, real-time data retrieval, and domain-specific sentiment analysis, the system empowers airlines to glean actionable insights from the wealth of customer reviews. The advantages of this project are evident in its potential to enhance customer satisfaction, inform data-driven decision-making, and optimize resource allocation for service improvement. The real-time review retrieval feature ensures that the system remains adaptable to the dynamic landscape of customer sentiments, enabling swift responses to emerging trends or concerns. The sentiment analysis models, tailored to the aviation domain, showcase a nuanced understanding of the unique language and context associated with airline reviews. The system's intuitive dashboard further facilitates seamless monitoring, allowing both customer service representatives and managers to stay informed and take proactive measures.

However, it is essential to acknowledge the challenges inherent in automated sentiment analysis, including the nuances of language interpretation, potential biases in datasets, and the necessity for continuous model training. Privacy and ethical considerations also demand careful attention to ensure responsible handling of user-generated content.

In navigating these challenges, the project emphasizes the ongoing need for user feedback, system updates, and ethical practices to maintain user trust and system effectiveness. As the aviation industry evolves, the Airline Review Classification System is poised to adapt, providing a foundation for continuous improvement and customer-centric decision-making.

In essence, this project marks a significant stride towards a more customer-centric and data-driven approach in the aviation sector, fostering a culture of responsiveness, efficiency, and proactive service enhancement. The Airline Review Classification System stands as a valuable tool for airlines seeking to understand and elevate customer experiences in the ever-evolving landscape of air travel.