PROJECT REPORT

Date	7 NOVEMBER 2023
Team ID	592337 (Satvik Marwah and Himanshu Rana)
Project Name	Airline Review Classification Using Machine Learning

1. INTRODUCTION

1.1 Project Overview

In today's globalized world, the airline industry plays a pivotal role in connecting people, facilitating travel, and driving international business. As air travel continues to become more accessible, the quality of service provided by airlines is a critical factor in shaping passenger experiences. This project aims to develop an airline review classification system using various machine learning classification models, such as Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier, to analyze and classify airline reviews and determine whether a flight is recommended or not based on passenger feedback.

1.2 Purpose

The primary purpose of the "Airline Review Classification System for Flight Recommendation" project is to leverage machine learning and natural language processing techniques to serve the following key objectives:

Passenger Satisfaction Analysis: To analyze and classify airline reviews to determine whether a flight is recommended or not, based on the feedback provided by passengers. This purpose directly addresses the need to understand and improve passenger satisfaction, which is crucial for any airline's success.

Service Enhancement: To provide actionable insights to airlines on areas where they can improve their services and overall passenger experiences. By identifying specific aspects of flights that lead to recommendations or dissatisfaction, airlines can make data-driven decisions to enhance their offerings.

Customer Loyalty: To contribute to the growth of customer loyalty and retention by helping airlines identify and rectify issues that may lead to negative passenger sentiments. Happy passengers are more likely to become repeat customers, leading to long-term business success for airlines.

Data-Driven Decision-Making: To promote data-driven decision-making within the airline industry. The project aims to show how data analysis and machine learning can be harnessed to make informed choices regarding service improvements, marketing strategies, and resource allocation.

Competitive Advantage: To provide airlines with a competitive advantage by leveraging insights gained from the analysis of user-generated content. Airlines that can better understand their customers' needs and sentiments are better positioned to outperform their competitors.

Efficiency and Cost Reduction: By automating the review classification process, the project can help airlines efficiently process and analyze a large volume of reviews. This can lead to cost savings and improved resource allocation for addressing specific issues.

Improved Passenger Experiences: Ultimately, the project's purpose is to contribute to the overall improvement of passenger experiences in the airline industry. Satisfied passengers are more likely to promote an airline through word of mouth and positive online reviews, which can lead to increased business and a positive brand image.

In summary, the purpose of this project is to harness the power of machine learning and natural language processing to enhance the airline industry's ability to understand and respond to passenger feedback, thereby improving service quality, customer satisfaction, and overall business performance.

2. LITERATURE SURVEY

2.1 Existing problem

The existing problem in the field of sentiment analysis and recommendation systems for airlines revolves around the analysis of user-generated content, particularly airline reviews. Several challenges and issues have been identified:

Unstructured Data: Airline reviews typically consist of unstructured text data, making it challenging to extract meaningful insights from the vast amount of information available on the internet.

Sentiment Analysis: Determining whether a review is positive, negative, or neutral can be complex due to the nuanced and context-dependent nature of language. Identifying the underlying sentiment is critical for assessing passenger satisfaction.

Classification Accuracy: Existing classification models may not always provide accurate recommendations or satisfaction ratings. Improving the accuracy of these models is essential for actionable insights.

Feature Selection: Selecting relevant features or attributes from reviews that influence passenger recommendations is a challenge. It's crucial to identify which aspects of the flight experience are most significant.

Real-time Analysis: Airlines need real-time analysis to address immediate issues and trends, but traditional approaches may not provide timely insights.

2.2 References

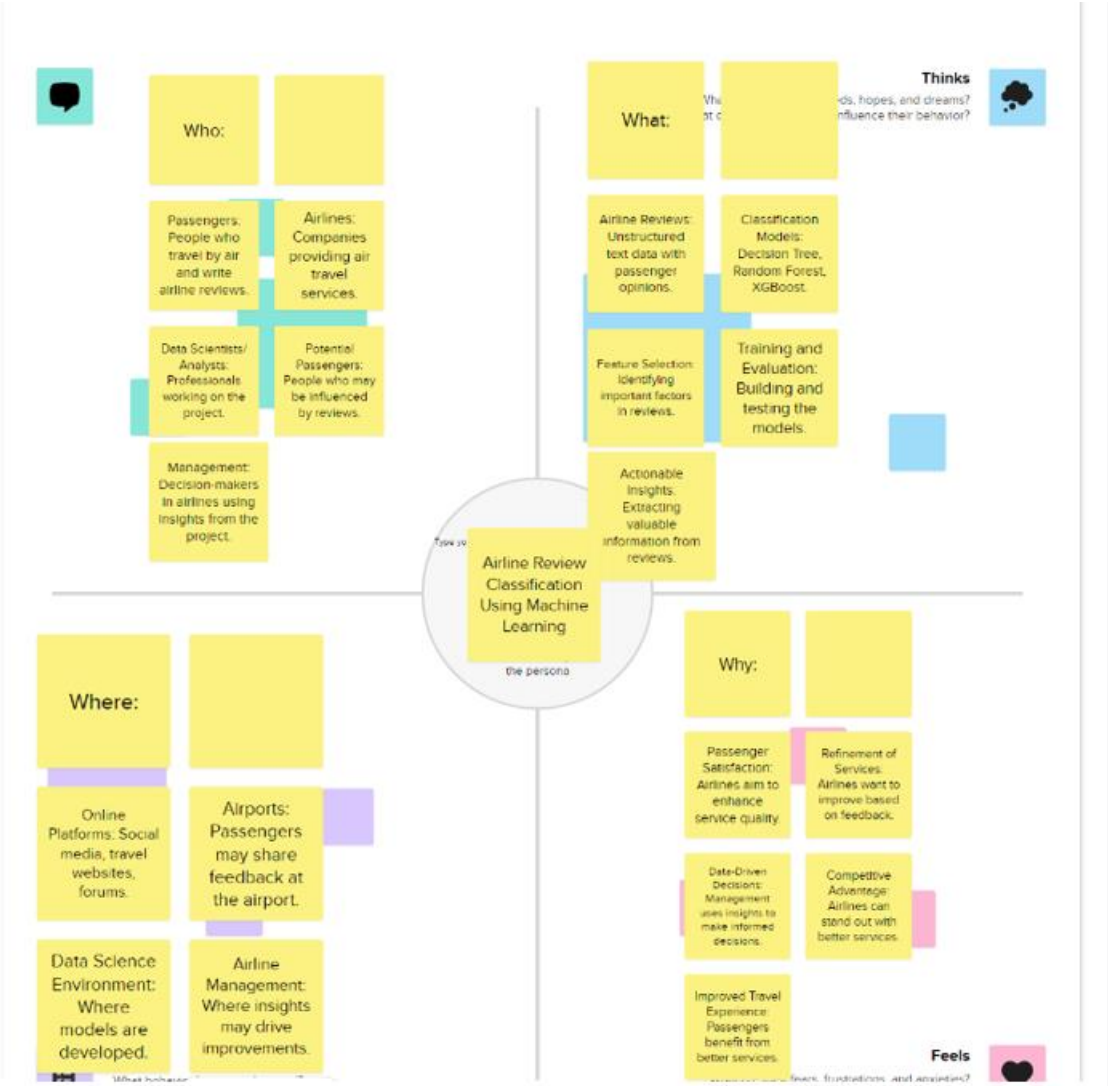
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. CRC Press.
Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

2.3 Problem Statement Definition

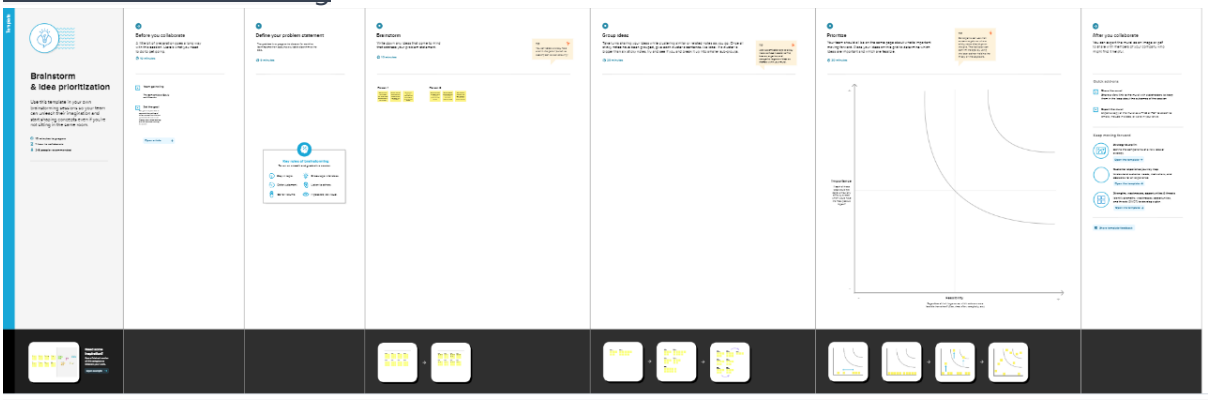
Develop a machine learning-based classification system that can process a wide range of airline reviews, extract valuable insights, and classify the reviews into two categories: "Recommended" and "Not Recommended." This system should consider the nuanced language used in reviews, identify the underlying sentiment accurately, and provide airlines with actionable recommendations for improving passenger satisfaction.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy map canvas



3.2 Ideation & Brainstorming



4. REQUIREMENT ANALYSIS

4.1 Functional requirement

1.Data Collection and Preprocessing:

The system shall collect airline reviews from various online sources.
It shall preprocess the raw text data to remove irrelevant information, perform text normalization, and handle missing data.

2.Feature Selection:
The system shall identify and select relevant features from the reviews to be used for classification.
It shall consider factors such as sentiment, keywords, and specific aspects of the flight experience.

3.Classification Models:
The system shall implement multiple classification models, including Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier.
It shall allow for the training and evaluation of these models using labeled data.

4.Real-time Analysis:
The system shall provide the capability to perform real-time or near-real-time analysis of incoming reviews.
It shall update its classification and recommendation predictions as new data becomes available.

5.User Interface:
The system shall have a user-friendly interface for data input and model evaluation.
It shall display classification results and recommendations in an understandable format.

6.Insight Generation:
The system shall generate insights and reports on the factors influencing flight recommendations.
It shall offer actionable recommendations for airlines based on the analysis.

4.2 Non-Functional requirements

1.Scalability:
The system shall be able to handle a large volume of reviews from different sources without performance degradation.

2.Accuracy:
The classification models shall achieve a high level of accuracy in determining flight recommendations, with a target accuracy rate of 90% or higher.

3.Real-time Responsiveness:
Real-time analysis should not exceed a specified time limit (e.g., 1 second) to ensure timely responses to new data.

4.Security:
The system shall ensure data privacy and security by implementing robust data encryption and access controls for sensitive information.

5.Robustness:
The system shall be resilient to noisy data, outliers, and potential biases in reviews to maintain the quality of insights.

6.Interoperability:
The system shall be compatible with various data sources and data formats, ensuring flexibility in data collection.

7.Usability:
The user interface shall be intuitive and user-friendly, requiring minimal training for users to operate the system.

8.Documentation:
The system shall have comprehensive documentation, including user manuals, technical guides, and code documentation for maintainability.

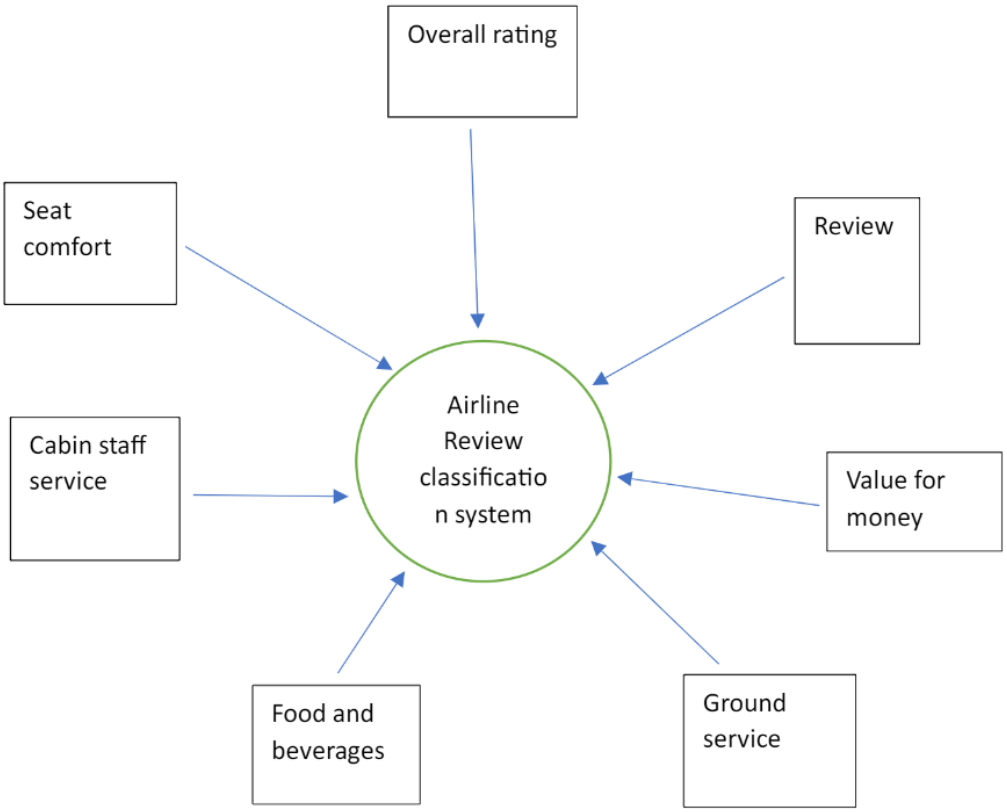
9.Scalable Infrastructure:
The system's infrastructure, including hardware and software, should be scalable to accommodate growing data volumes and user demands.

10.Maintenance and Support:
The project team shall provide ongoing maintenance and support, including addressing issues, updates, and improvements based on user feedback.

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories

DFD:



User stories:

Sprint	Functional Requirement	User story number	User story/task	Story points	Priority	Team members
Sprint-1	Registration	USN-1	As a user I can register by entering my personal information such as name, phone number and flight name.	1	High	2
Sprint-2	Review	USN-2	As a user I can enter my review of the flight which consists of rating several components as well as giving my own	1	High	2

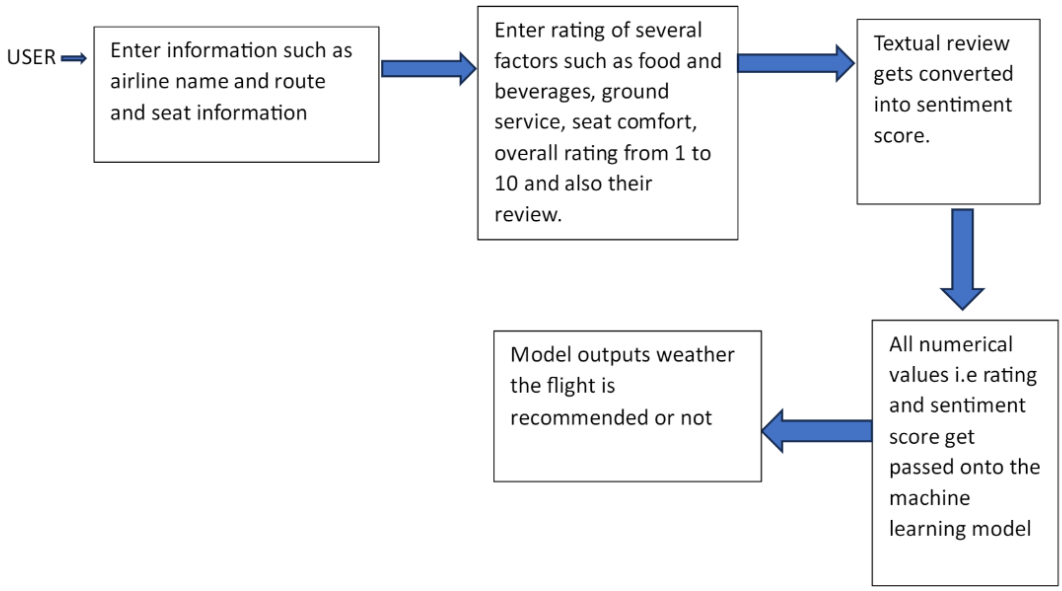
			experience of the flight			

5.2 Solution Architecture

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

1. Find the best tech solution to solve existing business problems.
2. Describe the structure, characteristics, behavior, and other aspects of the software to project stakeholders. Define features, development phases, and solution requirements.
3. Provide specifications according to which the solution is defined, managed ,and delivered.

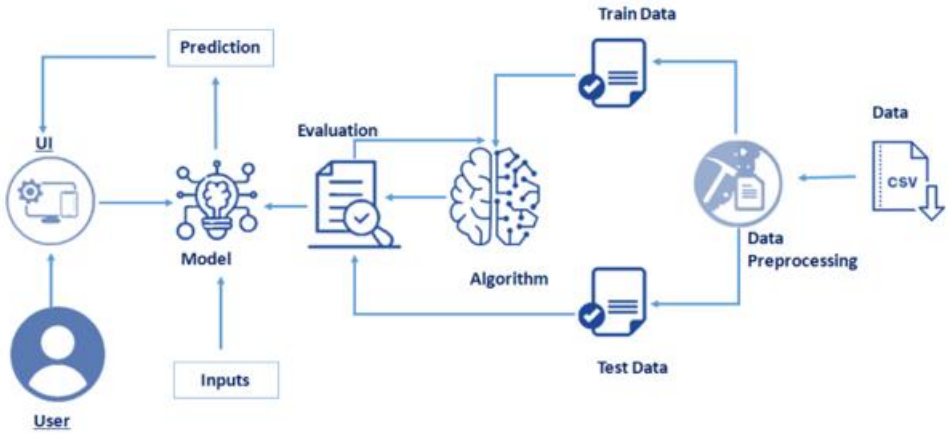
Solution Architecture Diagram:



6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture

Technical Architecture:



6.2 Sprint Planning & Estimation

7. Sprint	Functional Requirement	User story number	User story/task	Story points	Priority	Team members
Sprint-1	Registration	USN-1	As a user I can register by entering my personal information such as name, phone number and flight name.	1	High	2
Sprint-2	Review	USN-2	As a user I can enter my review of the flight which consists of rating several components as well as giving my own experience of the flight	1	High	2

6.3 Sprint Delivery Schedule

Sprint	Total story points	Duration	Sprint start date	Sprint end date (planned)	Story points completed (as on planned end date	Sprint release date actual
Sprint 1	20	1 day	25th October	26th October	20	26th October
Sprint 2	20	2 days	27th October	29th October	20	29th October

7. CODING & SOLUTIONING

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas. In pandas we have a function called read_csv() to read the dataset. As a parameter we have to give the directory of the csv file.

```
df2=pd.read_csv(r"C:\Users\Lenovo\Downloads\archive (14)\Airline_Reviews.csv")
print(df2.head())
```

We handle categorial data by one hot encoding it. We drop the columns that that too many categories to handle.

```
df4=df4.drop('Airline Name',axis=1)
df4=df4.drop('Route',axis=1)
df4=df4.drop('Date Flown',axis=1)
df4=df4.drop('Inflight Entertainment',axis=1)
```

Review column is neither numerical nor categorical. It is rather just a piece of text. Hence we combine the review and review summary column into a single text column and then calculate its sentiment score thus converting it into a numerical value.

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
def get_sentiment_score(text):
    sentiment_scores = analyzer.polarity_scores(text)
    return sentiment_scores['compound']
pd.options.mode.chained_assignment = None

df4['sentiment_score'] = df4['Review'].apply(lambda x: get_sentiment_score(x))
```

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas have a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features

	count	mean	std	min	25%	50%	75%	max
Seat Comfort	19016.0	2.618321	1.464844	0.0	1.0	3.0	4.0	5.0
Cabin Staff Service	18911.0	2.871609	1.604631	0.0	1.0	3.0	4.0	5.0
Food & Beverages	14500.0	2.553586	1.526314	0.0	1.0	2.0	4.0	5.0
Inflight Entertainment	10829.0	2.178964	1.488758	0.0	1.0	2.0	3.0	5.0
Ground Service	18378.0	2.353738	1.595747	1.0	1.0	1.0	4.0	5.0
Value For Money	22105.0	2.451120	1.594125	0.0	1.0	2.0	4.0	5.0

Now let’s split the Dataset into train and test sets. First split the dataset into X and y and then split the dataset. Here X and y variables are created. On X variable, nar is passed with dropping the target variable. And on y target variable is passed.

```
X = df4.drop('Recommended_yes', axis=1)  # Features
y = df4['Recommended_yes']  # Target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Now our data is cleaned and it’s time to build the model. We can train our data on different algorithms. We’re using 4 different classification algorithms: Decision tree, K-Nearesr neighbours, logistic regression and Random forest.

```
# Decision Tree Classifier
dt_classifier = DecisionTreeClassifier()
dt_classifier.fit(X_train, y_train)
y_pred_dt = dt_classifier.predict(X_test)

# K-Nearest Neighbors Classifier
knn_classifier = KNeighborsClassifier()
knn_classifier.fit(X_train, y_train)
y_pred_knn = knn_classifier.predict(X_test)

# Logistic Regression
logistic_regression = LogisticRegression()
logistic_regression.fit(X_train, y_train)
y_pred_lr = logistic_regression.predict(X_test)

# Random Forest Classifier
rf_classifier = RandomForestClassifier()
rf_classifier.fit(X_train, y_train)
y_pred_rf = rf_classifier.predict(X_test)
```

Connecting the model with flask:

```python
app = Flask(__name__)

@app.route('/')
def index():
    return render_template('index3.html')

@app.route('/get_recommendation', methods=['POST'])
def get_recommendation():
    overall_rating = int(request.form['overall_rating'])
    seat_comfort = int(request.form['seat_comfort'])
    cabin_staff_service = int(request.form['cabin_staff_service'])
    food_and_beverages = int(request.form['food_and_beverages'])
    ground_service = int(request.form['ground_service'])
    value_for_money = int(request.form['value_for_money'])
    seat_type = request.form['seat_type']
    review = request.form['review']

    recommendation = predict_recommendation(overall_rating, seat_comfort, cabin_staff_service,
                                            food_and_beverages, ground_service, value_for_money, review, seat_type)

    return jsonify({'recommendation': recommendation})

if __name__ == '__main__':
    app.run(debug=True,use_reloader=False)
```

8. PERFORMANCE TESTING

8.1 Performance Metrics

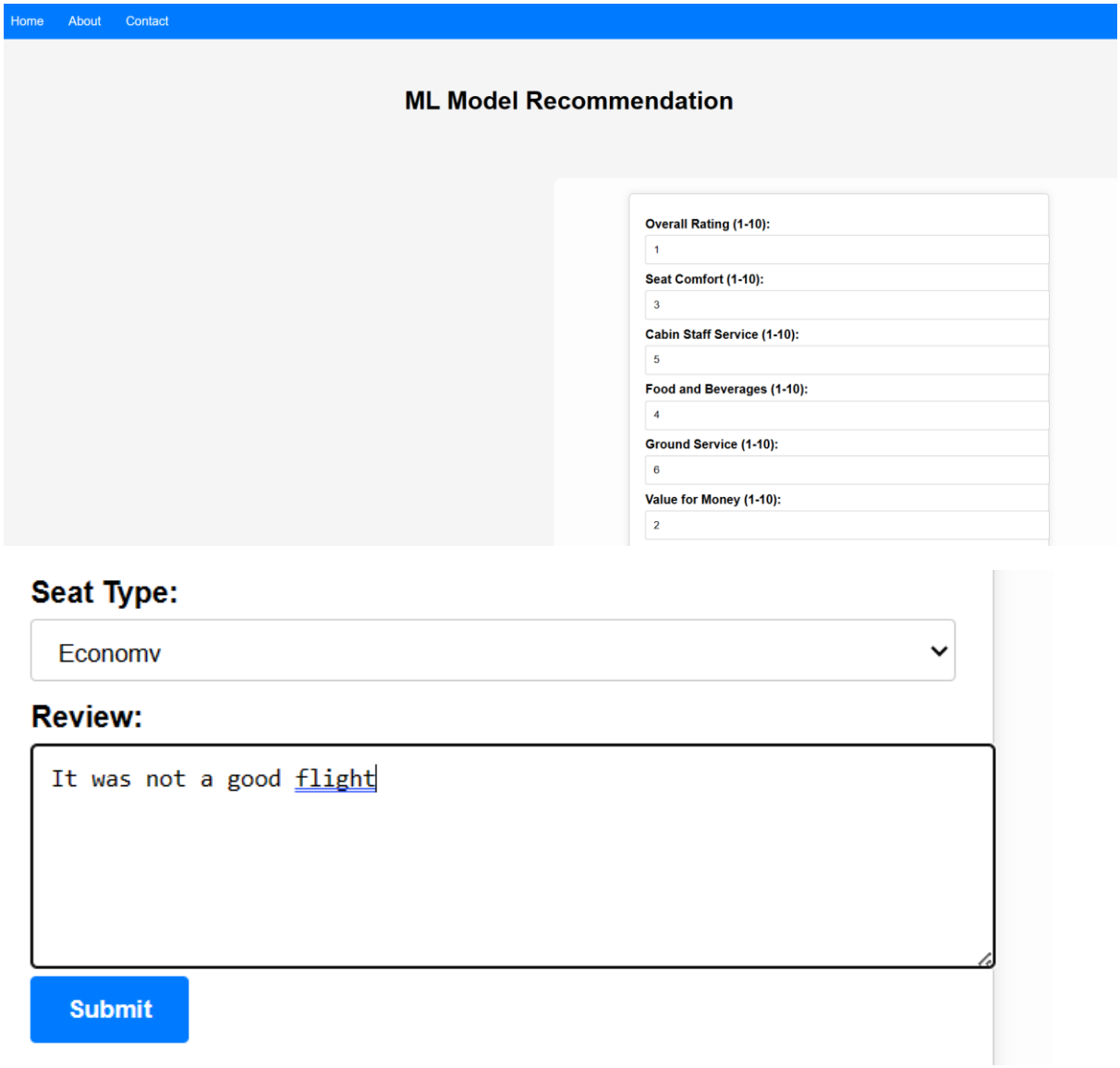
Accuracy Score: 0.9646170442286948

Classification Report:

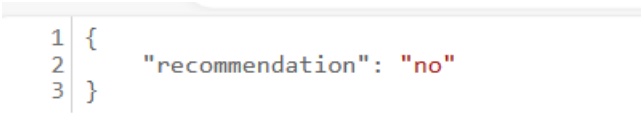
	precision	recall	f1-score	support
0	0.97	0.97	0.97	3064
1	0.95	0.94	0.95	1571
accuracy			0.96	4635
macro avg	0.96	0.96	0.96	4635
weighted avg	0.96	0.96	0.96	4635

9. RESULTS

9.1 Output Screenshots



Output:



10. ADVANTAGES & DISADVANTAGES

Advantages:

1.Improved Passenger Satisfaction: By accurately classifying airline reviews and providing recommendations, the system can help airlines identify areas for improvement. This, in turn, can lead to higher passenger satisfaction and improved customer loyalty.

2.Data-Driven Decision-Making: The system enables airlines to make data-driven decisions based on passenger feedback. This can lead to more informed choices in areas such as service enhancement, marketing strategies, and resource allocation.

3.Enhanced Customer Experience: By addressing passenger concerns and making necessary improvements, airlines can provide a better overall customer experience. Happy passengers are more likely to become repeat customers and promote the airline through positive word-of-mouth and online reviews.

4.Competitive Advantage: Airlines that utilize advanced classification systems to understand their customers' sentiments and preferences gain a competitive advantage. They can stay ahead of competitors by responding more effectively to passenger feedback.

Disadvantage: The website is not connected to a database to save customer information to further assist them and better their experience.

11. CONCLUSION

The "Airline Review Classification System for Flight Recommendation" project represents a valuable endeavor in the context of the airline industry. By leveraging machine learning and natural language processing techniques, this system provides airlines with a powerful tool for understanding passenger sentiments and feedback. The successful implementation of this system can lead to a host of benefits, including improved passenger satisfaction, data-driven decision-making, enhanced customer experiences, and competitive advantages.

Through the systematic analysis of unstructured text data from various online sources, the system can extract valuable insights and classify airline reviews into recommended and not recommended categories. These insights can guide

airlines in making informed decisions to enhance their services, address passenger concerns, and ultimately achieve long-term business success.

However, it's important to acknowledge that the effectiveness of the system depends on factors such as the quality of data, the selection of appropriate classification models, and the timely implementation of recommendations. Continuous monitoring and fine-tuning of the system are essential to maintain its relevance and effectiveness in an ever-evolving airline industry.

12. FUTURE SCOPE

Multilingual Support: Expanding the system to analyze reviews in multiple languages to cater to a broader audience.

Real-time Sentiment Monitoring: Enhancing real-time analysis capabilities to allow airlines to respond to passenger feedback in near real-time.

Personalized Recommendations: Developing personalized recommendations for passengers based on their historical feedback and preferences.

13. APPENDIX

The source code can be found in the git repo of this project and the video demonstration can be found in the google drive

Git repo link: https://github.com/smartinternz02/SI-GuidedProject-603225-1697565974

Google drive link: https://drive.google.com/drive/folders/1w_9dD7icrMa_kgqfy7y69UIFPIwIaibN?usp=sharing