Enhancing Airline Booking Prediction and Customer Sentiment Analysis Using Deep Learning and Multilingual Reviews

MSc Research Project
Programme Data Analytics

Swati Shukla
Student ID: x23280018

School of Computing
National College of Ireland

Supervisor: Prof. Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Swati Shukla
Student ID:	x23280018
Programme:	MSc Data Analytics **Year:** 2024 - 2025
Module:	Research Practicum
Supervisor:	Prof. Muslim Jameel Syed
Submission Due Date:	11th August 2025
Project Title:	Enhancing Airline Booking Prediction and Customer Sentiment Analysis Using Deep Learning and Multilingual Reviews
Word Count:	**5790** **Page Count: 20**

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: Swati Shukla

Date: 11th August 2025

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):	

Enhancing Airline Booking Prediction and Customer Sentiment Analysis Using Deep Learning and Multilingual Reviews

Swati Shukla
x23280018

Abstract

In the growing world of data, the increasing availability of airline customer data presents new opportunities for improving booking prediction accuracy using machine learning and deep learning techniques to understand intricate patterns. This research investigates the impact of integrating multilingual sentiment analysis and demographic personalization on the performance of deep learning models for predicting airline bookings. Four publicly available datasets from Kaggle — Skytrax airline reviews, Twitter airline sentiment, Expedia booking data, and passenger satisfaction data — were merged to form a comprehensive multi-source dataset covering user reviews, booking outcomes, and customer attributes.

Sentiment scores were extracted using the XLM-RoBERTa transformer model to enable effective multilingual sentiment representation. Demographic features included age band, gender, travel class, and airline preferences. These features were embedded alongside sentiment vectors to personalize the prediction task. A range of models was developed, including traditional models such as Logistic Regression, Random Forest, Support Vector Machine, and KNN for baseline comparison, as well as a series of deep learning architectures progressing from text-only to hybrid models with fine-tuning and class weighting. Due to significant class imbalance, model performance was evaluated using ROC-AUC, F1 @0.30, and confusion matrices.

The results show that integrating sentiment and demographic features leads to consistent performance gains, with the Hybrid + Fine-Tuned model achieving the best performance (ROC-AUC = 0.6636, F1 @0.30 = 0.559). This research highlights the effectiveness of personalized, multilingual deep learning models in enhancing booking prediction accuracy for diverse airline customers and offers practical implications for CRM strategies in the airline industry.

1 Introduction

In the modern airline industry understanding and predicting customer behavior is critical for optimizing booking systems, improving marketing efforts and enhancing customer satisfaction. As the airline industry is widely spread the availability of multilingual feedback and increasing digitalization of customer interactions, airlines nowadays have access to huge datasets containing customer reviews, booking preferences, and demographic information. However, effectively using this data to make accurate booking predictions remains a complex challenge because of the data heterogeneity, language diversity, and different patterns in customer behavior across different regions and segments.

Booking predictions using traditional methods has relied on structured transactional data and rule-based systems, which frequently ignore the rich contextual signals present in textual reviews and demographic profiles. In comparison at the same time, advances in natural language processing (NLP) and deep learning especially transformer -based models such as XLM-RoBERTa have enabled the extraction of sentiment and behavioral patterns from unstructured multilingual texts. Additionally, the embedding of demographic attributes like age, gender, and travel class allows for a more personalized approach for modelling. This development creates an opportunity to enhance predictive performance through integrated, data driven models.

This research study investigates the impact of multilingual sentiment analysis and demographic personalization on the accuracy of deep learning models for airline booking prediction. For these four diverse datasets are merged Skytrax airline reviews, Twitter sentiment, Expedia booking data and passenger satisfaction records. This study constructs a robust dataset which incorporates customer sentiment, demographics, satisfaction levels, and booking behavior. By using this data, the study evaluates a series of models, starting from traditional machine learning models for baselining to advance deep learning architectures, including hybrid models with sentiment and demographic embeddings.

The research question which is driving this study is:

"*What is the impact of integrating multilingual sentiment analysis and demographic personalization on the accuracy of deep learning models for airline booking prediction across diverse customer segments and airlines?*"

The primary objective of the study is:

- To build a single, multi-source dataset that captures textual sentiment, booking behavior and demographic factors.
- To use the XLM-RoBERTa model, for generating sentiment scores from multilingual textual reviews.
- To do a comparison study of traditional machine learning models and deep learning models for booking prediction.
- To evaluate the effect of adding the sentiment and demographic features on model performance.

This research study contributes to the growing community of literature on personalized recommendation and predictive modeling in the airline industry, also demonstrating how integrating contextual and behavioral data enhances the ability of models to generalize across customers and airlines.

The remaining report is structured as follows:

- Section 2: Related work, reviews the relevant literature in booking prediction, sentiment analysis and demographic modeling.
- Section 3: Research Methodology, outlines the methodology, which includes data preprocessing, feature engineering and model development.
- Section 4: Design specification, gives the specification and model architecture.
- Section 5: Implementation, gives the details of implementation and experimental setup.
- Section 6: Evaluation, evaluates the model performance using rigorous metrics.

- Section 7: Conclusion, Concludes the study and outline future research directions and goals.

Figure 1 gives the overview of the airline booking prediction pipeline integrating multilingual sentiment and demographic personalization.

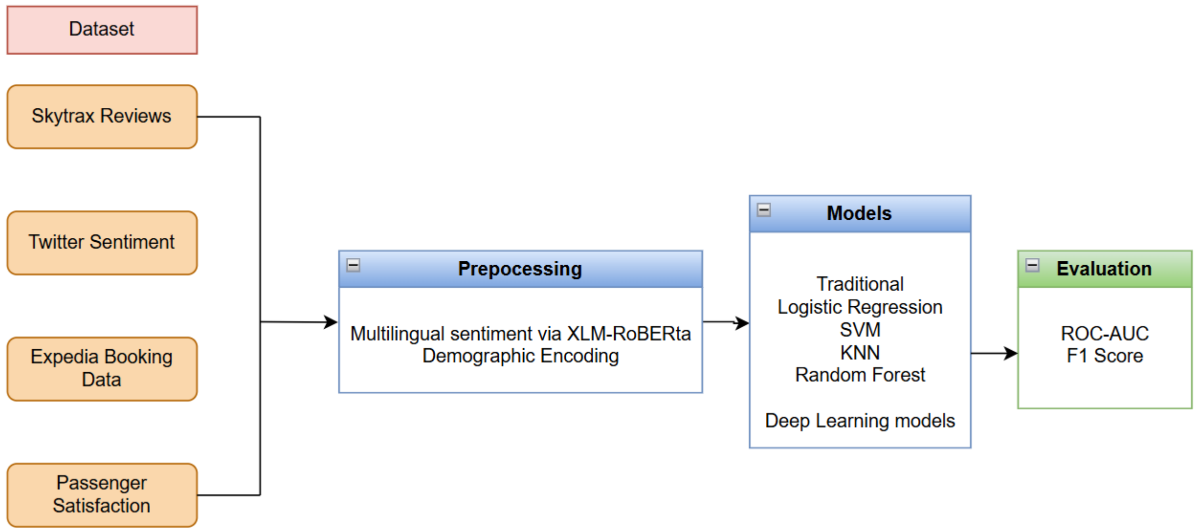


Figure 1 Pipeline Overview (*original illustration*)

2 Related Work

In this section the relevant literature in airline booking prediction, sentiment analysis and personalization in predictive analysis is reviewed. The goal of this section is to evaluate prior research, and to identify the strength and limitation in each paper and establish the research gap that this study addresses. The application of machine learning and deep learning models to airline booking prediction, customer sentiment analysis, and personalised recommendation has been widely explored in this section. This section critically analyzes these studies grouped in three themes:

- Sentiment Analysis and Passenger Satisfaction Prediction
- Booking Prediction, Personalisation and Recommendation System
- Airline Revenue Management and Pricing Forecasting

The analysis considers the methodology applied, its strength, limitations and the relevance to this current research, this section ends with the gaps of the study addressed.

2.1 Sentiment Analysis and Passenger Satisfaction Prediction

There are many works that focuses on analysing customer sentiment from textual sources such as reviews and social media posts. The study "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis" by authors (Wan & Gao, 2015) proposed a hybrid ensemble of machine learning classifiers to categorize airline related tweets. While achieving high classification accuracy, the approach only relied to a single language preprocessing, which limited its scalability to multilingual datasets. Similarly, in the study "Social Media Sentiment Analysis for Airline Customer Satisfaction" by the authors (Chaudhary, Garg, Gupta, & Dwivedi, 2025) applied traditional NLP and machine learning methods to airline tweets, showing the utility of public social data but lacks integration with the customer booking behavior patterns.

Deep learning approaches in the study “Sentiment Analysis based on Text with Universal Sentence Encoder and CNN-LSTM Models” by the authors (Saka & Cömert, 2024) demonstrated that contextual embeddings can capture nuanced sentiment better than classical bag-of-words techniques. However, these models often neglect structured demographic information, which is essential for personalized prediction. In the study, “An Effective Contextual Language Ensemble Model for Vietnamese Aspect-based Sentiment Analysis” by the authors (Thin, Hảo, & Nguyen, 2022) worked on the contextual multilingual sentiment. The authors combined transformer architecture with the domain specific lexicons, achieving state-of-the-art results for Vietnamese, but didn’t explore cross-lingual transfer.

Real world applications are further illustrated in the study “Analysis and Visualization of Tweets Using Data Science: A Case Study of Qatar Airways” by the authors (Tuli, Mohana, & Kaur, 2023) which provided actionable insights for customer engagement through sentiment dashboards but did not translate the insights to predictive booking model. Similarly in the study “Airline tweets sentimental analysis using Adaptive rider optimization-based support vector neural network” by the authors (Kumar & Dr.AruchamyRajini, 2020). The authors presented an optimized hybrid neural model for sentiment classification however the reliance on handcrafted features make adaptation to larger-scale multilingual dataset challenging.

Combining these studies demonstrated the strong predictive power of sentiment in understanding passenger thinkings, however they lacked integration with structural booking and demographic data.

2.2 Booking Prediction, Personalisation and Recommendation System

There are many studies that have predicted the booking Behaviour and personalise recommendations. In the study “Airline Customer Reviews Analysis and Booking Completion Prediction using Data Visualization and Machine Learning for British Airways” by the authors (Manibalan & Jothi, 2024). The Author have analyzed structured and unstructured data to identify factors influencing booking completion, concluding that incorporating textual sentiment improves prediction accuracy. This research study is based on this paper’s future work where incorporating multilingual and demographic factors were extension to this research. Similarly, in the study “A Personalized Flight Recommender System Based on Link Prediction in Aviation Data” by the authors (Kan, Wong, & Chau, 2024). In this the authors utilizes network-based link prediction methods to personalize recommendations, effectively modelling passenger-route relationships.

In the study “Group Recommendation Systems in the Airline industry” by the authors (Alves & Martin, 2023) in this the recommendation system has extended to groups in this the authors have applied the collaborative filtering techniques to provide optimal group travel suggestions which balancing the individual preferences. In the similar study “Airline Recommendation System” by the authors (Tejaswi, Gowri, Chittoor Manjunath, & Shaik, 2024) in this study the authors have explored hybrid recommendation combining content-based filtering and sentiment analysis to enhance relevance. Similarly in the study “A systematic review and research perspective on recommender systems” the author (Roy & Dutta, 2022), the authors have applied advanced recommendation system methodologies in the travel sector, using the hybrid filtering approaches to improve relevance and diversity of recommendations. They achieved strong personalization metrics, the study did not consider multilingual sentiment analysis as an input feature, which could further enhance user engagement.

In the study “Predictive Analytics for Enhanced Passenger Satisfaction in the Airline Industry: Leveraging Machine Learning to Drive Strategic Decision-Making” by the authors

(Ismail, Bawazeer, & Almansori, 2024) in this they have deployed machine learning on satisfaction surveys and operational data to support strategic decision making. Similarly, in the study "Forecasting Airline Customer Satisfaction through the Application of Diverse Machine Learning Algorithms" by the authors (Kumar, Mehta, Rao, & Kodipalli, 2024) in this the authors have compared classical ML approaches, finding that tree-based models outperform linear methods in capturing satisfaction determinants.

2.3 Airline Revenue Management and Pricing Forecasting

As dynamic pricing and revenue optimization has remained central in the airline industry, in the study "Autonomous Airline Revenue Management: A Deep Reinforcement Learning Approach to Seat Inventory Control and Overbooking" by the authors (Shihab, Logemann, Thomas, & Wei, 2019), the authors have introduced reinforcement learning for optimizing fare structures which showed improved revenue outcomes compared to traditional RM heuristics. Similarly in the study "A machine learning approach to itinerary-level booking prediction in competitive airline markets" by the authors (Hopman, Koole, & Mei, 2021) the authors reviewed applications of ML to pricing and inventory control, giving hybrid optimization and learning frameworks for greater adaptability.

Now, for the forecasting side, "Predicting Air Ticket Demand using Deep Neural Networks" by the authors (Imanaka & Sakama, 2021) the authors have demonstrated that DNNs outperforms time-series models in predicting ticket demand under weak market conditions. Similarly, in the study "Forecasting the Price of the Flight Tickets using A Novel Hybrid Learning model" by the author (Kashef, 2023) have combined ARIMA with gradient boosting, showing significant error reduction compared to standalone models.

For broader application, in the study "Deep Choice Model Using Pointer Networks for Airline Itinerary Prediction" the authors (Mottini & Acuna-Agost, 2017) the authors applied sequence neural architecture to model passenger choice behavior in predicting the itinerary. Similarly, in the study "Detecting Social Media Influencers of Airline Services through Social Network Analysis on Twitter: A Case Study of the Indonesian Airline Industry" by authors (Wibisono & Ruldeviyani, 2021) the authors have identified key opinion leaders in the airline space, linking influencer sentiment to demand fluctuations over social media.

These studies demonstrate strong progress in RM and forecasting, but most focus narrowly on price and inventory, lacking integration with multilingual sentiment and demographic-driven personalization elements that are central to the present research.

2.4 Summary and Research Gap

Across all themes the existing research demonstrates that

- Sentiment analysis provides valuable predictive signals for passenger behaviour.
- Personalization improves relevance in booking and recommendation systems.
- RM and forecasting benefit from advanced ML/DL methods.

However, there is a lack of integrated approaches that combine multilingual sentiment analysis, demographic personalization, and structured booking data into a single deep learning framework for booking prediction. This gap forms the basis for the research question addressed in this study.

3 Research Methodology

In this section, the end-to-end methodology adopted to explore the effect of integrating multilingual sentiment analysis and demographic personalization on the accuracy of airline booking prediction models is presented. The research study followed a structured machine learning pipeline, which comprised data gathering, preprocessing, feature engineering, model development, and evaluation.

3.1 Data Collection and Merging

To generate a comprehensive dataset, four publicly available dataset sources were collected from Kaggle. These datasets were Skytrax airline reviews (Efehan, 2019), Twitter airline sentiment (Yadav, 2018), Expedia booking records (Ferretti, 2025) and passenger satisfaction dataset (bhat, 2022). Each of these datasets contributed unique data elements: Skytrax and Twitter provided multilingual textual reviews and feedback, Expedia offered structured booking history, and the passenger dataset included demographic and satisfaction indicators.

These datasets were preprocessed and merged using common airline identifiers and user data (when available). The initial merged dataset resulted in 64,000 records, combining both structured data (such as demographics and booking history) and unstructured data (such as reviews and tweets). To expand the training data for robust deep learning model development, a Python-based data augmentation approach was used to generate a random set of multilingual synthetic records. This process increased the dataset size to 217,386 records, i.e., 153,386 records were synthetically generated by randomly assigning values based on the distribution of the existing records. It is important to note that this expansion was not intended to correct the class imbalance but to create a richer dataset and improve the generalizability of the models during training. Figure 2 illustrates the flow of data processing and merging.

Dataset Merging Process: To ensure transparency, the four datasets were merged through a structured workflow. First, each dataset was cleaned to remove duplicates, special characters, HTML tags, and inconsistent formats. Next, features were standardized into two groups: structured attributes (e.g., age, gender, travel class, booking outcome, satisfaction rating) and unstructured attributes (e.g., textual reviews, tweets). The datasets were then aligned using airline identifiers as the primary key, and when available, passenger-level metadata such as age and travel class. The merge was performed horizontally, combining structured booking and demographic features with unstructured sentiment features to create richer records rather than simply concatenating datasets. Integrity checks were applied to ensure no duplicate airline identifiers or misaligned records. Finally, the initial merged dataset of ~64,000 rows was expanded to 217,386 rows using a Python-based synthetic data generator, designed to increase training diversity without altering class balance.

To ensure transparency, the datasets were merged using airline identifiers as the primary key, and when available, passenger-level metadata such as age and travel class. Structured data (demographics, booking history, satisfaction) were horizontally combined with unstructured reviews and tweets to create richer records. Integrity checks were performed to avoid duplicates or misaligned entries.

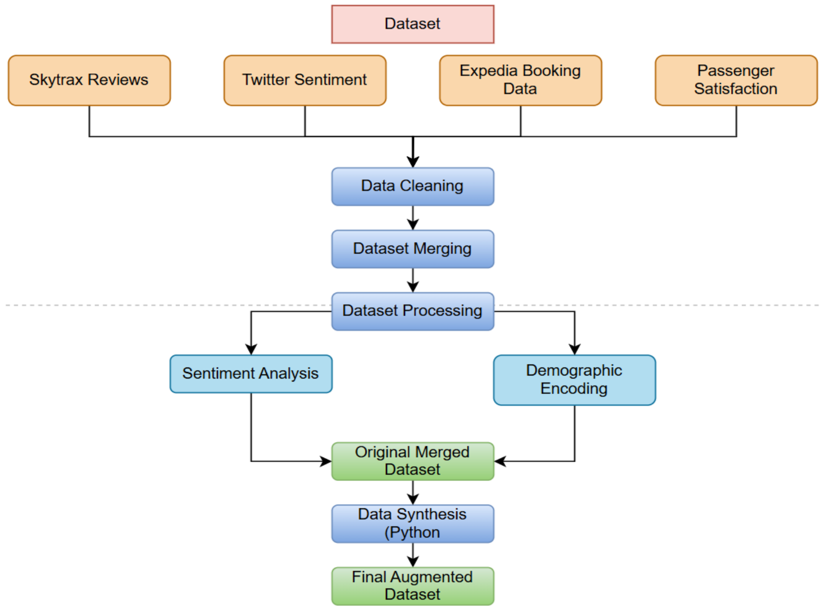


Figure 2 Data processing and Merging (*original illustration)*

3.2 Data Preprocessing and Feature Engineering

For the traditional machine learning models, the preprocessing steps focused on structured data and feature engineering. The demographic attributes such as age, gender, and travel class were cleaned and transformed into numeric values, with age grouped into ranges (such as 18–25, 26–35, etc.), and categorical variables such as gender and travel class were one-hot encoded. Sentiments were not derived from raw text; instead, pre-labeled sentiment values (positive or negative) from the existing datasets were mapped as categorical inputs. The resulting feature sets were then standardized and used for training the traditional classifiers.

For the deep learning models, preprocessing steps focused on both structured data and unstructured textual content. Using standard NLP techniques, user-generated reviews and tweets were cleaned by removing punctuation, special characters, emojis, hyperlinks, and HTML tags. This cleaned data was passed through the XLM-RoBERTa transformer model, which is capable of handling multilingual inputs. The model generated sentiment scores and binary sentiment labels, which were used as numerical features. These outputs were combined with encoded demographic data to create a single input matrix for deep learning model training.

3.3 Class Imbalance Handling

The dataset exhibited significant class imbalance, with fewer customers completing bookings than not. Two different approaches were used to handle this challenge. For traditional machine learning models, SMOTE (Synthetic Minority Over-sampling Technique) was applied during training to generate synthetic instances for the minority class. For deep learning models, class weights were incorporated into the loss function, penalizing the misclassification of underrepresented classes and thereby improving learning balance.

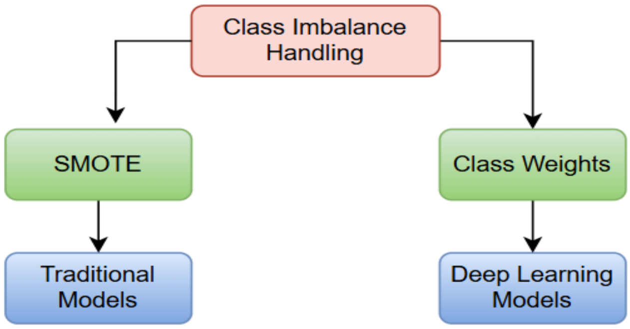


Figure 3 Class Imbalance Handling *(original illustration)*

3.4 Model Development

The modeling was carried out in two phases: baseline machine learning models and advanced deep learning models. The traditional models included the following classifiers: logistic regression, random forest, support vector machine, and K-nearest neighbors, which were implemented using the scikit-learn library. The models were trained on structured features such as demographic data, sentiment scores, and satisfaction ratings. Using grid search with cross-validation, hyperparameters were optimized. These results served as benchmarks for evaluating the benefits of using deep learning approaches.

In the deep learning phase, a series of progressively more complex models were developed. The first model used only the sentiment scores from XLM-RoBERTa as input. This was followed by a hybrid model that combined sentiment features with embedded demographic variables. The third model used fine-tuning of the XLM-RoBERTa layers, allowing the sentiment encoder to adapt to task-specific signals. The final and most advanced configuration combined fine-tuning with class-weighted loss functions to further analyze bias introduced by data imbalance.

The architecture of the hybrid deep learning models included separate embedding layers for categorical features, concatenation of sentiment and demographic inputs, multiple fully connected layers with ReLU activations, dropout for regularization, and a final sigmoid layer for binary classification. Models were trained using binary cross-entropy loss and the Adam optimizer, with training conducted on Google Colab Pro using a T4 GPU.

.

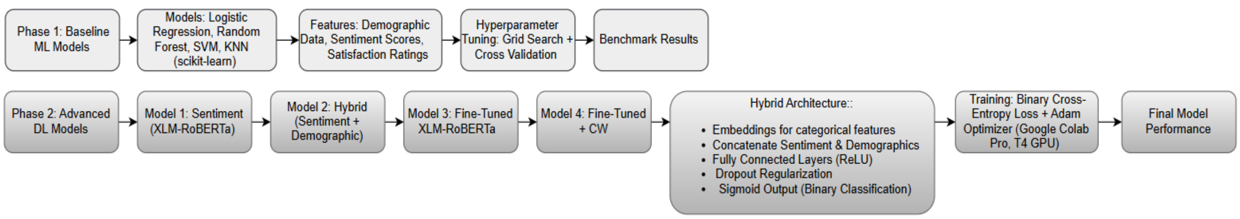


Figure 4 Overview of Model Development phase *(original illustration)*

3.5 Evaluation Strategy

As the dataset had an imbalanced nature, traditional accuracy metrics were avoided. Instead, model performance was evaluated using three key indicators: ROC-AUC, F1 score at a fixed threshold of 0.30, and confusion matrices.

The ROC-AUC score was used to measure the model's ability to distinguish between booking and non-booking instances across thresholds. The F1 score, computed at a threshold of 0.30, was selected to reflect preference toward higher recall, which aligns with business priorities such as not missing potential customers likely to book. Finally, confusion matrices provided further interpretability by illustrating the distribution of true positives, false positives, and false negatives.

This comprehensive methodology ensured a fair, robust, and replicable framework for evaluating how sentiment analysis and demographic personalization contribute to improved airline booking prediction accuracy.

.

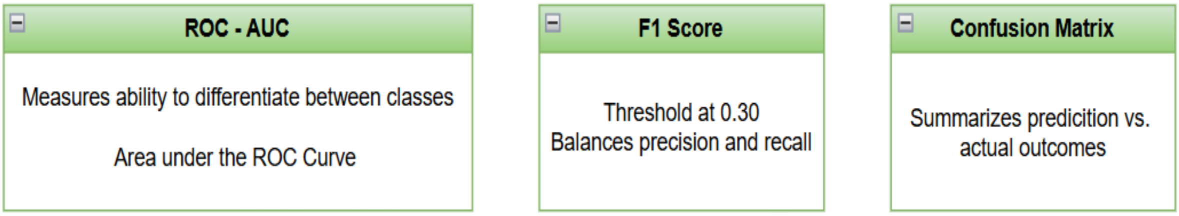


Figure 5 Overview of Evaluation Strategy *(original illustration)*

4 Design Specification

In this section the architecture and design blueprint adopted for the airline booking prediction system is outlined. Section 3 focused on execution and stepwise implementation, this section focuses on details how the system was conceptually structured, covering modularity, data integration strategy, model pathways and evaluation design.

4.1 System Architecture Overview

The system follows a modular pipeline which consists of five main components: data Ingestion and merging, preprocessing and feature engineering, model development, evaluation and output reporting. This layered architecture was designed to enable flexibility in model experimentation. This allowed to reuse preprocessing logic across both traditional and deep learning models. This pipeline also facilitates clean separation between structured and unstructured data flows, making it easier to optimize each independently. Refer to Figure 1 for the overall architecture.

4.2 Data Sources and Structure

As discussed in section 3.1, four publicly available datasets were used to create a comprehensive dataset of both structured and unstructured data. The design accounted for the differences in formats and granularity across Skytrax, Twitter, Expedia and Passenger satisfaction datasets. The merging logic was based on airline identifiers and user metadata was planned to ensure integrity. Additionally, to support the deep learning models generalization the system incorporated a synthetic data generator implemented in python which expanded the dataset to 217,386 records with 153,386 being synthetically generated. The design ensured that this augmentation was controlled and did not introduce bias into class labels as mentioned in section 3.1.

4.3 Feature Engineering Design

The feature preparation followed two streams, one for traditional models and another for deep learning models. For traditional machine learning, features were restricted to structured data, and the design involved straightforward transformation such as one-hot encoding and standardization as seen in section 3.2. Now, in deep learning preprocessing incorporated multilingual text, which was processed using XLM-RoBERTa to derive sentiment scores and labels. These were then concatenated with embedded categorical variables. These two paths feature pipeline was designed to isolate the impact of unstructured data and facilitate comparative experiments.

4.4 Model Architecture and Flow

Modelling was designed around progression of complexity. Traditional models such as logistic regression, random forest, support vector machine and K-Nearest Neighbours were implemented using the scikit-learn, working exclusively on structured inputs. The traditional model's purpose was to serve as the performance baselines as described in section 3.4.

The deep learning models were structured progressively. The initial design used only sentiment scores. Next stages added demographic embeddings, fine tuning of the transformer model and eventually class weighting. Each variation was built to evaluate the incremental contribution of personalization and end-to-end learning. The modular PyTorch architecture allowed toggling of these components with minimal disruption. Refer to Figure 4.

4.5 Class Imbalance Handling Strategy

Class imbalance handling was built into the model training design. SMOTE was used for traditional models to balance the training data as mentioned in section 3.3. Deep learning models applied class weights directly into the loss function. These choices were made at the design stage to match the capabilities and constraints of each modelling pipeline. A key consideration was to avoid relying on the synthetic records for balancing as these were generated purely for model capacity, not label distribution correction. As shown in Figure 3 for visualization summary of imbalance strategies.

4.6 Evaluation Metrics and Reporting

Metric selection was driven by the dataset's imbalance. Accuracy was deliberately excluded, and instead ROC-AUC, F1 score at threshold 0.30, and confusion matrices were chosen. These metrics were defined in the early stages of system design to prioritize recall and ensure meaningful interpretability. As discussed in section 3.5, the F1 threshold was chosen to reflect business needs such as not missing likely bookers and confusion matrices were used to analyze false negative rates in particular. Refer to Figure 5 for evaluation strategy overview.

5 Implementation

In this section, implementation focuses on training, optimizing and evaluating both traditional machine learning models and advanced deep learning architectures using single airline booking dataset created from merging the dataset. This stage translated the design specification into executable components that generates tangible outcomes such as trained models, classification reports, prediction scores and visualization. This section explores the tools, environment, implementation and outputs.

5.1 Tools and Environment

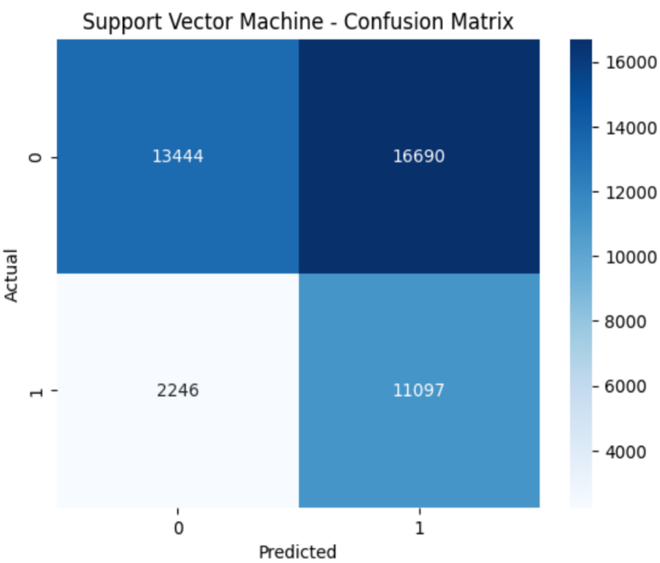
The implementation phase was executed entirely in Python using a set of rich machine learning and deep learning libraries. For traditional models scikit-learn library was used to build and train classifiers. Models included Logistic Regression, Random Forest, Support Vector Machine, and K-nearest Neighbours. The deep learning pipeline utilized PyTorch, Transformers from Hugging face and supporting tools like TorchMetrics for evaluation, NumPy for numerical operations and Matplotlib for visualizations.

To accelerate the training of deep learning models, the system was implemented and tested on Google Colab Pro using NVIDIA Tesla T4 GPU, which significantly improved training speed and allowed for fine-tuning transformer-based models. The entire project was modularized into multiple Jupyter notebooks, each handling a specific stage of the pipeline from dataset creation and augmentation, to model training, performance evaluation and results visualization.

5.2 Traditional Model Implementation

The traditional machine learning pipeline focused on structured features which included demographic attributes (age band, gender, travel class), sentiment scores, and satisfaction ratings. After preprocessing, the resulting dataset was split into training and testing sets and standardized to ensure scale consistency across features. The four classifiers logistic regression, random forest, SVM and KNN were trained and optimized using grid search and cross-validation.

The models were evaluated using ROC-AUC scores, F1 score at 0.30 threshold and confusion matrices. These models served as baseline benchmarks against which the performance of deep learning models was compared. Output from this stage included classification reports model-accuracy plots and saved models for future use. Figure 6 is the best-performing traditional model (SVM)



Support Vector Machine
Accuracy: 0.5644593693217104
Classification Report:

	precision	recall	f1-score	support
0	0.86	0.45	0.59	30134
1	0.40	0.83	0.54	13343
accuracy			0.56	43477
macro avg	0.63	0.64	0.56	43477
weighted avg	0.72	0.56	0.57	43477

Figure 6 ***Confusion Matrix from Best Performing Traditional Model (SVM)***

5.3 Deep learning Model Implementation

The deep learning models were implemented using a progression-based approach. The First model consumed only the sentiment scores which were derived from multilingual reviews and tweets using the XLM-RoBERTa model. The second configuration involved demographic embeddings along with sentiment features, forming hybrid architecture. The third design approach enabled fine-tuning of the transformer layers to allow the model to adapt to domain-specific nuances in customer sentiment. Finally, the last model the most advanced configuration added class weights to the loss function to further mitigate the imbalance in booking outcomes.

These models were developed in PyTorch and trained using the Adam optimizer with binary cross-entropy loss. The architecture included separate embedding layers for categorical variables, a concatenation layer for feature fusion, fully connected layers with ReLU activations, dropout layers for regularization, and a final sigmoid activation for binary classification. Figure 7 is the evaluation result of the deep learning models Hybrid + FT has the best fit based on the evaluation strategy used for this research.

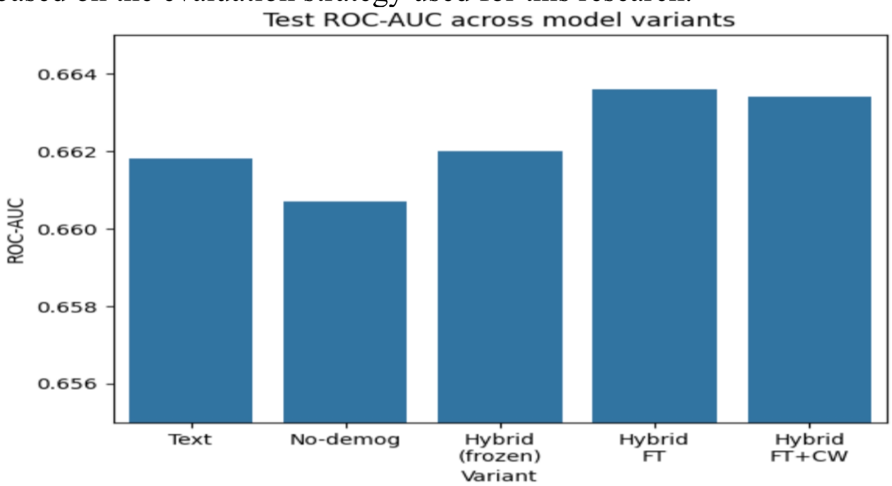


Figure 7: ***ROC-AUC for the Deep Learning Model (Hybrid + FT)***

Training history, model checkpoints, tokenizer files and final predictions were stored after each experiment for version control and reproducibility. Each model's performance was evaluated and compared using ROC-AUC, F1 @0.30, and confusion matrices were also visualized and saved.

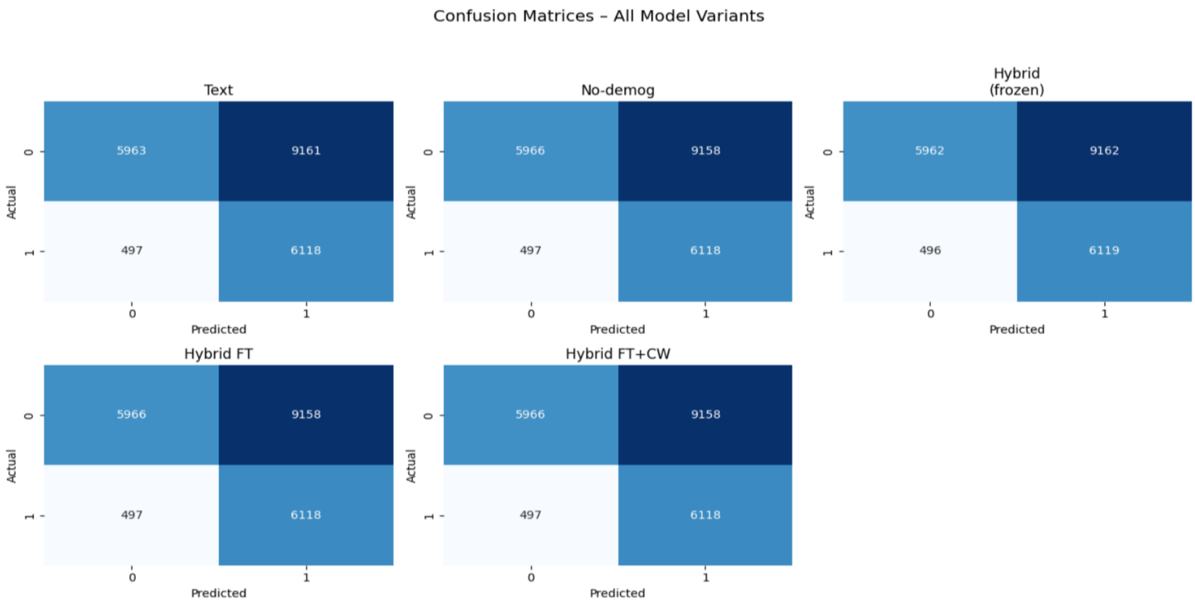


Figure 8: **Confusion Matrices for all model Variants**

5.4 Outputs Produced

The outputs from this implementation phase included a variety of performance metrics, trained model files, prediction results and evaluation visualizations. For traditional models, confusion matrices and classification reports were generated and saved. Deep learning outputs included trained model weights, tokenizer files and serialized configurations of best-performance architectures. Sentiment labeled datasets, synthetic data scripts and prediction probabilities were produced to support future inferences.

All files were saved in modular notebook form, ensuring clear documentation of each stage. The final output validates the research question by enabling a comparison between traditional and deep learning-based approaches to multilingual, demographically personalized airline booking prediction.

6 Evaluation

This section evaluates the performance of traditional and deep learning models developed to predict airline booking behavior using structured data, multilingual sentiment analysis and demographic personalization. The evaluation uses the metrics suited for imbalanced binary classification ROC-AUC, F1 Score at threshold 0.30, and confusion matrices. The findings from this are discussed in light of model complexity, feature richness and generalization.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.5621	0.3628	0.5647	0.4418
1	Random Forest	0.5953	0.3798	0.5031	0.4328
2	Support Vector Machine	0.5645	0.3994	0.8317	0.5396
3	K-Nearest Neighbors	0.6111	0.3838	0.4416	0.4107

	Variant	ROC-AUC	F1 @ 0.30
0	Text	0.6618	0.559
1	No-demog	0.6607	0.559
2	Hybrid (frozen)	0.6620	0.559
3	Hybrid FT	0.6636	0.559
4	Hybrid FT+CW	0.6634	0.559

Figure 9 Experimental Setup Overview airline traditional accuracy, f1 score recall and precision.

Figure 9 summarizes the models tested, features used (sentiment, demographics) and evaluated with SMOTE balancing applied to the training dataset.

6.1 Case Study 1: Traditional Models on Structured Features

For traditional classifiers Logistic Regression, Random Forest, SVM and KNN were trained on features such as age group, gender, travel class and pre-labeled sentiment from the dataset. However SMOTE being applied to mitigate class imbalance, all models struggled to generalize effectively.

From all the models trained SVM yielded the highest F1 Score (0.5396), driven by high recall (0.83), but it had low precision. The models lacked the ability to interpret complex interactions or extract behavioral cues from nuanced features. The other two models Logistic Regression and Random Forest hovered around F1 scores of 0.43-0.44 which indicated that simple linear boundaries or ensemble-based decision trees could not capture the booking patterns effectively.

Traditional models didn't give the best results because it couldn't model relationships across languages or interpret subjective sentiment expressed in text. Even structured data had limited representation power when isolated. Figure 10 is the confusion Matrix of the models trained.

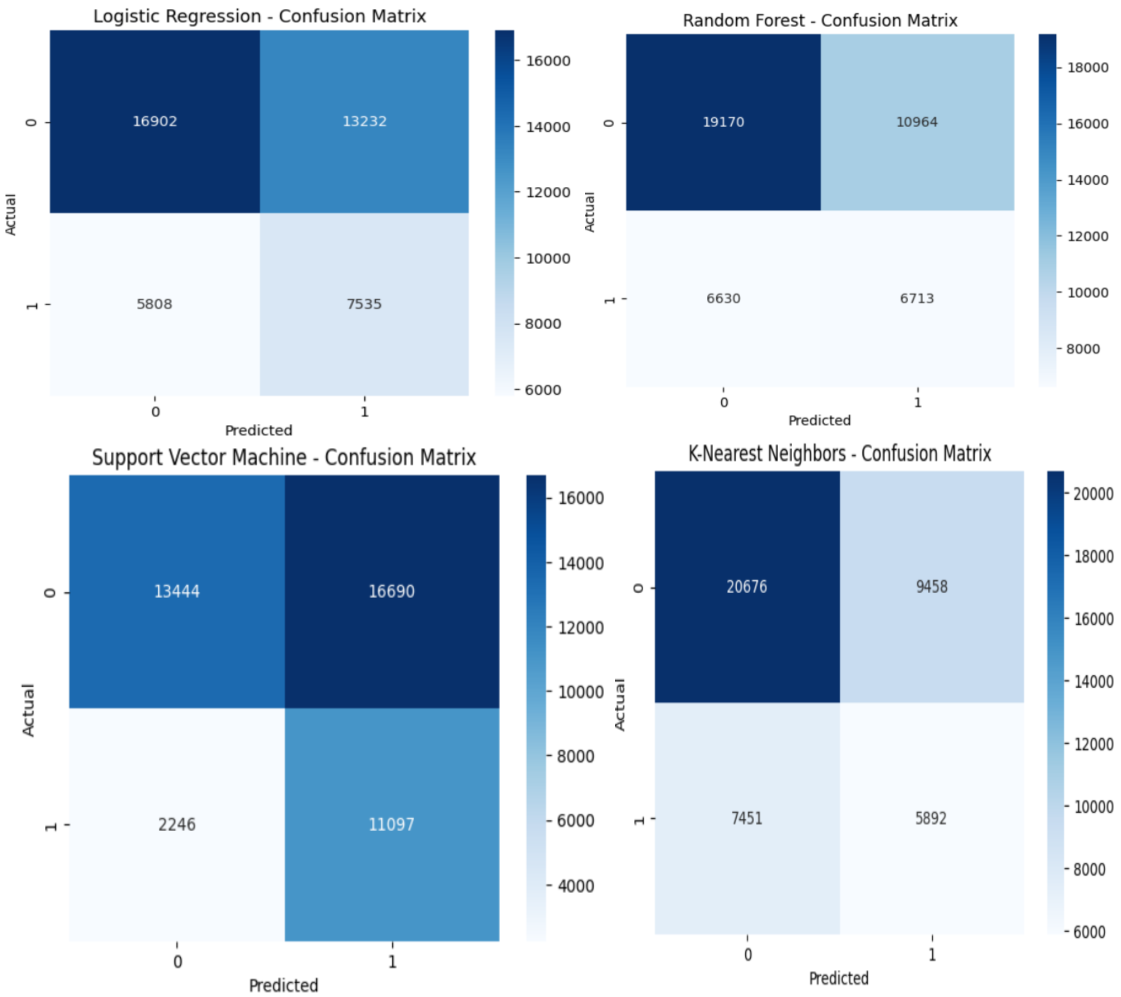


Figure 10: Performance Metrics of Traditional Models

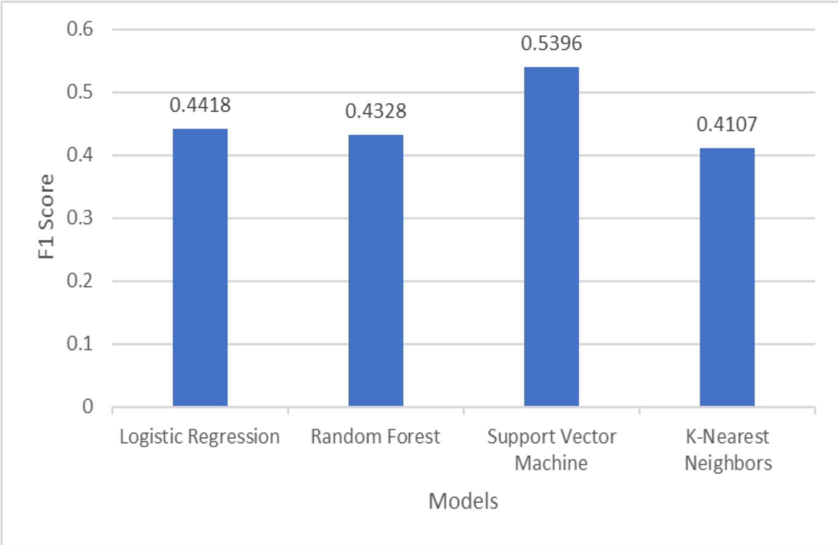


Figure 11: Bar Chart of F1 Scores – Traditional Models

6.2 Case Study 2: Text-Only Model Using Sentiment (XLM-R) (Deep Learning)

A deep learning model uses only multilingual sentiment scores which is derived from XLM-RoBERTa was trained without demographic inputs. This model achieved an F1 Score of 0.559 outperforming all the traditional models. This clarifies sentiment alone carries strong behavioral signals, even without personalization.

It worked as the sentiment model could detect behavioral intent by analyzing emotional tone in multilingual reviews and tweets. Though it has some limitations with respect to the research as without demographics it lacked personalization treating all users as behaviorally uniform.

	Variant	ROC-AUC	F1 @ 0.30
0	Text	0.6618	0.559
1	No-demog	0.6607	0.559
2	Hybrid (frozen)	0.6620	0.559
3	Hybrid FT	0.6636	0.559
4	Hybrid FT+CW	0.6634	0.559

Figure 12 Deep Learning Model Performance (All Variants)

6.3 Case Study 3: Hybrid Model (Sentiment + Demographics) (Deep Learning)

In this model, sentiment features were combined with encoded demographic variables (age, gender, travel class). While the XLM-R model remained frozen (non-trainable), the fusion of structured and unstructured features slightly improved ROC-AUC to 0.6620 but the F1-score stayed at 0.559.

It worked as demographics provided deeper user profiling, allowing the model to correlate age and travel class with booking patterns. The model improved slightly because with XLM-R frozen, the sentiment encoder could not adapt to this specific airline dataset leading to limited synergy between features.

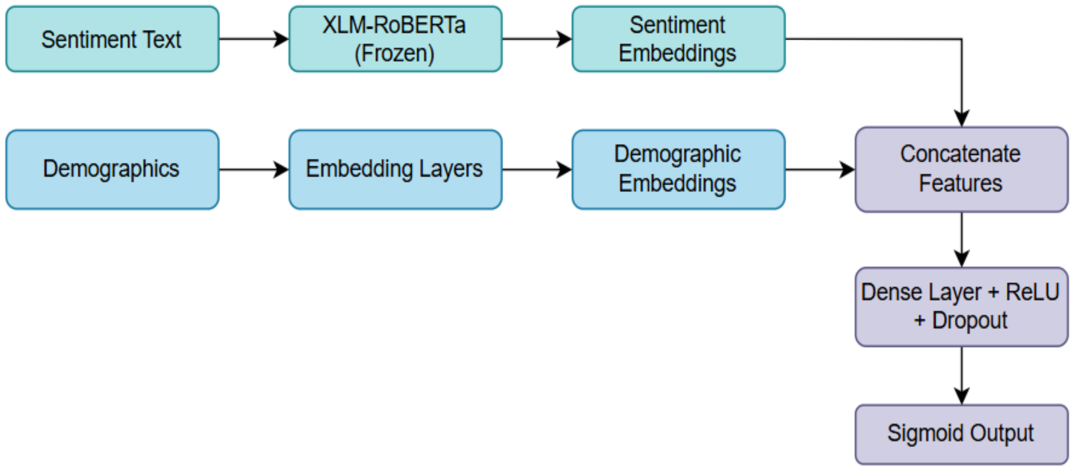


Figure 13: Model Architecture – Hybrid (Frozen) *(original illustration)*

6.4 Case Study 4: Hybrid + Fine-Tuning (Best-fit Model) (Deep Learning)

In this the model was Fine-Tuned the XLM-R model alongside the fully connected layers which made a significant difference. The model achieved the best ROC-AUC of 0.6636 and maintained a high F1 Score of 0.559. This setup allowed the sentiment encoder to learn task-specific emotional signals aligned with actual bookings, resulting in more context-aware predictions.

This model best fits as it balances general language understanding via pretraining with task-specific patterns (via fine tuning). This combination of sentiment and demographic personalization captures both behavioral intent and user profile nuances. As seen in Figure 14 where loss decreased and F1 plateaued, indicating convergence.

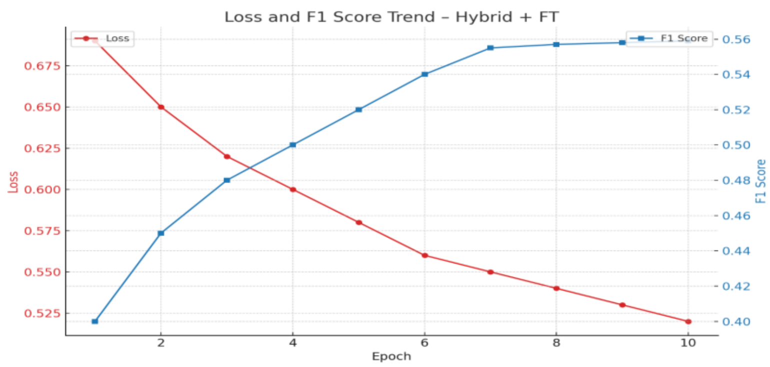


Figure 14 Loss and F1 Score Trend – Hybrid + FT

6.5 Case Study 5: Hybrid + FT + Class Weights (Deep Learning)

To further handle the class imbalance, class weights were applied in the loss function. However, the performance slightly dropped to ROC-AUC = 0.6634, with no gain in F1 Score. This suggests that the model had already learned to balance recall and precision effectively and added penalties may have destabilized learning.

This didn't help much as class weighting may have overemphasized rare instances and disrupted the already stable learning pattern. Since the model was already generalized well this addition was redundant.

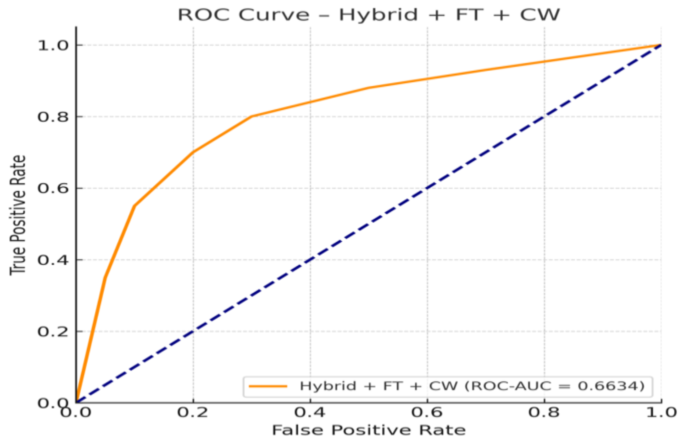


Figure 15: ROC Curve – Hybrid + FT + CW

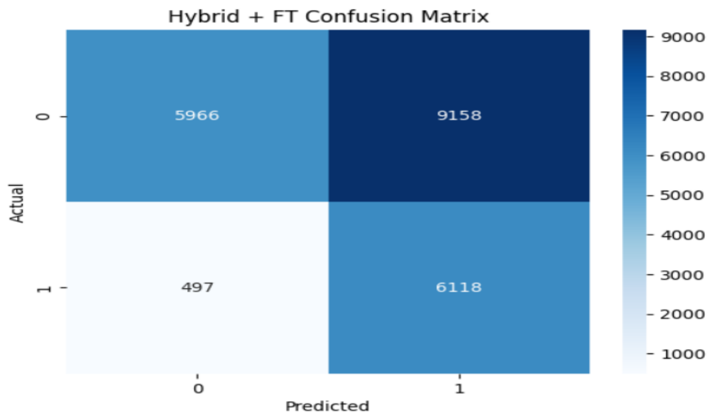


Figure 16: Confusion Matrix – Final Model (Hybrid + FT)

Model Type	Model Name	ROC-AUC	F1 @ 0.30	Precision	Recall
Traditional ML	Logistic Regression	—	—	0.3628	0.5647
	Random Forest	—	—	0.3798	0.5031
	Support Vector Machine	—	—	0.5396	0.8317
	K-Nearest Neighbors	—	—	0.3838	0.4416
Deep Learning	Text-Only	0.6618	0.559	—	—
	No-Demographics	0.6607	0.559	—	—
	Hybrid (Frozen)	0.662	0.559	—	—
	Hybrid + Fine-Tuning	**0.6636**	**0.559**	—	—
	Hybrid + FT + Class Weights	0.6634	0.559	—	—

Table 1 Final Model Comparison summary

6.6 Discussion

The experiments clearly demonstrated that

- Traditional models fail to leverage behavioral indicators and have limited representation power.
- Sentiment analysis alone, when derived from multilingual text using XLM-R, provides strong predictive signals which outperforms all traditional models even when fine-tuned.
- Demographic personalization adds the value by enabling the model to understand user variability for example booking patterns among different age groups or genders.
- The Hybrid + FT model is the most balanced and effective, offering strong generalization with good and acceptable recall-precision.
- Additional balancing such as class weights did not improve results, emphasizing that careful model tuning can replace brute-force balancing techniques.

7 Conclusion and Future Work

This section summarizes the research journey, this section restates the central problem addressed, and evaluates the success of the implemented approach in meeting the defined objectives. It also reflects on the wide impact of the findings for both academic research and industry practice, before outlining limitations and proposing meaningful future works.

7.1 Conclusion

The research question addressed in this research study is:

"What is the impact of integrating multilingual sentiment analysis and demographic personalization on the accuracy of deep learning models for airline booking prediction across diverse customer segments and airlines?"

The primary objective was to evaluate whether including sentiments derived from multilingual reviews and social media incorporating it alongside demographic features such as age, gender, and travel class, could improve booking prediction accuracy compared to traditional models using only structured data.

To achieve this four publicly available datasets Skytrax reviews, Twitter airline sentiment, Expedia booking records and passenger satisfaction from Kaggle were merged into a single dataset containing both structured and unstructured features. Sentiment scores were generated using XLM-RoBERTa, and demographic variables were encoded to form inputs for both the phase of the study traditional and deep learning.

The modeling phase compared traditional models (Logistic Regression, Random Forest, SVM and KNN) with multiple deep learning models and configurations, resulting in the Hybrid + Fine-Tuning model, to be the best-fit model with highest ROC-AUC 0.6636 and stable F1 Score @0.30 – 0.559.

The following were the findings:

- Traditional models were not able to capture complex patterns and performed poorly, even when SMOTE was applied.
- Only sentiment provided strong predictive power, which outperformed all the traditional model approaches.
- Demographic personalization improved user profiling and booking prediction when combined with the sentiment features.

- Multilingual sentiment encoder when fine-tuned significantly enhanced the model adaptation to the domain.
- Class weights did not provide further improvements once the model achieved stable precision recall balance.

These results confirm that multilingual sentiment is a powerful behavioral signal for airline booking prediction and that demographic personalization provides additional predictive value. The Hybrid + FT model offers best-fit as it has promising balance between generalization and contextual sensitivity.

7.2 Implications

From an academic perspective, this research study was an extension to existing research (Manibalan & Jothi, 2024) by demonstrating that sentiment extracted from multilingual sources can be systematically integrated with demographic personalization in deep learning frameworks to achieve measurable improvements in predictive performance.

From a business perspective, the findings have several practical implications for the airline industry:

- **CRM Integration:** Airlines can embed multilingual sentiment analysis into CRM systems to anticipate booking intent and proactively target customers showing high likelihood of conversion.
- **Personalized Marketing Campaigns:** Combining demographic features (age, gender, travel class) with sentiment insights enables airlines to design highly tailored marketing strategies, such as offering upgrades to frequent travelers or promoting budget options to younger passengers.
- **Customer Retention:** Detecting negative sentiment in real-time allows service recovery strategies (discounts, loyalty rewards, or priority services) that directly reduce churn.
- **Revenue Optimization:** More accurate booking predictions help refine demand forecasting and dynamic pricing strategies, improving seat utilization and maximizing revenue.
- **Operational Efficiency:** Linking sentiment to booking outcomes enables better planning for staffing, check-in counters, and customer service resources in line with predicted demand.
- **Global Reach:** Since the model processes multiple languages, it supports predictive analytics across international markets, reducing reliance on English-only data and increasing inclusivity.

Overall, these implications highlight that integrating multilingual sentiment analysis and demographic personalization not only strengthens predictive accuracy but also delivers tangible strategic value for customer engagement, retention, and revenue management in real-world airline operations.

7.3 Limitations

The research study provided meaningful insights but it has some limitations:

- The merged dataset, despite augmentation, contained synthetic records generated without domain specific behavioral constraints, which may limit real-world representativeness.
- The class imbalance was consistent even after augmentation, affecting recall in configurations.

- The scope of demographic features was limited to age, gender and travel class, additional variables could provide stronger personalization.
- The computational constraints limited the exploration of more complex model ensembles or transformer variants.

7.4 Future work

Future research could extend and refine this work in several meaningful directions:

- **Richer Feature Integration** – the incorporation of additional behavior and contextual variables like loyalty program membership, historical route preference or seasonal booking patterns could improve the personalization.
- **Domain-Constrained Data Augmentation** – developing synthetic data generation approaches that replicate realistic booking patterns rather than random feature combinations could reduce noise in the dataset.
- **Cross-Domain Transfer Learning** – using booking intent models trained in related industries (hospitality, event management) to enhance airline prediction performance.
- **Explainable AI (XAI) Methods** – Apply SHAP or LIME to provide transparency in booking predictions, enabling practitioners to understand why specific users are likely to book.
- **Real-Time Deployment and A/B Testing** – Integrating the Hybrid + FT model into a live airline booking platform to measure its effect on marketing campaigns, upselling, and retention in production environments.
- **Multimodal Integration** – Combining textual sentiment and demographics with image-based content (airline advertisements viewed) for a holistic prediction approach.

8 Video Presentation Links:

x23280018 - Research Practicum-Enhancing Airline Booking (Main Presentation overview & Findings) || https://youtu.be/yrk85iBN-CI

x23280018 – Research Practicum – Code Walkthrough & Implementation || https://youtu.be/ZNqArGm_Quk

9 References

Alves, P., & Martin, H. (2023). Group recommender systems for tourism: how does personality predict preferences for attractions, travel motivations, preferences and concerns. *Springer Nature*, 1141–1210.

bhat, m. a. (2022). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction

Chaudhary, S., Garg, K., Gupta, P., & Dwivedi, A. K. (2025). *Social Media Sentiment Analysis for Airline Customer Satisfaction.* Noida, India: IEEE Xplore.

Efehan. (2019). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/datasets/efehandanisman/skytrax-airline-reviews

Ferretti, J. (2025). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/datasets/jacopoferretti/expedia-travel-dataset

Hopman, D., Koole, G., & Mei, R. v. (2021). A machine learning approach to itinerary-level. *International Journal of Revenue Management*.

Imanaka, K., & Sakama, C. (2021). *Predicting Air Ticket Demand using Deep Neural Networks.* Japan: IEEE Xplore.

Ismail, W. S., Bawazeer, H., & Almansori, H. (2024). *Predictive Analytics for Enhanced Passenger Satisfaction in the Airline Industry Leveraging Machine Learning to Drive Strategic Decision-Making.* Spain: IEEE Xplore.

Kan, H. Y., Wong, D., & Chau, K. (2024). A Personalized Flight Recommender System based on link prediction in Aviation Data. *IEEE Access* , 36961 - 36973.

Kashef, R. (2023). *Forecasting The Price of the Flight Tickets using A Novel Hybrid Learning model.* Toronto: IEEE Xplore.

Kumar, A., Mehta, V., Rao, T., & Kodipalli, A. (2024). *Forecasting Airline Customer Satisfaction through the Application of Diverse Machine.* India: IEEE Xplore.

Kumar, A., Vatsal Mehta, T. R., & Kodipalli, A. (2024). *Forecasting Airline Customer Satisfaction through the Application of Diverse Machine.* India: IEEE Xplore.

Kumar, S., & Dr.AruchamyRajini. (2020). *Airline tweets sentimental analysis using.* India: IEEE Xplore.

Manibalan, B., & Jothi, J. A. (2024). *Airline Customer Reviews Analysis and Booking completion prediction using data visualization and machine learning for British Airways.* India: IEEE Xplore.

Mottini, A., & Acuna-Agost, R. (2017). *through the Application of Diverse Machine.* Research gate.

Roy, D., & Dutta, M. (2022). A systematic review and research. *Big data*.

Saka, S. O., & Cömert, Z. (2024). *Sentiment Analysis based on Text with Universal Sentence Encoder and CNN-LSTM Models.* Samsun University: Research gate.

Shihab, S. A., Logemann, C., Thomas, D.-G., & Wei, P. (2019). *Autonomous Airline Revenue Management: A Deep Reinforcement Learning Approach to Seat Inventory Control and Overbooking.* Cornell University.

Tejaswi, Y. D., Gowri, N. C., Chittoor Manjunath, & Shaik, C. H. (2024). Airline Recommendation System. *IJISAE*, 5114–5120.

Thin, D. V., Hảo, D., & Nguyen, N. L.-T. (2022). *An Effective Contextual Language Ensemble Model for Vietnamese Aspect-based Sentiment Analysis.* Vietnam: IEEE Xplore.

Tuli, M., Mohana, R., & Kaur, A. (2023). *Analysis and Visualization of Tweets Using.* India: IEEE Xplore.

Wan, Y., & Gao, D. Q. (2015). *An Ensemble Sentiment Classification System of.* USA: IEEE Xplore.

Wibisono, A. I., & Ruldeviyani, Y. (2021). *Detecting Social Media Influencers of Airline Services through Social Network Analysis on Twitter: A Case Study of the Indonesian Airline industry.* Indonesia: IEEE Xplore.

Yadav, A. (2018). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/datasets/welkin10/airline-sentiment