National College of Ireland

Configuration Manual

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		
Lecturer:	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:	**576**	**Page Count: 3**	

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):	

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

Swati Shukla
Student ID: x23280018

1 Introduction

1.1 Background

In the modern airline industry, there is a huge amount of data available which can help aviation industry in predicting customer booking behaviour as it is essential for optimizing revenue and enhancing customer experience. Additionally, it also helps in improving marketing strategies. This project is a research study on the impact of integrating multilingual sentiment analysis with demographic personalization to predict airline bookings more accurately. It uses transformer-based model XLM-RoBERTa for extracting sentiment from multilingual textual data and combines it with structured demographic attributes.

1.2 Aim of the Study

To enhance the accuracy of airline booking predictions by combining multilingual sentiment analysis with demographic personalization using deep learning models.

1.3 Research Objectives

- Building a single dataset which contains both textual sentiment and structured demographic data.
- Generating multilingual sentiment scores using the XLM-RoBERTa transformer model.
- Compare traditional machine learning models (Baseline) with deep learning architectures.
- Evaluating the effect of integrating sentiment and demographic features on predictive performance

1.4 Research Question

“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?

2. Environmental Setup

2.1 Hardware Requirements

- 16GB RAM
- 500GB Storage
- Intel Core i5 or above
- NVIDIA Tesla T4 GPU (Google Colab Pro)

2.2 Software Requirements

- Operating System: Windows 11 / Linux
- Python 3.x
- Google Colab Pro environment

2.3 Programming Tools

- IDE: Jupyter Notebook / Google Colab
- Language: Python
- GPU: NVIDIA Tesla T4
- Visualization: Matplotlib, Seaborn

3. Libraries Required

Library	Purpose
pandas	Data loading and manipulation
numpy	Numerical operations
scikit-learn	Machine learning models and evaluation
imblearn	SMOTE for class balancing
matplotlib	Data visualization
seaborn	Advanced data visualization
transformers	XLM-RoBERTa for multilingual sentiment extraction
torch	Deep learning model development
torchmetrics	Evaluation metrics for PyTorch models

4. Dataset Details

4.1 Dataset Sources

• Skytrax Airline Reviews – Multilingual textual reviews.
• Twitter Airline Sentiment – Social media feedback.
• Expedia Booking Data – Structured booking history.
• Passenger Satisfaction Dataset – Demographic and satisfaction ratings.

4.2 Merging Strategy

Datasets were merged using airline identifiers and user metadata. Synthetic data augmentation was done using python which increased the dataset to 217,386 records to enhance generalization. SMOTE was used for class balancing in traditional models, while class weights were applied in deep learning.

Each dataset was first cleaned (duplicates, HTML, special characters removed) and mapped into a common schema: structured attributes (age, gender, travel class, booking outcome, satisfaction ratings) and unstructured text (reviews, tweets). Records were aligned using airline identifiers as the primary key and, when available, passenger metadata (e.g., age, travel class). We then merged horizontally so that structured booking/demographic fields are paired with textual sentiment for the same airline context (not just concatenated rows). Integrity checks removed duplicate airline identifiers and misaligned records. The initial merged set (~64,000 rows) was then augmented to 217,386 using a Python generator; augmentation was for training diversity only, not class rebalancing. SMOTE was used only for traditional models; deep learning used class weights.

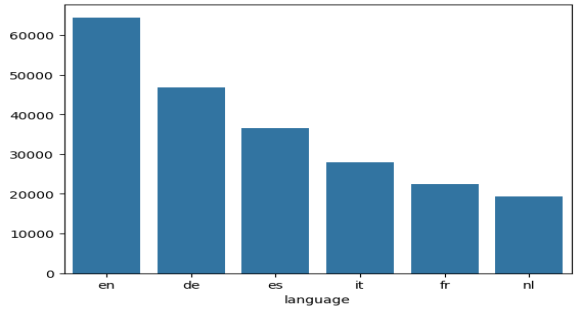

Figure 1: Dataset preprocessing and merging overview.

5. Project Flow

Phase 1: Traditional Models

- Data cleaning, feature encoding, and scaling.
- Application of SMOTE for class balancing.
- Training Logistic Regression, Random Forest, SVM, and KNN models.

Phase 2: Deep Learning Models

- Data cleaning, Sentiment extraction, and demographic encoding
- Tokenization and using XLM-RoBERTa
- Model configurations – sentiment-only, Hybrid, Hybrid + Fine Tuning, Hybrid + FT + CW (class weights)
- Training with Adam optimizer and binary cross-entropy loss
- Evaluation using ROC-AUC, F1 @0.30 and confusion matrix.

6. Repository Contains

- dataset_merge.ipynb : Merges airline datasets (refer section 4) into a single master dataset
- creating_synthetic_data.ipynb: Generates synthetic multilingual review and demographic data to augment data (not for rebalancing)
- final_airline_dataset.ipynb: prepares the final single data for modelling.
- airline_traditional.ipynb: implements models such as logistic regression, SVM, KNN, Random Forest for baseline models
- deep_learning_airline_model.ipynb: Implements multiple deep learning models configurations as text-only, no-demographics, Hybrid (frozen), Hybrid + Fine-Tuning, and Hybrid +FT+ CW (Class Weights)

7. Setup and Installation

Pip install the following:

pip install pandas==2.2.2 numpy==1.26.4 scikit-learn==1.4.2 imbalanced-learn==0.12.2 \
torch==2.2.2 transformers==4.41.2 torchmetrics==1.3.2 matplotlib==3.8.4 seaborn==0.13.2 faker==24.9.0

8. Run the Project

- dataset_merge.ipynb → merged_dataset.csv (~64k rows)
- creating_synthetic_data.ipynb → merged_dataset_augmented.csv (~217,386 rows)
- final_airline_dataset.ipynb → final_airline_dataset.parquet/csv (train/test splits)
- airline_traditional.ipynb → classification reports, confusion matrices (SVM best baseline)
- deep_learning_airline_model.ipynb → ROC-AUC curves, F1 @ 0.30, loss/F1 trends, final checkpoints (Hybrid + FT best)