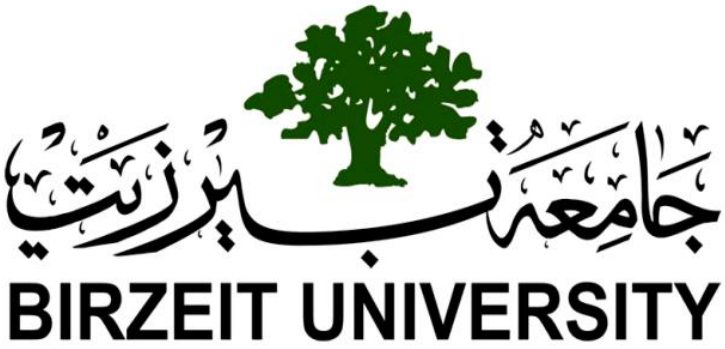

Faculty of Engineering & Technology

Department of Electrical and Computer Engineering

Machine Learning and Data Science

ENCS5341

Assignment 3

Prepared by:

Mohammad Abu Hijleh **ID: 1221350**
Hamed Musleh **ID: 1221036**

Instructor: **Dr. Yazan Abu Farha**

Section: **1**

Date: **13/1/2026**

Table of Contents

Table of Figure:

List of Table:

Introduction

Task Selection

For this assignment, we selected the task of **Multi-Class Text Classification**. The objective is to predict the country of a travel destination based solely on its textual description.

This task is significant for automated content tagging in travel applications. It presents specific Natural Language Processing (NLP) challenges, primarily because travel descriptions often share generic vocabulary (Like "beautiful beach," "historic ruins") across different countries, making simple keyword matching insufficient.

Models Explored

We implemented and evaluated three distinct types of machine learning models to solve this problem:

1. **Baseline:** k-Nearest Neighbors (k-NN) using TF-IDF features and Cosine distance.
2. **Tree-Based Model:** Random Forest Classifier (Ensemble method).
3. **Transformer (DistilBERT)**

Evaluation Metrics

Given the imbalance in the dataset (class sizes ranging from 15 to 64 samples), accuracy alone is a misleading metric. Therefore, we used the following metrics:

- **Macro F1-Score:** The primary metric for optimization and comparison, as it treats all classes equally regardless of size.
- **Accuracy:** Used for a general performance overview.
- **Confusion Matrix:** Used to analyze specific misclassifications between countries.

Data Preprocessing and Exploratory Data Analysis (EDA):

Data Preprocessing and Cleaning

The raw dataset contained 127 unique countries. To ensure statistical validity and model stability, the following preprocessing steps were applied:

1. **Filtering:** Countries with fewer than 15 samples were removed, reducing the dataset to **21 classes** and **656 total samples**. This step was crucial to avoid "few-shot" learning scenarios where the model would have insufficient data to generalize.
2. **Text Cleaning:** Descriptions were normalized by stripping leading/trailing whitespace and reducing multiple spaces to single spaces.
3. **Handling Missing Data:** Rows with empty descriptions or missing labels were excluded.

Dataset Characteristics

Post-filtering analysis revealed several challenges:

- **Class Imbalance:** The dataset remains highly imbalanced with a ratio of **4.27:1**. Japan is the most represented class (64 samples, 9.8%), while the UK, UAE, and USA are the least represented (15 samples, 2.3% each).
- **Description Length:** The text inputs are generally short, with a mean length of **25.08 words**. Short descriptions (some as short as 1 word) pose a challenge as they provide limited context for the model.

Interesting Trends and Challenges

Beyond standard statistics, the EDA revealed four critical challenges that complicate classification:

- **Generic Vocabulary Overlap:** Descriptors like *"beautiful"*, *"ancient"*, and *"famous"* appear across almost all classes. This creates significant semantic overlap, making a beach in **Greece** statistically similar to one in the **Maldives** without specific named entities.
- **Geographical Clustering:** Errors are not random; they cluster among geographically similar nations. For instance, Mediterranean countries (Italy, Spain, Greece) share architectural terms like *"ruins"* and *"cathedral"*, leading to frequent confusion.

- **The "Micro-Description" Problem:** Several samples contain fewer than 10 words (like, *"A beautiful view"*). These lack sufficient entropy for prediction and effectively act as noise in the training data.
- **Persistent Imbalance:** Even after filtering, the 4.27:1 imbalance ratio biases the model. Majority classes like **Japan** and **Italy** provide significantly richer vocabulary patterns than minority classes like the **UK** or **UAE**, making the latter harder to predict correctly.

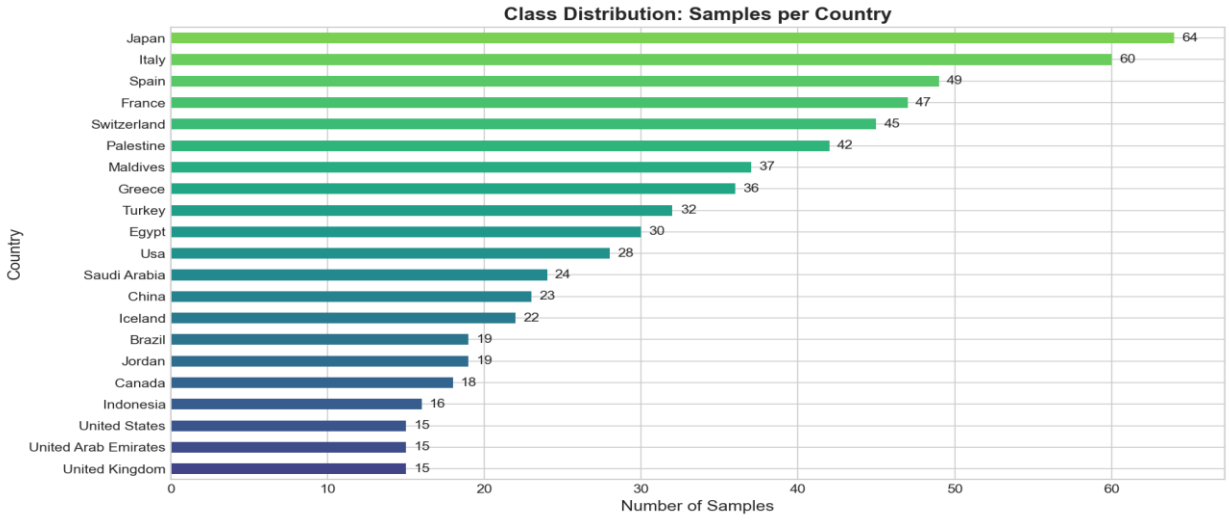

Figure 1: Distribution of samples per country after filtering. Note the dominance of Japan and Italy.

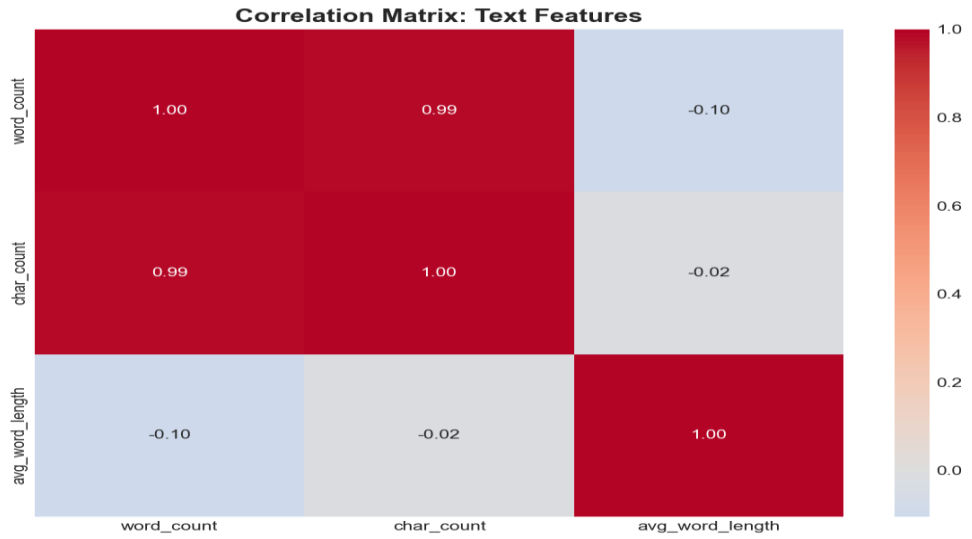

Figure 2: Distribution of description lengths. Most descriptions are concise (under 30 words).

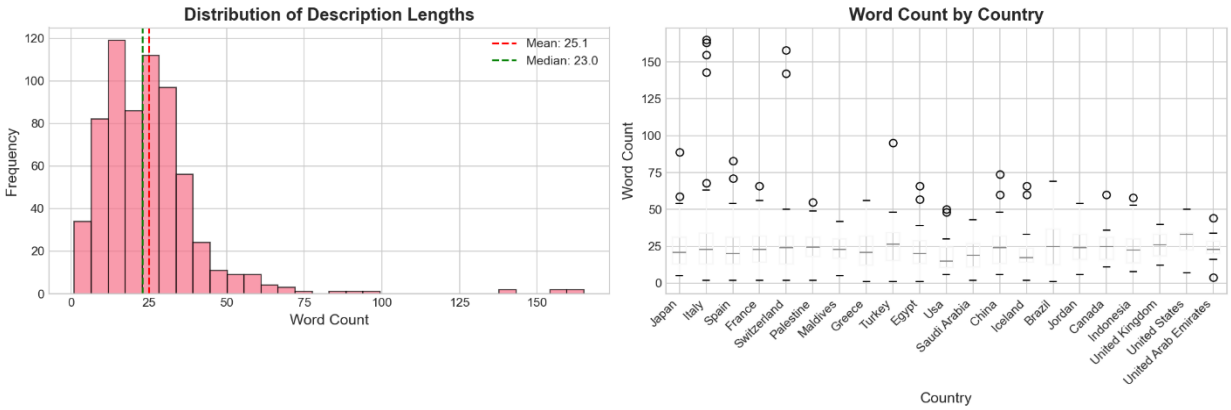

Figure 3: Correlation analysis of text length features (word count vs. character count).

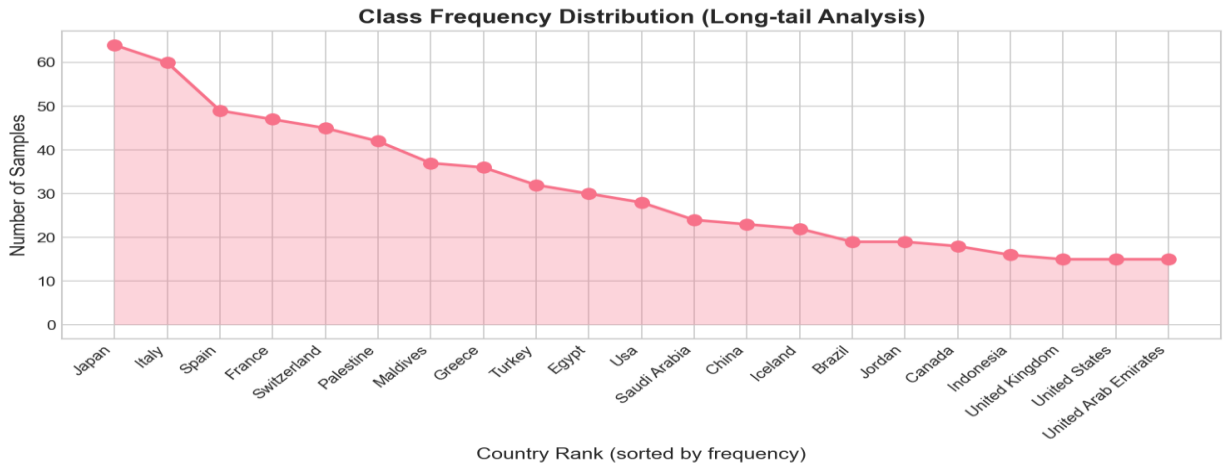

Figure 4: Sample count per country after filtering (min. 15 samples).

Methodology and Models

Feature Engineering

Text data was converted into numerical vectors using **TF-IDF (Term Frequency-Inverse Document Frequency)**.

- **N-grams:** Unigrams and bigrams (1, 2) were used to capture local context (like, "ancient ruins" vs. "modern ruins").
- **Vocabulary:** Limited to the top 5,000 features to reduce dimensionality and overfitting.
- **Stop Words:** Standard English stop words were removed to focus on content-bearing terms.

Baseline Model: k-Nearest Neighbour (k-NN)

A k-NN classifier was chosen as the baseline to establish a performance benchmark based on similarity in the vector space.

- **Distance Metric: Cosine Distance** was selected instead of Euclidean distance. In high-dimensional sparse spaces (like TF-IDF), the angle between vectors (cosine) is a more reliable measure of semantic similarity than magnitude.
- **Configurations Tested:** k=1 and k=3.

Proposed Model 1: Random Forest Classifier

Why Selected: Random Forest is an ensemble learning method that is robust to overfitting and can capture non-linear relationships between features. It handles high-dimensional data well and provides feature importance, which is useful for interpretability.

- **Hyperparameter Tuning:** A Grid Search with 5-fold cross-validation was performed.
    - n_estimators: Tested [50, 100, 200, 300]
    - max_depth: Tested [10, 20, 30, None]
    - class_weight: Set to 'balanced' to explicitly counter the dataset imbalance.
- **Best Configuration:** n_estimators=200, max_depth=30, class_weight='balanced'.

Proposed Model 2: Transformer (DistilBERT)

Why Selected: To overcome the limitations of bag-of-words models (like k-NN and Random Forest) which ignore word order and context, we implemented a Transformer-based model. We selected **DistilBERT** (distilbert-base-uncased), a distilled version of BERT that retains 97% of BERT's performance while being 40% lighter and 60% faster. Unlike TF-IDF, DistilBERT utilizes self-attention mechanisms to capture deep semantic relationships and context (like, distinguishing "resort" in a snowy context vs. a beach context).

- **Hyperparameter Tuning:**

Unlike the Random Forest model where we performed an exhaustive Grid Search, for the Transformer, we utilized **manual hyperparameter optimization** combined with **Early Stopping**.

Training Transformers is computationally expensive, so we selected hyperparameters known to be effective for fine-tuning distilbert-base-uncased on small datasets:

- **Optimizer & Learning Rate:** We used the **AdamW** optimizer with a low learning rate of **1e-5**. This is crucial for "fine-tuning" to avoid destroying the pre-trained weights of the model.

- **Scheduler:** A linear learning rate scheduler with a **warmup ratio of 0.1** (10% of training steps) was implemented. This gradually increases the learning rate at the start, stabilizing the training process.
- **Regularization:**
    - **Weight Decay:** Set to 0.01 to penalize large weights.
    - **Dropout:** Set to 0.1 to prevent overfitting on our small dataset.
- **Early Stopping (Implicit Tuning):** We implemented Early Stopping with a **patience of 3 epochs** monitoring the Validation Macro F1-Score.
    - *Result:* Although set to run for 10 epochs, the model stopped training at **Epoch 7**, effectively "tuning" the training duration to the point of maximum generalization before overfitting began.

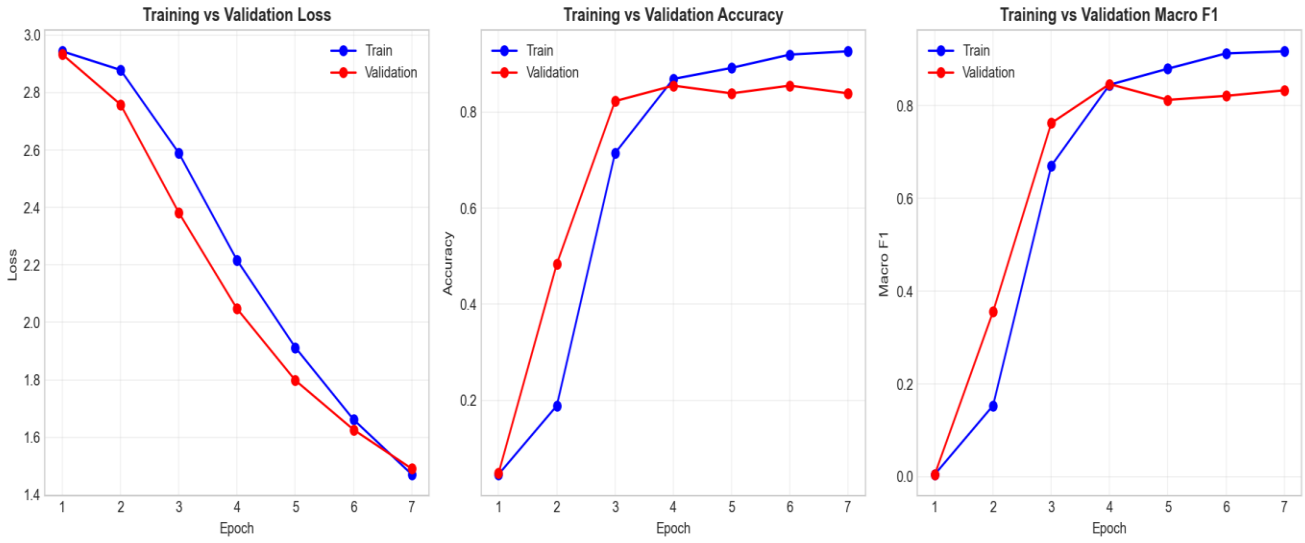

Figure 5: Transformer Model Training History

Experiments and Results

Quantitative Results

The experiments were conducted on a held-out test set (20% of data, stratified). The Transformer (DistilBERT) model achieved the highest performance across all metrics.

Model	Accuracy	Macro F1-Score	Weighted F1-Score
k-NN (k=1)	0.7419	0.7125	0.7319
k-NN (k=3)	0.7581	0.7302	0.7473
Transformer (DistilBERT)	**0.8710**	**0.8636**	**0.8711**
Random Forest	0.8065	0.7758	0.7913

Table 1: Models Evaluation Metrics Results

Interpretation:

- **Superior Accuracy:** DistilBERT achieved **87.10% accuracy**, outperforming Random Forest (80.65%) by **6.45%**.
- **Minority Class Gains:** The **Macro F1-Score** rose to **0.8636** (+8.78%), with significant improvements in underrepresented classes like **USA** (0.29 to 0.60) and **China** (0.40 to 0.86).
- **Baseline Viability:** k-NN (k=3) remained competitive (**75.81%**), confirming that simple keywords explain most data, while deep learning resolves the complex cases.

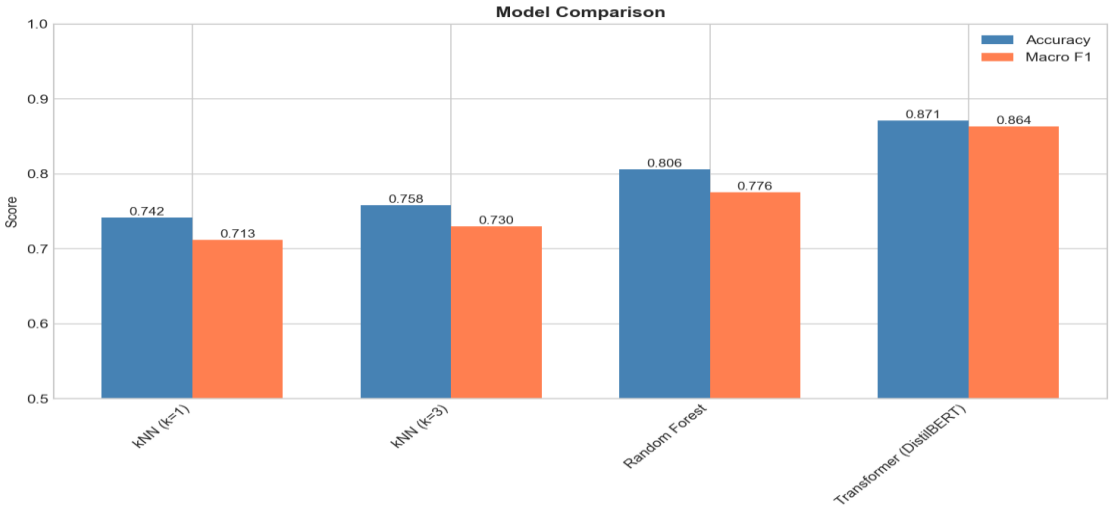

Figure 6: Models Evaluation Comparison

Performance Analysis

Error Patterns and Confusion Analysis

We compared the error profiles of the best traditional model (Random Forest) and the deep learning model (Transformer). The Transformer reduced the total number of misclassifications from **24** to **16**.

1. Mitigation of "Generic Vocabulary" Issues: The Random Forest model frequently struggled with short, generic descriptions (e.g., descriptions containing just "resort" or "beautiful view"). The Transformer's self-attention mechanism allowed it to infer context even from limited text, significantly improving performance on these "micro-descriptions."

2. Persistent Semantic Confusion: Despite the improvements, some confusion pairs remain difficult for both models, likely due to extreme semantic overlap:

- **Switzerland vs. Iceland:** Both models struggle to distinguish these based on descriptions of mountains and nature.
- **Greece vs. Turkey:** Shared Mediterranean features (like , "ancient ruins," "coastal") continue to cause minor confusion.

3. Impact on Minority Classes: The Transformer demonstrated superior generalization for classes with few samples. In the Random Forest model, **China** and **USA** had very low recall (incorrectly predicting them as other countries). The Transformer corrected most of these errors, proving that pre-trained embeddings effectively compensate for data scarcity.

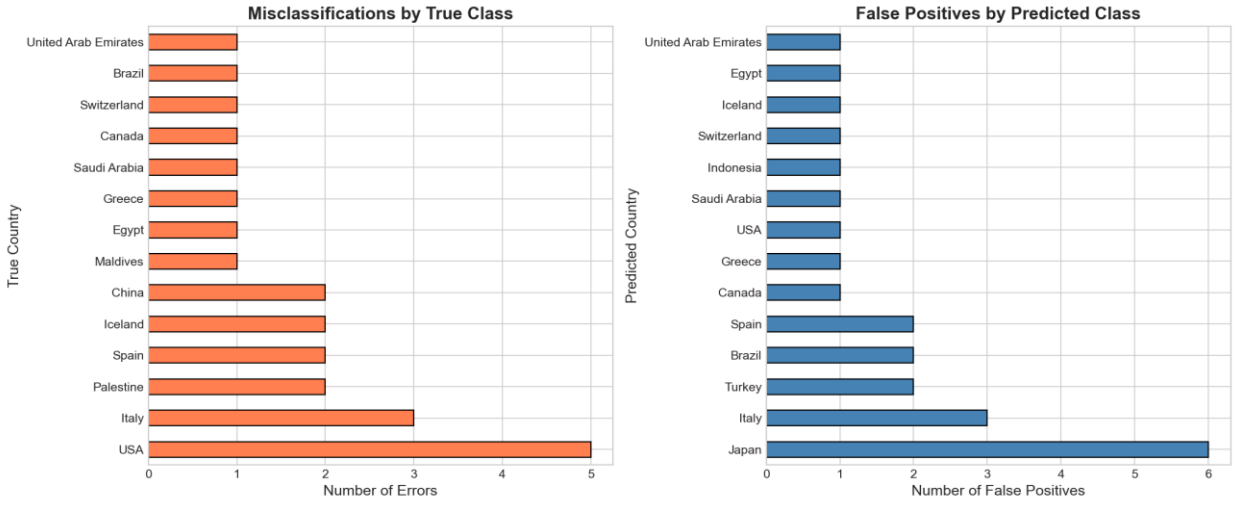

Figure 7: Error Distribution Per Class

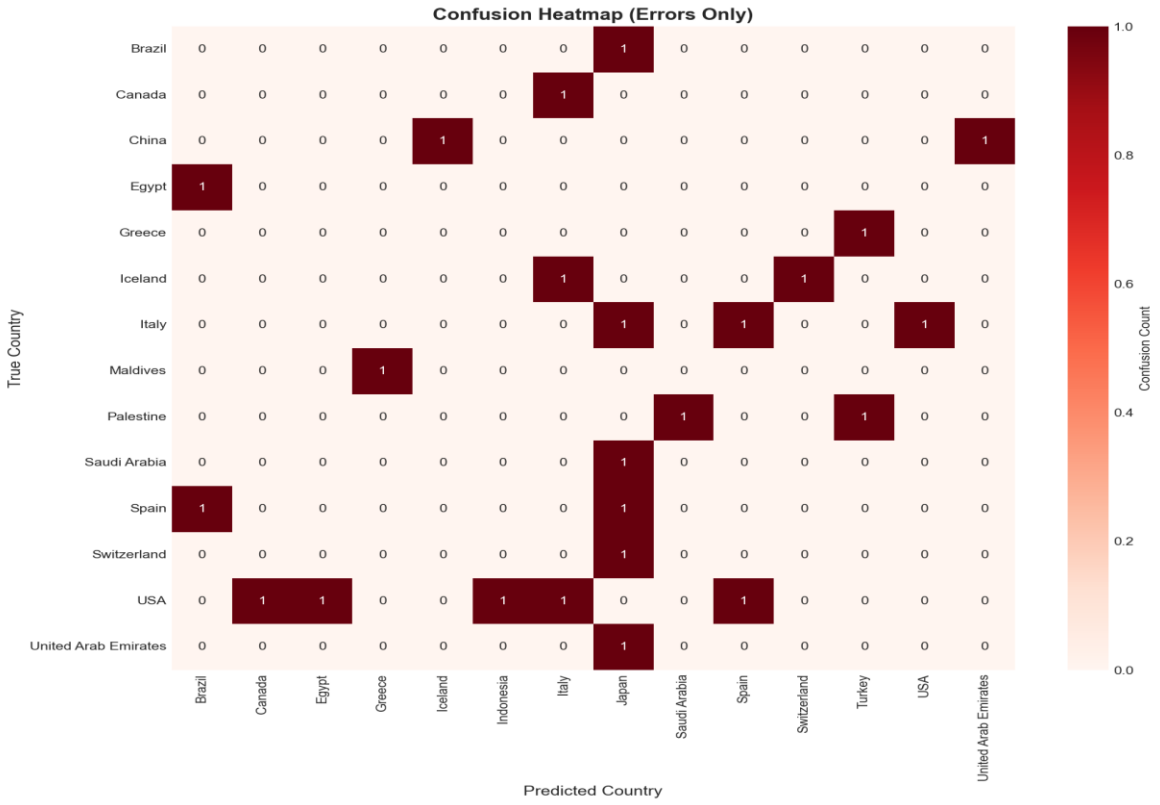

Figure 8: Confusion Heatmap Errors

Insights

- **Syntax vs. Semantics:** The shift to Transformers proved that previous errors were syntactic (missing words). The remaining errors are "semantic dead ends" ambiguous descriptions applicable to multiple destinations that require external knowledge to resolve.
- **Solving Imbalance:** Pre-training outperformed algorithmic class-weighting. The Transformer's prior knowledge of English allowed it to generalize on minority classes (USA F1: 0.29 to 0.60) without the aggressive tuning required by the Random Forest.

Conclusions and Discussion

Summary of Findings

This project successfully demonstrated that travel destination classification can be significantly improved by moving from bag-of-words models to deep learning. While the **Random Forest** classifier performed well (80.65%), the **DistilBERT Transformer** established a new benchmark with **87.10% accuracy**. The Transformer's ability to understand semantic context allowed it to solve difficult cases such as distinguishing semantically similar countries and correctly classifying minority classes that traditional models missed.

Limitations

- **Computational Cost:** The Transformer is computationally expensive (10 mins training vs. seconds for RF), making the Random Forest preferable for low-latency edge applications.
- **Interpretability:** Unlike the Random Forest's clear feature importance lists, the Transformer is a "black box," making it difficult to diagnose specific misclassifications without complex attention mapping.
- **Data Ceiling:** Accuracy plateaued at 87%, suggesting a quality ceiling caused by "noisy" samples (URLs, single words) that contain no predictive signal regardless of model complexity.

Future Work

To improve the system, we thought of:

- **Data Augmentation:** Generate synthetic training samples via LLMs to directly address imbalance in minority classes (like, UAE, USA).
- **Ensemble Methods:** Combine Random Forest (keywords) and Transformer (context) predictions to capture both syntactic and semantic signals.
- **Hierarchical Classification:** Implement a two-stage strategy (Continent → Country) to minimize confusion between geographically similar neighbors like Greece and Turkey.