

NOMMON

A semi-supervised
approach to multi-label
classification of NOTAMs
using BERT



INTRODUCTION

METHODOLOGY

EXPERIMENTS

CONCLUSIONS



Introduction

About the project

- **Project:** GenAI models for ATM
- **Call:** SESAR3 ENGAGE2 KTN 1st Call for Catalyst Funding
- **Thematic Challenge:** Disruptive ATM system modernisation
- **Start:** 9th September 2024
- **Duration:** 12 months
- **Partners:** Nommon (Coordinator), Tecnalia

NOMMON

tecnalia

Project objectives



To explore how **generative AI (GenAI)** can **enhance the ATM** system by:

Analyse ATM challenges suitable for GenAI

Review operational needs and identify ATM processes where generative AI (e.g. LLMs, GANs) can provide clear added value, selecting the most promising use cases for further investigation

Develop and demonstrate a GenAI proof-of-concept

Build prototype solution for the selected GenAI use case and validate their potential through a dedicated case study applied to real or realistic ATM scenarios

Project objectives



To explore how **generative AI (GenAI)** can **enhance the ATM** system by:

Analyse ATM challenges suitable for GenAI

Review operational needs and identify ATM processes where generative AI (e.g. LLMs, GANs) can provide clear added value, selecting the most promising use cases for further investigation

Develop and demonstrate a GenAI proof-of-concept

Build prototype solution for the selected GenAI use case and validate their potential through a dedicated case study applied to real or realistic ATM scenarios

Use case: Annotating NOTAMs with a tag system using Gen AI



Goal: Automatically contextualize NOTAMs through multi-label tagging to reduce misunderstandings and ease information overload.



ATM challenge addressed: NOTAMs are numerous, complex, and often cover multiple topics. Single-label approaches are not expressive enough and limit automation and filtering.



How it helps: A GenAI-powered multi-label BERT model assigns accurate, structured tags even for multi-topic and rare NOTAM cases, improving readability and supporting safer decision-making.



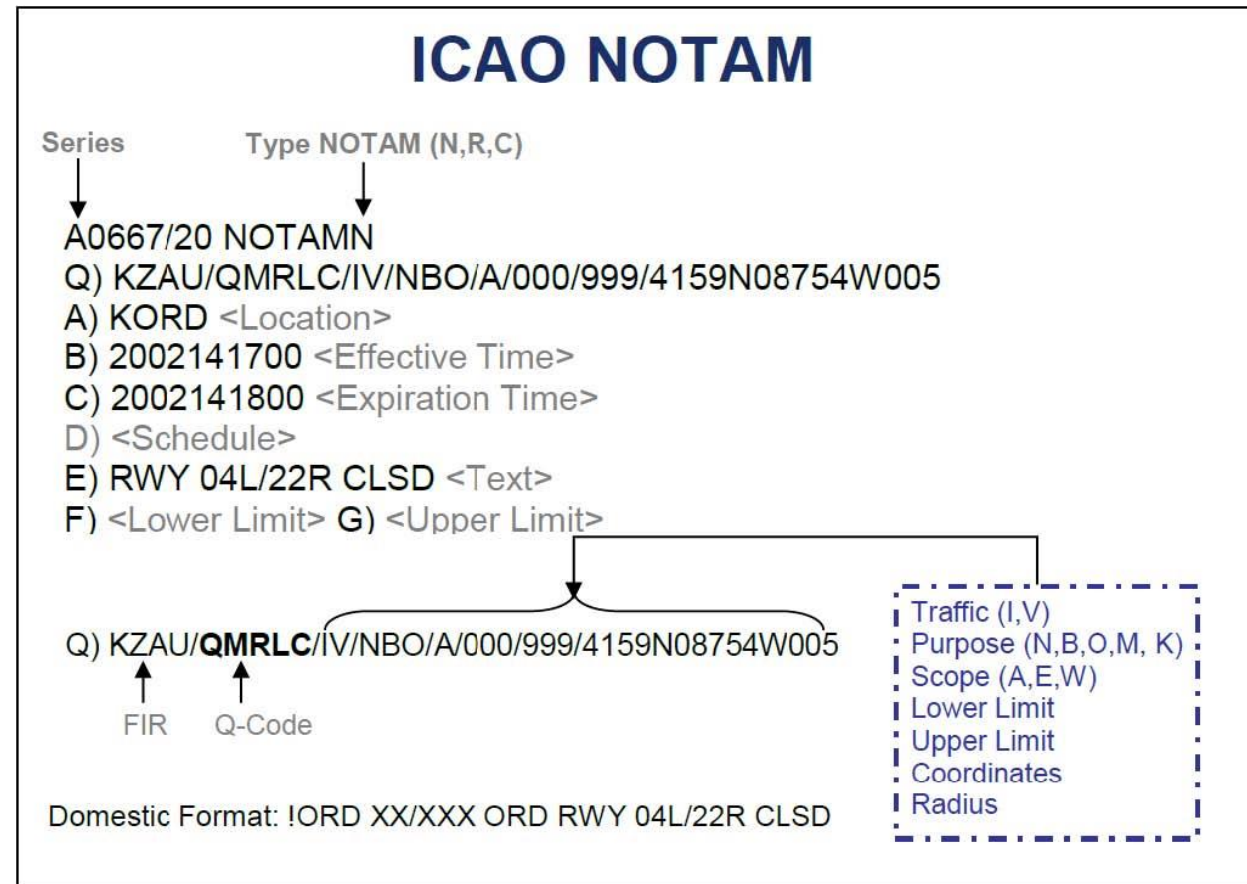
GenAI added value:

- Enables non-mutually-exclusive tags, better reflecting real NOTAM semantics (validated with Lufthansa Systems).
- Leverages unlabelled data to improve performance
- Reduces dependence on expert annotation while capturing complex label interactions

What is a NOTAM?

A **NOTAM** (Notice to Air Missions) is an **official message used to communicate** time-critical information that is essential for flight operations. It typically reports temporary changes affecting aerodromes, airspace, navigation aids, procedures, or infrastructure.

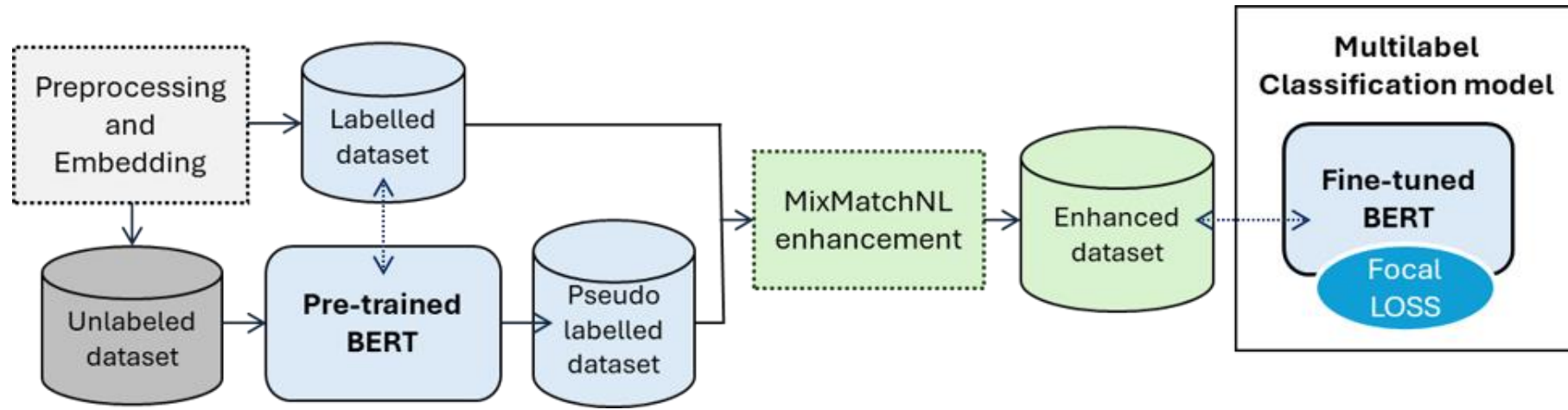
NOTAMs follow a **structured ICAO format** (series, location, validity, and free-text description), but the information is often presented as dense, unstructured text, making automatic processing and classification challenging.



A photograph of an airport tarmac under a clear blue sky. In the foreground, the large engine and wing of a white commercial airplane are visible. Several ground crew members in orange and blue uniforms are working around the aircraft. In the background, other planes and airport buildings are visible. A semi-transparent green rectangular box is overlaid on the right side of the image, containing the word 'Methodology' in white text.

Methodology

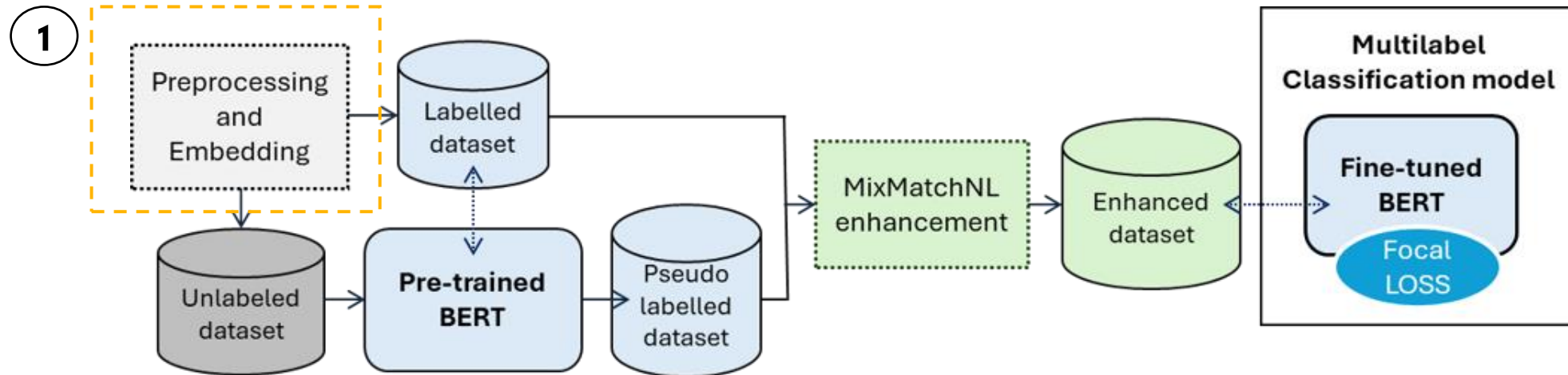
Model overview



Why the semi supervised approach?

- Most **NOTAMs** are **unlabeled**, and expert **annotation** is **costly and time-consuming**.
- The **label distribution** is highly **imbalanced** (some categories are rare), making pure supervised learning insufficient.
- **MixMatchNL leverages unlabelled data** to improve generalization and handle multi-topic NOTAMs more robustly.
- Combining pseudo-labels with **focal loss boosts performance**, especially for **underrepresented or complex classes**.

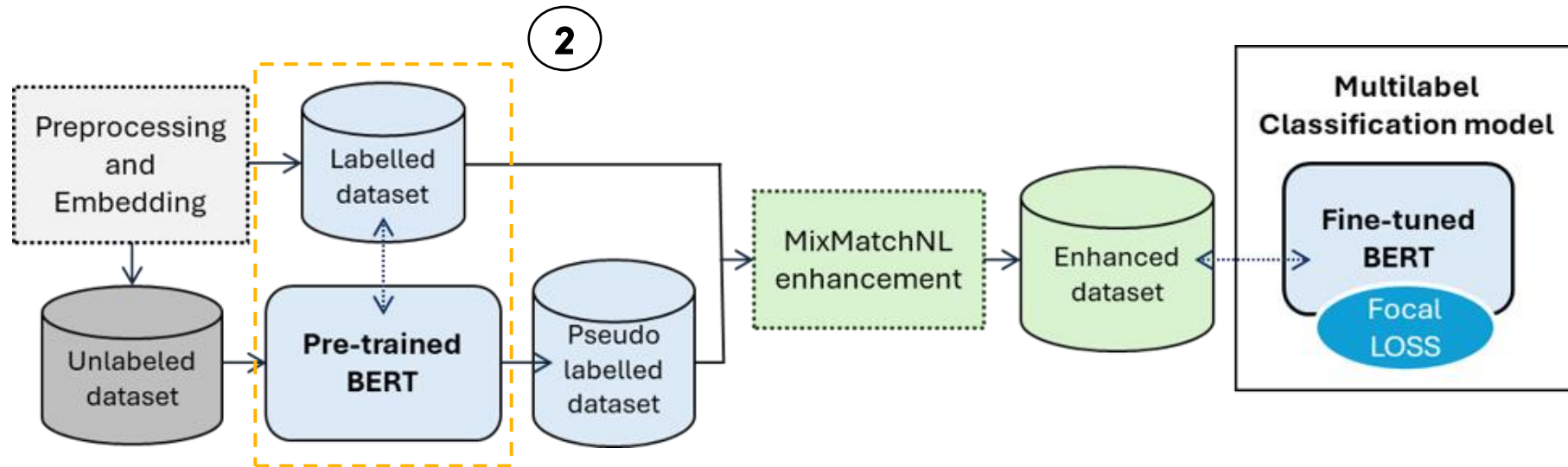
1. Preprocessing and embedding



- **Expand aviation abbreviations** and to provide BERT with full-word forms and reduce out-of-vocabulary tokens.
- **Normalize and clean the NOTAM** text (tokenization, lowercasing, removal of formatting artefacts)
- Convert each NOTAM into **BERT-compatible tokens** using the bert-base-uncased tokenizer.
- Extract **contextual BERT embeddings** to represent both labelled and unlabelled NOTAMs in a unified feature space.

Expanding abbreviations and producing consistent embeddings improves BERT's ability to understand domain-specific terminology and supports later semi-supervised mixing between labelled and unlabelled data.

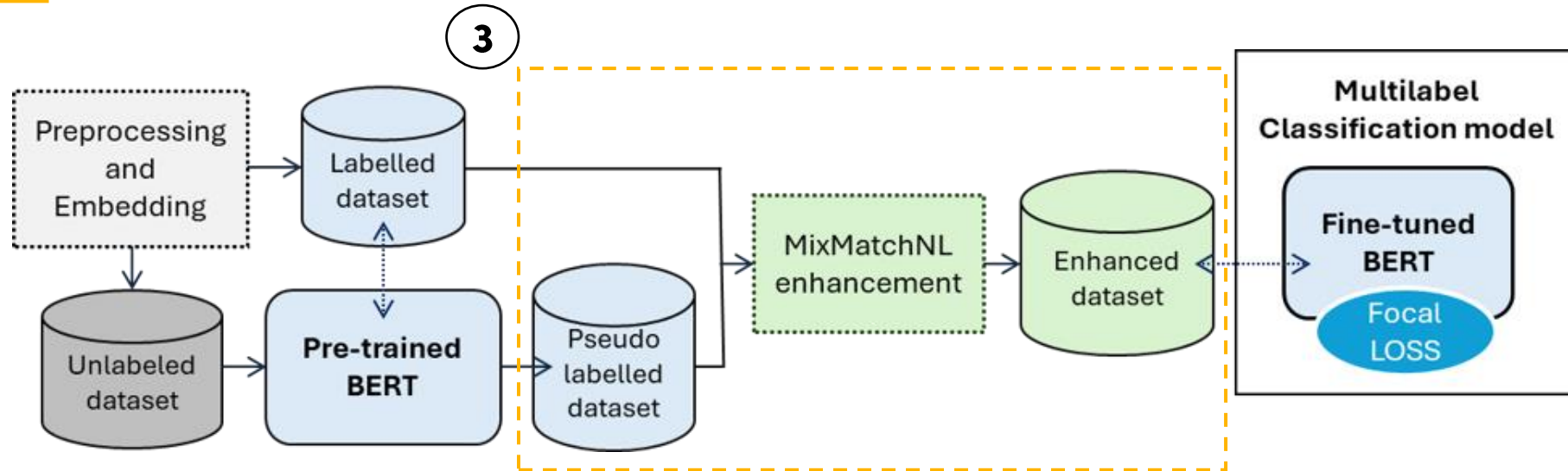
2. Supervised pretraining



- **Fine-tune the pretrained BERT** model using the manually labelled NOTAM dataset
- Train the model to predict **multiple tags** simultaneously, adapting the output layer for multi-label classification
- Use **binary cross-entropy loss** and a small batch size to ensure stable training under limited data.
- Extract contextual BERT embeddings to represent both labelled and unlabelled NOTAMs in a unified feature space.

This supervised model becomes the **baseline** and is used to produce pseudo-labels for the unlabelled NOTAMs in the next stage.

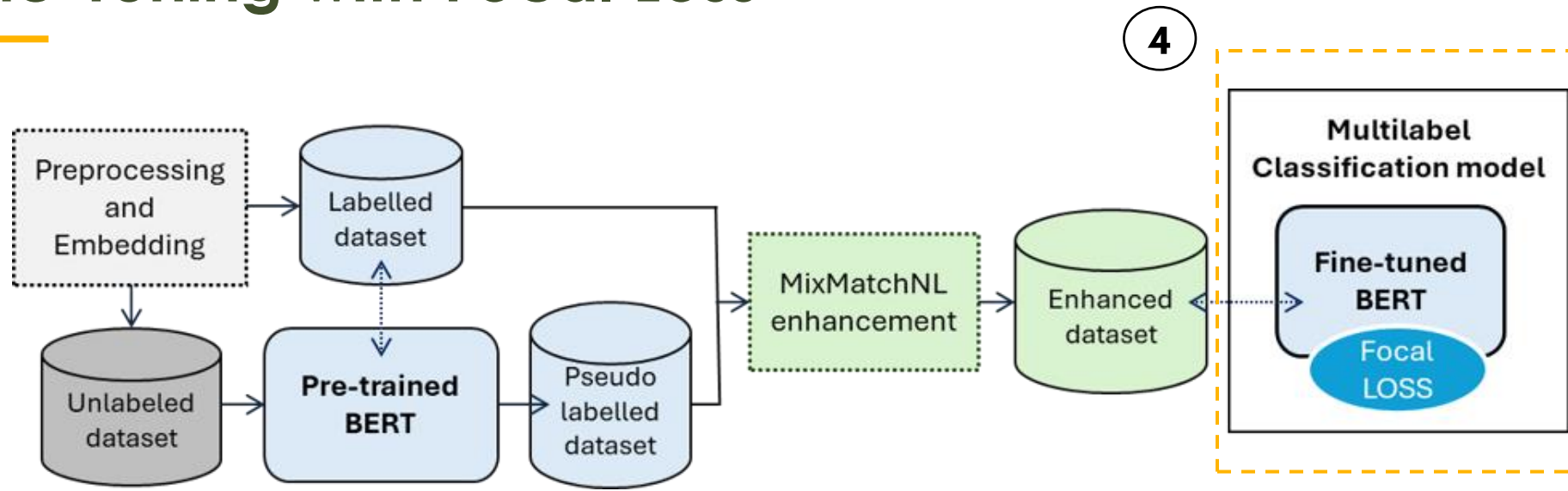
3. Semi-Supervised data enhancement with Mix-MatchNL



MixMatchNL is a semi-supervised method that blends labelled and unlabelled NOTAMs to exploit the structure of the large unlabelled dataset, improving generalization and boosting performance on multi-topic and rare categories.

- Generate **pseudo-labels** for unlabelled NOTAMs using the supervised BERT model and apply **sharpening** to increase consistency.
- Extract BERT embeddings for labelled and pseudo-labelled NOTAMs and **mix them using MixUp interpolation**
- Use the resulting synthetic samples to **augment the training set**, improving generalization and robustness to rare or multi-topic NOTAMs.

4. Fine-Tuning with Focal Loss



- **Fine-tune the BERT model on the enhanced dataset** (labelled + MixMatchNL synthetic samples).
- Replace standard Binary Cross-Entropy with **Focal Loss** to handle the strong label imbalance.

Focal Loss prioritizes hard and underrepresented NOTAM labels by down-weighting easy examples, addressing the strong label imbalance that standard BCE cannot handle.



Experiments

Dataset

To train and evaluate our model, we compiled a large dataset of NOTAMs from FAA and ICAO sources, focusing on the free-text field where operational information is encoded.

Total NOTAMs collected: 107,371

Sources:

- FAA-format dataset
- ICAO-format dataset

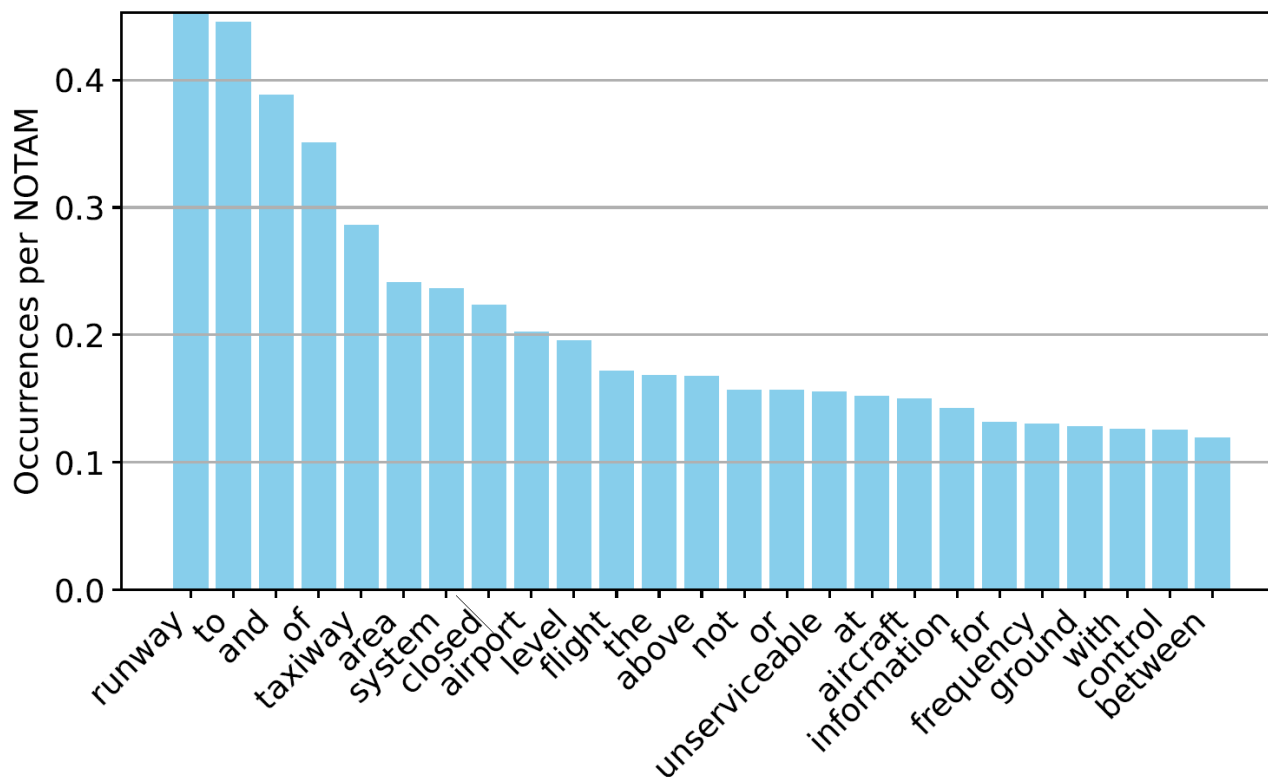
Field used: Free-text section (ICAO E)

Language preprocessing:

- Expansion of abbreviations & airport codes (iterative regex rules)
- Reduces out-of-vocabulary terms for BERT

Manually labelled NOTAMs: 8,497

Test set: 1,700

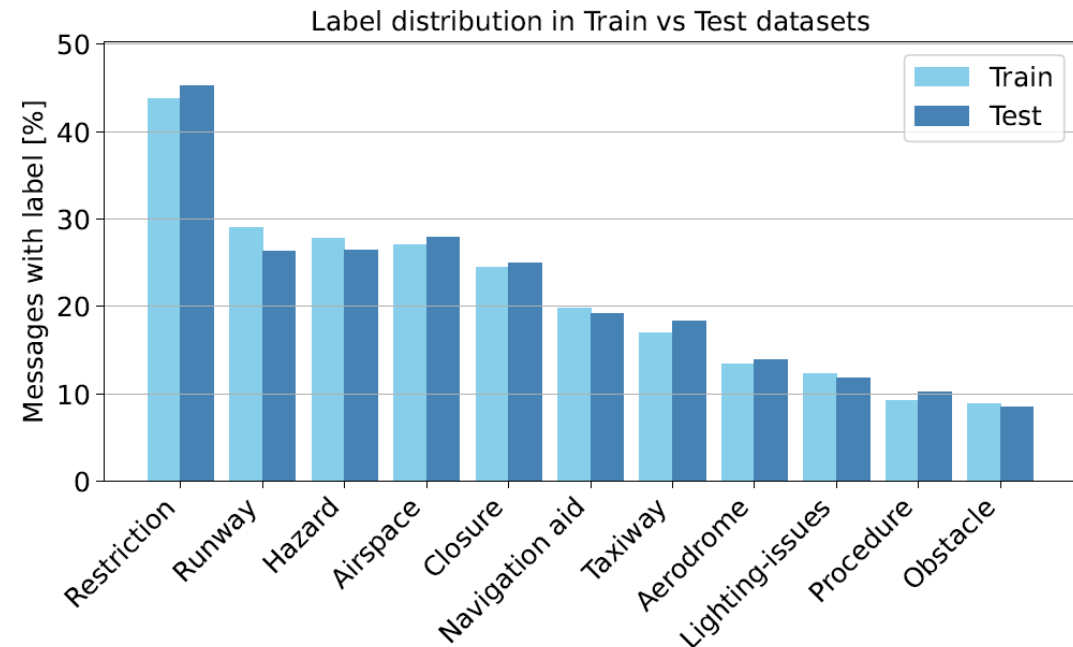
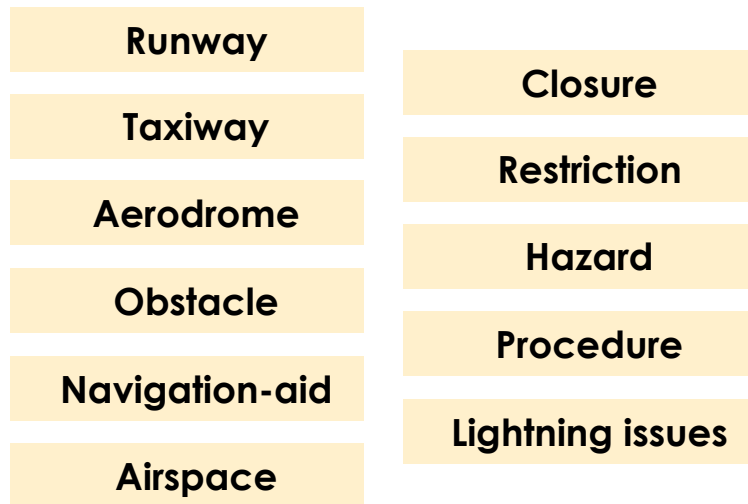


Most frequent words after abbreviation expansion

Labelling system

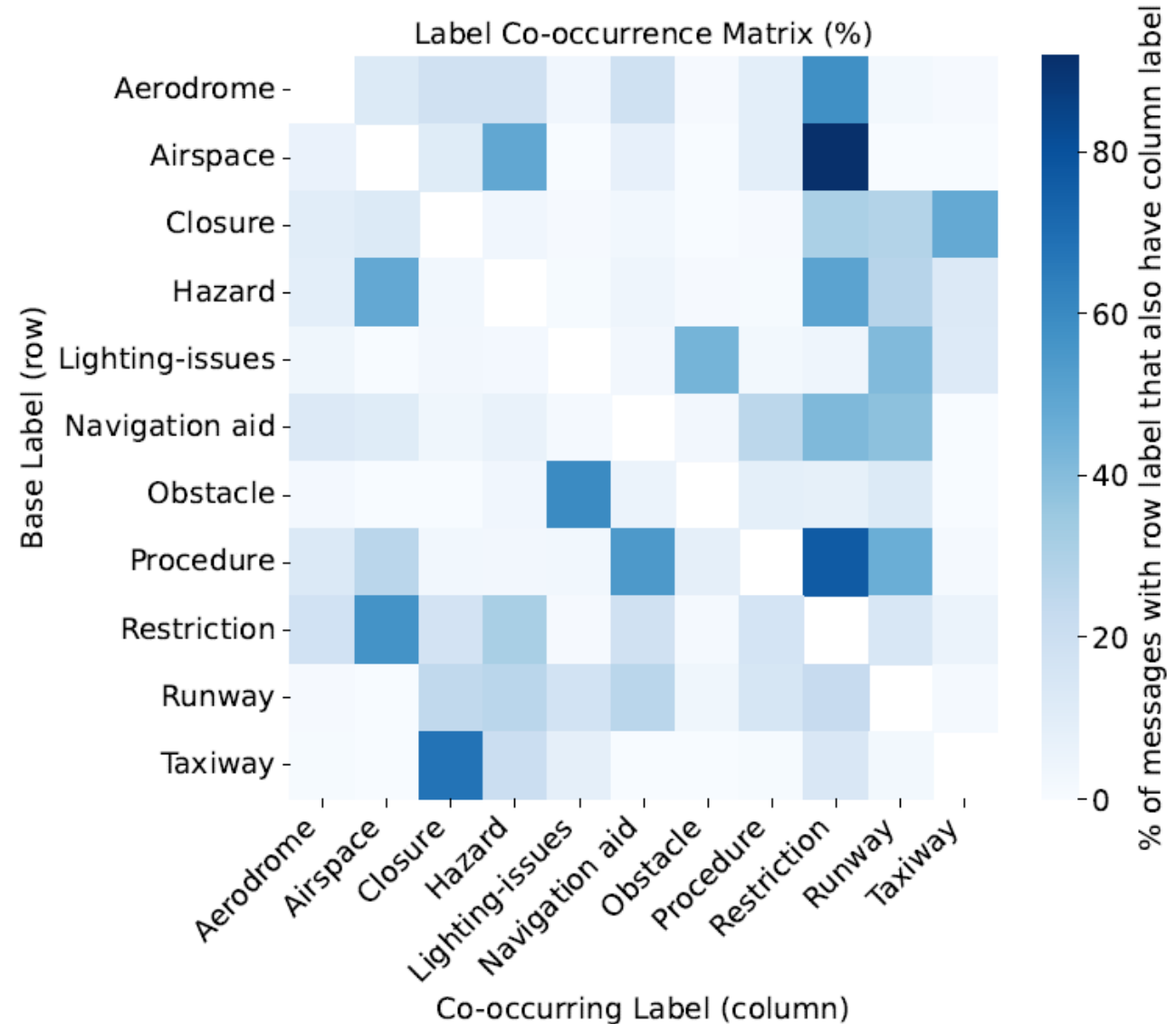
We designed an **11-label multi-label system** that captures the key operational aspects of NOTAMs while keeping the taxonomy expressive yet simple and non-exclusive.

- Based on FAA categories and OPSGROUP's 50-tag proposal, **reduced to 11 for clarity and usability**.
- Allows **non-mutually exclusive** combinations: a NOTAM may concern both a subject (e.g., *Runway*) and a state (e.g., *Closure*).
- Validated with an end-user (Lufthansa Systems): **supports operational workflows** efficiently.

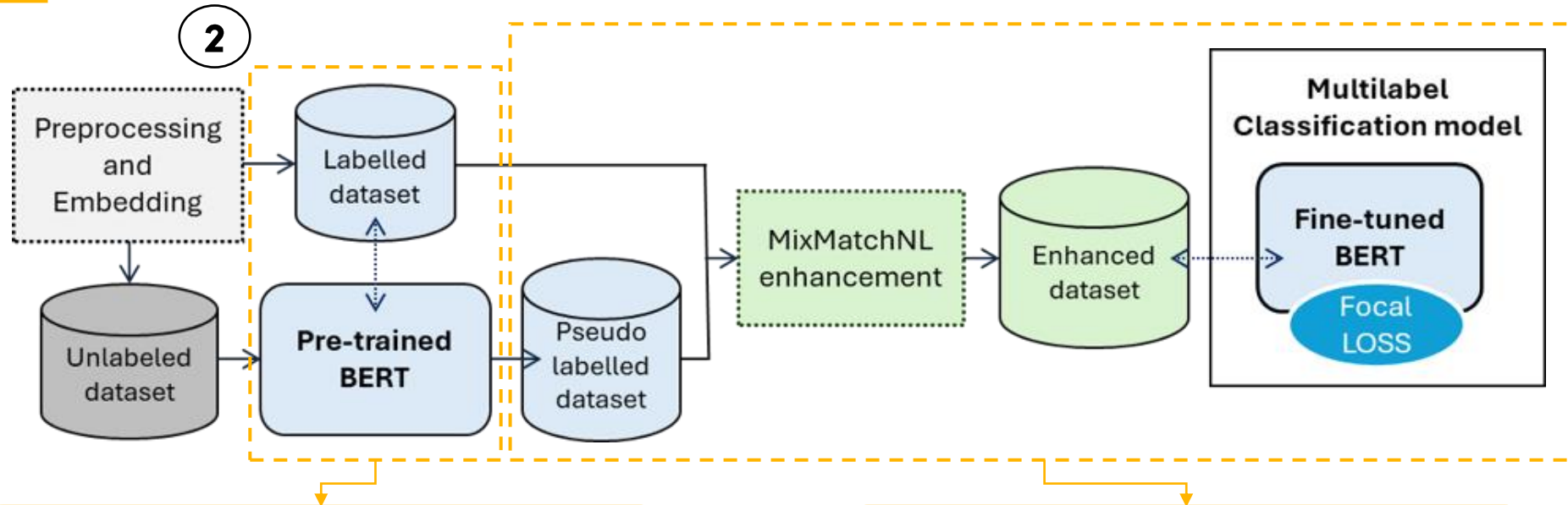


Labelling system

- The matrix values represent the **percentage of NOTAMs with the row label that also contain the column label**.
- Clear patterns emerge: subject-related labels (e.g., *Runway*, *Airspace*) frequently co-occur with state labels (e.g., *Closure*, *Restriction*).
- However, **many label combinations are rare or nearly absent**, resulting in a highly sparse output space and making the multi-label classification task more challenging.



Experimental setup



Supervised training (baseline)

- Training set: **6,797 labelled NOTAMs**
- Test set: **1,700 NOTAMs**
- Loss: **Binary Cross-Entropy**
- Optimizer: **AdamW**, learning rate **1e-5**
- Batch size: **4**, epochs: **8**
- Purpose: learn initial decision boundaries and generate **pseudo-labels** for unlabelled data

Semi-supervised training

- Unlabelled NOTAMs used: **98,750**
- Method: **MixMatchNL**
 - Pseudo-labelling + sharpening
 - MixUp interpolation ($\alpha = 0.75$)
- Final fine-tuning with **Focal Loss** ($\gamma = 2$) to address label imbalance
- Learning rate: **1e-5**, epochs: **8**

Evaluation metrics

To assess the performance of our multi-label classification model, we employ a comprehensive set of evaluation metrics that capture both label-level and instance-level behaviour.

Label-based metrics (multi-label quality per tag):

We compute **precision**, **recall** and **F1-score** for each label using standard multi-label averaging strategies:

- **Micro** – Global performance giving more weight to frequent labels.
- **Macro** – Averages performance across labels equally (good for rare labels).
- **Weighted** – Like macro, but weighting by label frequency.

Instance-based & ranking metrics (quality per NOTAM)

- **Hamming Loss** – fraction of incorrect label assignments.
- **Subset Accuracy** – all labels correct for the whole NOTAM.
- **Jaccard Index** – similarity between predicted and true label sets.
- **Coverage Error** – how deep we must go in the ranked list to cover all true labels.
- **LRAP** – ranks true labels above non-true labels.

Results

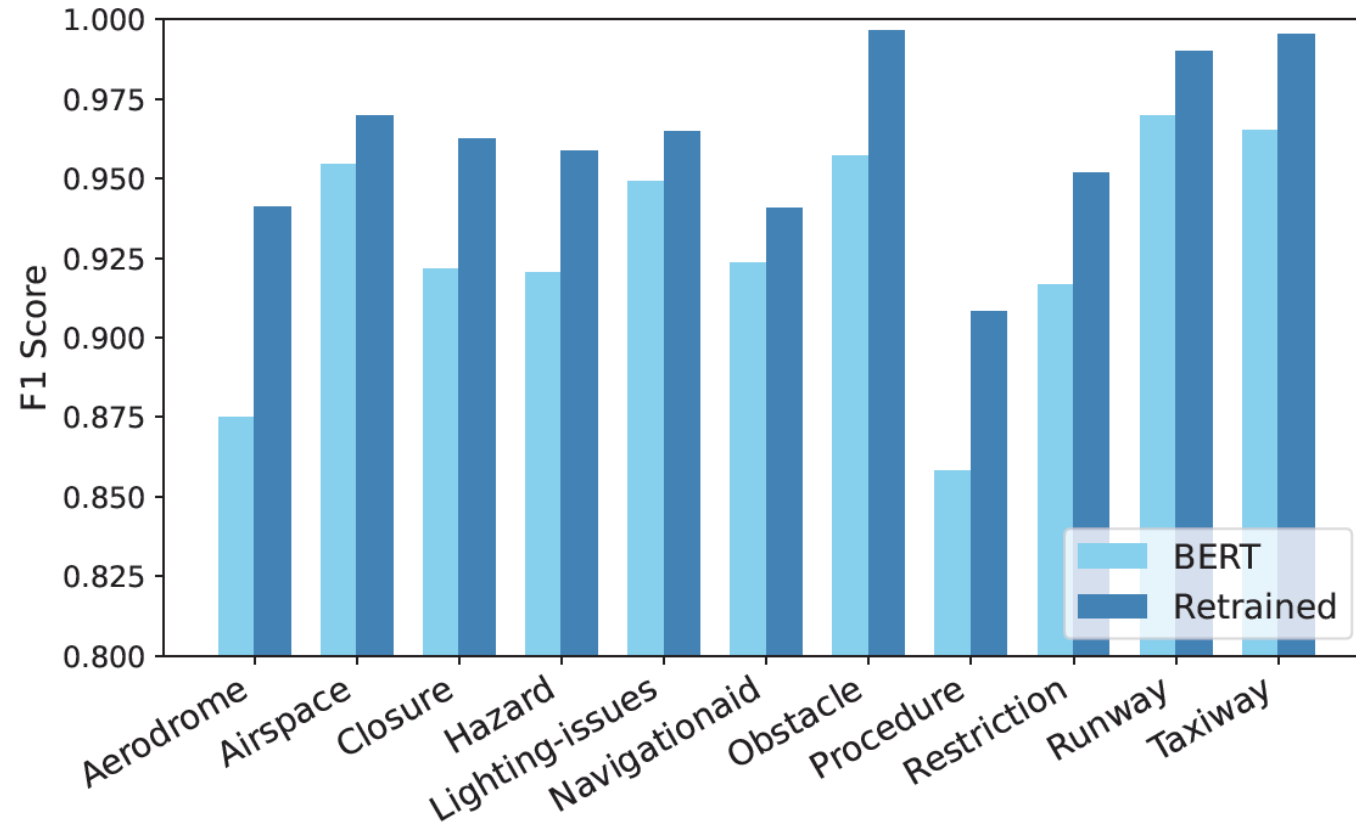
We compare the baseline BERT model with the semi-supervised version enhanced with MixMatchNL and Focal Loss. The semi-supervised approach yields consistent improvements across all metrics.

- **Both models evaluated on the same test set** of 1,700 labeled NOTAMs.
- The semi-supervised model (**MixMatchNL + Focal Loss**) **outperforms** the baseline BERT **across all metrics**.
- **Micro F1-score:** 0.93 → **0.96** (clear global improvement).
- **Macro F1-score:** 0.93 → **0.96** (better performance even on rare labels).
- **Hamming Loss reduced by half:** 0.0294 → **0.0159**.
- **Subset accuracy:** 0.78 → **0.88**, showing more exact multi-label predictions.
- **Jaccard Index:** 0.901 → **0.945** (higher overlap with true label sets).
- **Coverage Error:** 3.54 → **3.02**, meaning fewer labels to rank before covering all true labels.
- **LRAP:** 0.91 → **0.95**, indicating improved prioritization of relevant labels.

Metric	BERT + (Pretrained)	BERT + MixMatchNL + Focal Loss
Micro-Averaged		
Precision	0.9316	0.9640
Recall	0.9295	0.9608
F1-score	0.9306	0.9624
Macro-Averaged		
Precision	0.9338	0.9651
Recall	0.9231	0.9591
F1-score	0.9282	0.9617
Weighted-Averaged		
Precision	0.9320	0.9646
Recall	0.9295	0.9608
F1-score	0.9305	0.9624
Instance-Based metrics		
Hamming Loss	0.0294	0.0159
Subset Accuracy	0.7812	0.8753
Jaccard Index	0.9010	0.9452
Coverage Error	3.5453	3.0241
LRAP	0.9122	0.9515

Results

Comparison of **per-label F1-scores** on the test dataset between the pretrained BERT model and the retrained model using MixMatchNL and focal loss.

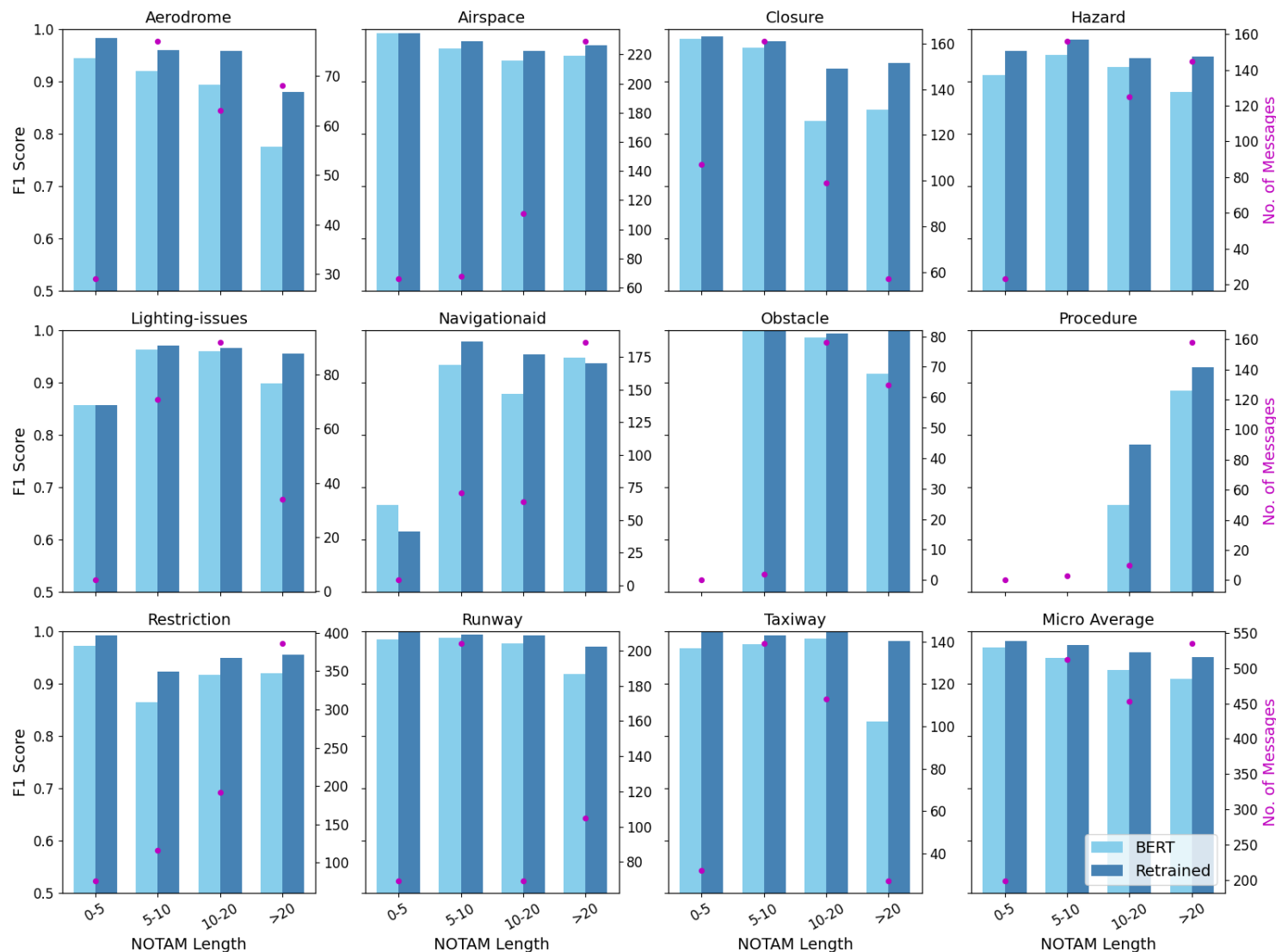


Per-label F1-score comparison (BERT vs. retrained)

- Retrained model improves F1 on **almost all labels**.
- Strong gains in **Aerodrome, Closure, Obstacle, Taxiway** (last two reach **F1 = 1.00**).
- Frequent labels (**Lighting-issues, Navigation aid, Runway**) also improve.
- **Procedure** and **Restriction** show better precision-recall balance.
- **Procedure** remains hardest label (F1 $\approx$ 0.91).

Results

- Both models maintain **high F1** on all label-length groups with **≥ 5 samples** (mostly **>0.90** for the retrained model).
- The retrained model **outperforms BERT** consistently across all lengths.
- Gains grow with message length:** biggest improvements in NOTAMs **>20 words** (micro-F1: **$0.91 \rightarrow 0.96$**).
- Short NOTAMs show **small improvements**, but long NOTAMs benefit significantly from MixMatch + Focal Loss.
- Despite improvements, **performance still drops** for longer NOTAMs in both models.



Per-label F1-scores by NOTAM length (BERT vs. retrained)

A photograph of an airport at sunset. A large commercial airplane is on the runway, with its shadow cast long on the tarmac. In the background, a control tower and airport buildings are visible. A green rectangular overlay is positioned in the upper right, containing the word "Conclusions" in white text.

Conclusions

Conclusions

- We presented a **semi-supervised multi-label approach** for annotating NOTAMs, combining BERT, MixMatchNL and focal loss
- Our **two-phase training strategy** (supervised + semi-supervised with pseudo-labels) consistently improves classification performance
- The **raw data**, processed data and results are publicly **available** in a public repository
- **Future work** includes:
 - Refine the labelling system and explore hierarchical classification to improve structure.
 - Test stronger transformer models (RoBERTa, Longformer).
 - Use aviation-specific corpora for domain pretraining.
 - Investigate active learning to reduce annotation costs.

CONTACTO

David Mocholí González

Aviation Director

david.mocholi@nommon.es

www.nommon.es

