A Semi-supervised Approach to Multi-label Classification of NOTAMs using BERT

O. Rey Orozco,
L.M. Viso Dominguez
Tecnalia Research and Innovation
Derio, Spain

A. Montoya Santamaría, D. Mocholí Gonzalez,
O. García Cantú-Ros
Nommon Solutions and Technologies
Madrid, Spain

***Abstract*—Notices to Air Missions (NOTAMs) are key for the safe operation of commercial and civil aviation flights around the globe, as they provide up-to-date information on any disturbance on aerodromes and in the airspace. In this work, we propose a framework to train bidirectional encoder representations from transformers (BERT) to assign labels to NOTAM messages using a multi-label classification approach where each NOTAM can be assigned multiple labels. To deal with the scarcity of labeled data for this task, we propose a semi-supervised learning framework using the MixMatch algorithm to allow the leveraging of unlabeled NOTAMs, and reduce the need for expert labeled NOTAMs. We demonstrate that the MixMatch algorithm combined with focal loss improves the performance of BERT on the multi-label classification of NOTAMs, with the micro-averaged F1-score improving from 0.93 to 0.96 on a publicly available test dataset.**

***Keywords*—Notices to Air Missions, Large Language Models, BERT, Multi-label classification, Semi-supervised learning**

I. Introduction

Notices to Air Missions (NOTAMs) are documents delivered by aviation authorities, which aim to rapidly and concisely communicate updates concerning regulations or disruptions to pilots and other airspace users. Some examples of possible situations that can be reported in a NOTAM includes restrictions on runways and taxiways, changes in approach procedures, or new obstacles near an aerodrome. NOTAMs have adopted two standarized formats; one defined by International Aviation Organization (ICAO) and the other defined by FAA. Common to both formats there is a free text section where the issue reported is explained. This is written in capital letters and and frequently uses contractions and abbreviations to keep the text compact.

As the number of daily flights increases year by year, the number of NOTAMs emitted also increases [1]. Due to the large volume of NOTAMs that have to be processed currently, it is more likely that mistakes are made during flight planning, such as NOTAMs being overlooked or misunderstood. This has resulted in dangerous situations in the past that could have been avoided if no errors had been made during the processing of NOTAMs [2], [3]. Therefore, it is in the interest of aviation authorities and airspace users to develop tools that can process and analyze NOTAMs to contextualize their content and improve their clarity to prevent more of these occurrences.

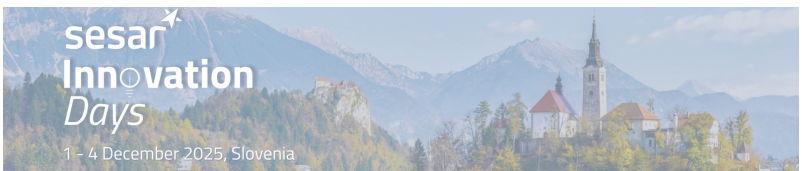
While most NOTAMs are relatively simple and could be handled by a single-label classification system with a large number of labels, there are cases where this approach falls short. Longer or more intricate NOTAMs often contain multiple pieces of information that cannot be adequately represented by a single label. To address this, we propose a multi-label classification approach that gives the model the flexibility to assign multiple tags to a single NOTAM. This design enables better handling of unorthodox or complex NOTAMs and provides richer contextualization, which can be leveraged for filtering and organizing NOTAMs through a tagging system. Such contextualization supports downstream applications such as flight planning, flight briefing, AIP automation, and coding of FMS data. Furthermore, allowing multiple tags mitigates the risk of misclassification caused by forcing the model to choose only one label when several are relevant. We explicitly validated this need with an end-user, Lufthansa Systems, during the design process, confirming that multi-label tagging aligns with operational requirements and offers better usability than single-label alternatives. The results were also presented to Eurocontrol experts, who provided very positive feedback on the methodology.

A. Related Work

The emergence of Large Language Models (LLMs) using the transformer neural network architecture has revolutionized Language Natural Processing (NLP) [4], increasing its effectiveness over a variety of different tasks. NOTAM contextualization is one of these tasks, as it has been an increasingly investigated topic in the literature in recent years.

The problem of NOTAM contextualization has been framed with a number of different approaches. Clarke et al. [5] compared the performance of various Large Language Models (LLMs) in question and answer tasks aimed at extracting relevant information from NOTAMs such as the affected region or the status of runways or taxiways.

Several authors have approached the problem as a classification task. Baigang Mi et al. [6] used a supervised training framework to train a wide array of deep learning models



for NOTAM classification, obtaining its best results with an attention-based Recurrent Neural Networks (RNN) architecture. Szeto et al. [7] also used an attention-based architecture, bi-directional encoder representations from transformers (BERT) [8], to classify NOTAMs using a simplified version of the FAA tagging system with 5 labels.

Recently, Rosa [9] and Mogillo-Dettwiller [10] classified NOTAMs in two categories "suppress" and "do not suppress". The aim is to filter out irrelevant NOTAMs to reduce the workload of flight operators. For this task, a traditional XGboost classifier using Word2vec for text vectorization was found to achieve higher accuracy compared to BERT. Rosa also substitutes the abbreviations in NOTAMs with their full definition, since there is no guarantee that these abbreviations would be present in the training data of BERT, thus making the performance comparison to the XGboost classifier more fair [9].

B. Contribution

To the best of our knowledge, the work available in the literature so far assigns a single label to each NOTAM. Instead, we propose to use a model for multi-label classification of NOTAMs. In a multi-label classification problem, a data point can be assigned several labels simultaneously, in contrast to standard classification models, which assign a single label to each data point. Having the flexibility to combine multiple labels should result in a more descriptive contextualization of the NOTAM. This also results in a more complex output space due to the increased number of possible combinations [11]. Due to this added complexity, we consider this application perfect for a LLM-based NLP algorithm, such as BERT. Its large number of trainable parameters combined with the generative transformer architecture should give it an edge over traditional classification models for this task, unlike single-label classification tasks where the advantage is not as clear [9].

A limitation in developing a solution to the multi-label annotation of NOTAMs is the lack of labeled data. Supervised learning techniques for large models, such as BERT, require large amounts of high-quality labeled data, which can be costly to acquire [12]. As a solution to this, we propose using semi-supervised learning to limit the amount of labeled data that is required to train the model. This Generative Artificial Intelligence (Gen-AI) technique uses a dataset that is only partially labeled. Thus, it alleviates the weaknesses of supervised and unsupervised learning, since a lesser amount of labeled data is required, while still retaining the domain knowledge embedded in the labeled data [13]. Hence, the contributions of this paper can be summarized as follows:

- We present a holistic methodology for the multi-label classification of NOTAMs, starting from the processing of the abbreviated text and ending with the annotation of each NOTAM with multiple labels from a predetermined set of possible tags.
- The semi-supervised multi-label classification method based on MixMatch data enhancement [14] is used to overcome limitations on the quantity of available labeled training data for NOTAM classification.
- We implement a first version of the proposed pipeline and demonstrate the improvement it provides in BERT's performance for the multi-label classification of NOTAMs. For this purpose, we use a publicly accessible dataset, part of which was manually labeled by the authors.

II. Method

A. Model

Multi-label text classification (MLTC) is a challenging task in natural language processing, where each input sample may be associated with multiple labels from a predefined set. Traditional supervised approaches often struggle when labeled data is scarce or when the label distribution is highly imbalanced. To address these limitations, we adopt a semi-supervised learning strategy based on the MixMatch algorithm and focal loss, which has shown promising results in combining labeled and unlabeled data to improve generalization.

The proposed model tackles the multi-label classification problem using a semi-supervised learning pipeline that integrates BERT-based embeddings, MixMatchNL data augmentation and focal loss optimization.

The process is illustrated in Figure 1 and consists of four main stages:

- **Preprocessing and Embedding:** Raw NOTAMs are cleaned and tokenized. Tokenization refers to splitting the text into smaller units (tokens), such as words or subwords, which are compatible with BERT's vocabulary. These tokens are then embedded using a BERT encoder, which provides contextualized representations for downstream classification.
- **Supervised Pretraining:** the BERT model is initially fine-tuned on labeled data to learn task-specific features. This step establishes a strong baseline and prepares the model for semi-supervised enhancement.
- **Semi-Supervised Data Enhancement with MixMatchNL**: the MixMatchNL model is applied to incorporate unlabeled data into training. It generates pseudo-

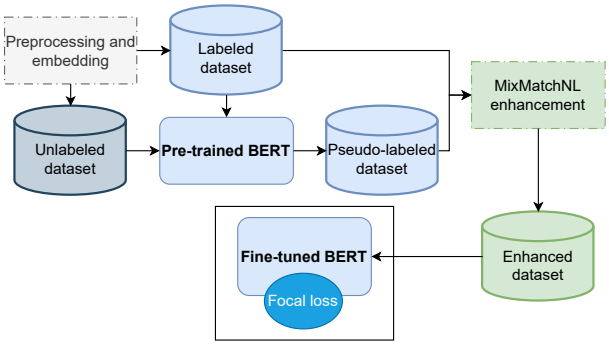


Figure 1. The architecture of the semi-supervised multi-label NOTAM classification model.

labels and mixes labeled and unlabeled samples to improve generalization and robustness.
- **Fine-Tuning with Focal Loss:** Finally, the Pretrained BERT model is fine-tuned using Focal Loss to address class imbalance and improve performance on underrepresented labels.

1) Semi-supervised Data Enhancement with MixMatchNL: To exploit the large volume of unlabeled NOTAMs, we integrate the MixMatch algorithm into our training pipeline. MixMatch [14] is a semi-supervised learning technique that combines pseudo-labeling, data augmentation, and consistency regularization to improve model generalization. It is particularly effective in low-resource settings where labeled data is scarce. Although unlabeled data does not explicitly contain label information, it provides valuable structural information about the input space. As shown by Zou and Wang [12], semi-supervised approaches leverage assumptions such as cluster consistency and low-density separation, which posit that samples close in feature space tend to share labels and that decision boundaries should lie in regions of low data density. Incorporating unlabeled NOTAMs helps the model learn this underlying distribution, improving representation learning and reducing overfitting to the limited labeled set. Furthermore, pseudo-labeling expands the effective training set, while consistency regularization enforces stable predictions under perturbations, both of which enhance generalization.

Our implementation is inspired by Zou and Wang [12], who demonstrated the effectiveness of BERT-based models in multi-label classification. We adapt MixMatch to our multi-label setting as follows.

The process begins by generating pseudo-labels for the unlabeled NOTAMs. For each unlabeled input u_j, the current model produces a probability vector $\hat{y}_j = f_\theta(u_j)$, where each component $\hat{y}_{jk} \in [0,1]$ represents the predicted likelihood of label k being present. These soft predictions are then sharpened using temperature scaling to reduce entropy and increase confidence. This is, for each unlabeled input u_j, the sharpened pseudo-label $\tilde{y}_j$ is computed as:

$$\tilde{y}_j = \text{Sharpen}(\hat{y}_j, T) = \frac{\hat{y}_j^{1/T}}{\hat{y}_j^{1/T} + (1-\hat{y}_j)^{1/T}},$$

where $\hat{y}_j$ is the model's soft prediction and $T < 1$ is the temperature parameter applied element-wise.

Next, we extract the BERT embeddings for both labeled and unlabeled inputs. Let x_i be a labeled input with label y_i, and u_j an unlabeled input with pseudo-label $\tilde{y}_j$. Their embeddings are obtained from the BERT encoder:

$$\text{emb}_i = \text{BERTEmb}(x_i), \quad \text{emb}_j = \text{BERTEmb}(u_j).$$

We then apply MixUp to pairs of labeled and unlabeled embeddings. Given two embedding-label pairs (emb_i, y_i) and (emb_j, y_j), the mixed sample is computed as:

$$\tilde{\text{emb}} = \lambda \text{emb}_i + (1-\lambda)\text{emb}_j, \quad \tilde{y} = \lambda y_i + (1-\lambda) y_j,$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ and α controls the degree of interpolation. The choice of α influences the shape of the distribution:

- When $\alpha < 1$, the distribution is U-shaped, favoring values close to 0 or 1, which results in more extreme mixing.
- When $\alpha = 1$, the distribution is uniform, allowing all mixing ratios equally.
- When $\alpha > 1$, the distribution is peaked around 0.5, promoting a balanced mixing between samples.

This parameter plays a critical role in controlling the diversity and smoothness of the synthetic examples used during training.

This process yields a stream of synthetic training examples in the form of mixed embeddings and interpolated labels. These are used to train the model in a semi-supervised fashion, encouraging it to learn smoother decision boundaries and better utilize the structure of the unlabeled data. The final training phase, described in Section II-A2, applies focal loss to address class imbalance.

2) Fine-Tuning with Focal Loss: After the data enhancement phase, the model is fine-tuned using a supervised learning strategy tailored to the characteristics of the multi-label classification task. In particular, we address the issue of class imbalance by replacing the standard binary cross-entropy loss with focal loss.

Focal loss modifies the standard binary cross-entropy loss by introducing a modulating factor that reduces the contribution of well-classified examples and focuses learning on harder, misclassified instances. This is especially beneficial in multi-label settings where certain labels are significantly underrepresented.

For each label $k \in \{1, \ldots, K\}$ and sample i, the focal loss is defined as:

$$\begin{aligned}\text{FL}(p_{ik}, y_{ik}) = &-\alpha_k (1-p_{ik})^\gamma y_{ik} \log(p_{ik}) \\ &- (1-\alpha_k) p_{ik}^\gamma (1-y_{ik}) \log(1-p_{ik}),\end{aligned}$$

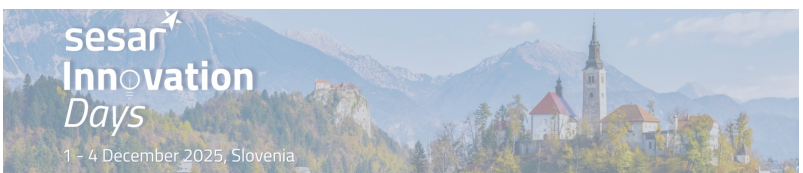
where:

- $p_{ik} = f_\theta(x_i)_k$ is the predicted probability for label k,
- $y_{ik} \in \{0,1\}$ is the ground truth for label k,
- γ is the focusing parameter (typically $\gamma = 2$),
- α_k is a balancing factor for label k, proportional to class frequency.

The total loss is computed by summing over all labels and samples:

$$\mathcal{L}_{\text{focal}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \text{FL}(p_{ik}, y_{ik}).$$

This formulation ensures that the model pays more attention to underrepresented and difficult labels, improving recall and



overall performance in imbalanced multi-label settings. By applying focal loss during the final fine-tuning stage, the model is able to refine its predictions and achieve better generalization across all classes. This final training step complements the MixMatch enhancement by ensuring that the model not only generalizes well but also handles label imbalance effectively.

B. Evaluation criteria

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 behavior. These metrics are computed using the predicted label sets and the associated probability scores.

1) Label-Based Metrics: We report precision, recall, and F1-score using three averaging strategies commonly used in multi-label evaluation:

- **Micro-average**: Aggregates the contributions of all labels to compute the average metric. It is sensitive to the performance on frequent labels.
- **Macro-average**: Computes the metric independently for each label and then takes the average. It treats all labels equally, regardless of their frequency.
- **Weighted-average**: Similar to macro-average but weights each label by its support (i.e., the number of true instances), balancing the influence of frequent and rare labels.

2) Instance-Based and Ranking Metrics: To further evaluate the model's behavior at the instance level and its ability to rank relevant labels, we include the following metrics:

- **Hamming Loss**: Measures the fraction of incorrect labels to the total number of labels.

$$\text{Hamming Loss} = \frac{1}{N \cdot K} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}[y_{ik} \neq \hat{y}_{ik}].$$

- **Subset Accuracy**: Computes the proportion of samples for which the predicted label set exactly matches the true label set.
- **Jaccard Index (samples)**: Measures the similarity between predicted and true label sets for each instance.

$$\text{Jaccard}_i = \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}.$$

- **Coverage Error**: Indicates how far we need to go down the ranked list of labels to cover all true labels.
- **Label Ranking Average Precision (LRAP)**: Evaluates how well the model ranks relevant labels higher than irrelevant ones.

Together, these metrics provide a robust and multi-faceted evaluation of the model's performance, accounting for both classification accuracy and ranking quality across diverse label distributions.

III. Experiments

A. NOTAM Datasets and labeling system

The NOTAM datasets used for this work are publicly available in repositories of previous projects on similar topics. We included a dataset that uses the Federal Aviation Administration (FAA) format [7] and a dataset that uses the ICAO format [15]. For the tagging task, we have decided to use only the NOTAM free text field, which in ICAO format belongs in section E) [9]. In total, we collected a dataset of 107371 NOTAMs. To minimize the number of words that BERT has not seen in its pre-training, we expand all the annotations present in the text, including those for airports. This was done using the regular expression library in Python [16] in an iterative manner, since the definition of certain abbreviations can contain another abbreviation. The most frequent words in the dataset are plotted in the bar chart in Figure 2.

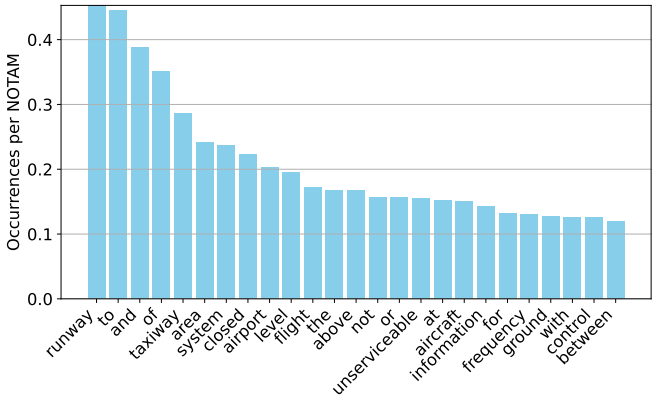


Figure 2. The 25 words that appeared most frequently on the NOTAM dataset after expanding all abbreviations and contractions.

From this dataset, 8,497 NOTAMs were manually labeled by the authors of this work, following a first iteration of the labeling system for the multi-label NOTAM classification. Of these, 1,700 were reserved for testing. The labeling system consists of 11 labels. This number aims to achieve a good balance between expressability and complexity, keeping in mind the volume of data available. The chosen labels are based on the 50 labels that OPSGROUP proposed in their NOTAM sprint [17] and the FAA's labeling system [18]. However, rather than having a lot of mutually exclusive labels, we opt to have a lower amount of non-mutually exclusive labels. In the list below, these labels are presented along with what they convey about a NOTAM.

- Runway: NOTAMs directly concerning a runway.
- Taxiway: NOTAMs directly concerning a taxiway.
- Aerodrome: NOTAMs concerning airports or terminal maneuvering areas, excluding those concerning a specific runway or taxiway.
- Obstacle: NOTAMs concerning obstacles in the vicinity of aerodromes or relating to minimum obstacle avoidance altitudes.

- Lighting-issues: NOTAMs concerning issues on lights that communicate information to pilots, such as those on obstacles or runways.
- Hazards: NOTAMs concerning potentially dangerous occurrences in the airspace or aerodromes that do not qualify as obstacles (wildlife, drones, airdrops, fireworks…).
- Navigation aid: NOTAMs concerning or referring to navigation aids, including radio, visual, and electronic aids.
- Procedure: NOTAMs giving instructions to pilots on how to overcome an unexpected occurrence.
- Airspace: NOTAMs concerning unexpected events in the airspace, such as failures in navigation aids or other disruptions.
- Closure: The complete absence of a service or infrastructure (runway, taxiway, aerodrome, etc.).
- Restriction: The partial and/or conditional closure of a service, infrastructure, airspace, or air route. Can also be used in the context of a procedure.

The frequency of the labels in the dataset is not balanced, as can be observed in Figure 3, with *restriction* being the most common label and *obstacle* being the least common.

Furthermore, certain combinations of labels are much more frequent than others. This is shown graphically in Figure 4

Observing Figure 4, a pattern can be noticed where labels that are related to the ”subject” of the NOTAM are often paired with labels that concern the ”description” of the state of that subject. For instance, the most frequent pairing for ”Airspace” is ”Restriction”, and the most frequent pairing for ”Taxiway” is ”Closure”. It is also interesting to note that ”Restriction” is also paired with ”Closure” quite frequently. The reason for this is that this pairing was used for situations in which the ”subject” infrastructure was only closed to some airspace users. For example, a taxiway closed to aircraft with wingspan greater than 30 meters. An exception for this is the ”Procedure” label, which is most often paired with ”Restriction”, a label used also for description. Generally, what can be observed is that many combinations are extremely rare, making the output space unnecessarily large, and therefore making model training more challenging. Based on this analysis, some modifications for future iterations of the labeling system are described in Section IV. The entire NOTAM dataset and the results of the paper are available in Zenodo (https://doi.org/10.5281/zenodo.17208970).

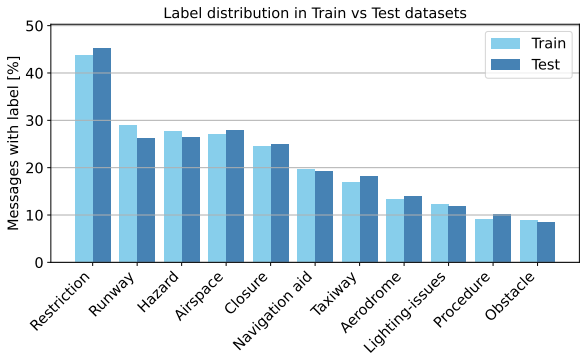


Figure 3. Label frequencies for each class on the labeled NOTAMs.

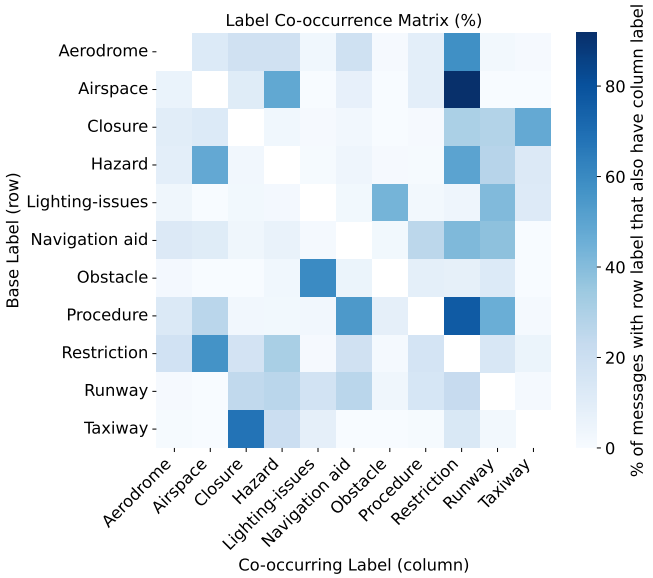


Figure 4. Label co-occurrence matrix on the combined test and train datasets. The numbers shown are the percentage of the labeled NOTAMs with the row label that also contain the column label.

B. Model training

The training process consists of two main phases: supervised fine-tuning of a BERT model on labeled data, followed by semi-supervised training using MixMatch and focal loss. Both phases are designed to address the challenges of multi-label classification, including limited labeled data and class imbalance.

1) Supervised Training on Labeled Data: In the first phase, we fine-tune a BERT model using the labeled subset of the NOTAM dataset.The model parameters are initialized from the pre-trained BERT base (uncased) checkpoint, and adapted for multi-label classification by configuring the output layer to match the number of target labels.

The training set consists of 6,797 labeled samples, while the test set includes 1,700 samples. The model is trained using the binary cross-entropy loss applied to the logits produced by the final layer. Optimization is performed using the AdamW optimizer with a learning rate of 10^{-5} and a batch size of 4. The training is conducted for 8 epochs. A batch size of 8 is commonly used in BERT fine-tuning when computational resources are limited or when working with moderately sized datasets. Smaller batch sizes tend to produce noisier gradient estimates, which can help the model escape sharp local minima and improve generalization. Additionally, this size fits well within the memory constraints of typical GPU setups used in academic and applied research.The learning rate of 3×10^{-5} is one of the most widely recommended values for BERT fine-tuning across various NLP tasks [19]. It provides a good balance between convergence speed and training stability.

During training, the model receives tokenized inputs and learns to predict the presence or absence of each label independently.

2) Semi-Supervised Training with MixMatch and Focal Loss: In the second phase, we enhance the model using a semi-supervised strategy that combines MixMatch and focal loss. This phase leverages both the labeled dataset and an additional set of 98,750 unlabeled NOTAMs.

The MixMatch procedure operates in batches of size 8. For each batch, we extract BERT embeddings from both labeled and unlabeled inputs. Pseudo-labels for the unlabeled data are generated using the current model and sharpened using temperature scaling. These embeddings and labels are then mixed using the MixUp algorithm explained in II-A1. In our experiments, the parameter γ is set to an empirical value of 2 and the MixUp interpolation parameter to $\alpha = 0.75$. This value favors moderately extreme mixing, allowing synthetic examples to lean more toward one of the original samples while still maintaining a degree of balance. This choice was made to encourage diversity in the training data without introducing excessive noise, and it aligns with common practice in semi-supervised learning setups.

The resulting mixed embeddings and labels are used to train the model using focal loss, which emphasizes learning from hard-to-classify and minority-label instances. The model is trained for 8 epochs with a learning rate of 10^{-5} and gradient clipping applied to stabilize training.

This semi-supervised phase allows the model to benefit from the structure of the unlabeled data while addressing the imbalance in label distribution, resulting in improved performance across all classes.

C. Model performance / Results

To evaluate the effectiveness of our multi label classification strategy, we compare the performance of the pretrained BERT model (fine-tuned only on labeled data) with the retrained BERT model enhanced through MixMatch and focal loss. Both models are evaluated on the same test set consisting of 1,700 labeled NOTAMs.

The comparison is based on the evaluation metrics defined in Section II-B.

The results presented in Table I demonstrate a clear performance improvement achieved through the integration of semi-supervised learning via MixMatchNL and loss reweighting using focal loss. Across all classification metrics—micro, macro, and weighted averages—the retrained model consistently outperforms the baseline pretrained BERT. Notably, the micro-averaged F1-score increased from 0.9306 to 0.9624, while the macro-averaged F1-score rose from 0.9282 to 0.9617, indicating enhanced performance both globally and across individual classes, including those with fewer samples. Weighted averages also reflect this trend, confirming improved handling of class imbalance. Furthermore, instance-based metrics reveal substantial gains: Hamming loss was reduced by nearly half (from 0.0294 to 0.0159), and subset accuracy improved from 0.7812 to 0.8753, suggesting more precise multi-label predictions. The Jaccard index increased from 0.9010 to 0.9452, and coverage error decreased from 3.5453 to 3.0241, indicating better label ranking and reduced prediction uncertainty. Finally, the label ranking average precision (LRAP) rose from 0.9122 to 0.9515, further confirming the model's enhanced ability to prioritize relevant labels.

TABLE I. COMPARISON OF EVALUATION METRICS BETWEEN THE PRETRAINED BERT MODEL AND THE RETRAINED MODEL WITH MIXMATCH AND FOCAL LOSS ON THE TEST SET.

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

Table II presents a detailed comparison of precision, recall, and F1-score for each label between the pretrained BERT model and the retrained model enhanced with MixMatchNL and focal loss. The retrained model demonstrates consistent improvements across nearly all labels. For instance, Aerodrome shows notable gains in all three metrics, with F1-score increasing from 0.88 to 0.94. To better illustrate the performance improvements across individual labels, Figure 5 presents a comparison of per-label F1-scores between the pretrained BERT model and the retrained model.

TABLE II. PER-LABEL COMPARISON OF PRECISION, RECALL, AND F1-SCORE BETWEEN THE PRETRAINED BERT MODEL AND THE RETRAINED MODEL.

Label	BERT - Pretrained			BERT - MixMatchNL - Focal Loss		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Aerodrome	0.89	0.86	0.88	0.97	0.91	0.94
Airspace	0.96	0.95	0.95	0.96	0.99	0.97
Closure	0.94	0.90	0.92	0.99	0.94	0.96
Hazard	0.91	0.93	0.92	0.96	0.96	0.96
Lighting-issues	0.97	0.93	0.95	0.98	0.95	0.96
Navigationaid	0.92	0.92	0.92	0.96	0.92	0.94
Obstacle	0.99	0.93	0.96	0.99	1.00	1.00
Procedure	0.86	0.86	0.86	0.88	0.94	0.91
Restriction	0.90	0.93	0.92	0.94	0.96	0.95
Runway	0.97	0.97	0.97	0.99	0.99	0.99
Taxiway	0.95	0.98	0.97	0.99	1.00	1.00

Similarly, Closure, Obstacle, and Taxiway exhibit substantial improvements, with Obstacle and Taxiway achieving perfect F1-scores of 1.00. Labels such as Lighting-issues,

Navigation aid, and Runway also benefit from enhanced precision and recall, leading to higher F1-scores. Even for more challenging categories like ”Procedure” and ”Restriction”, the retrained model achieves better balance between precision and recall, resulting in improved overall performance. Notably, ”Procedure” is the least predicted label in the retrained model, with the lowest F1-score of 0.91. While this still reflects strong performance, it is comparatively lower than other labels and may indicate challenges in consistently predicting this category, possibly due to greater variability in its features or fewer representative samples in the training data.

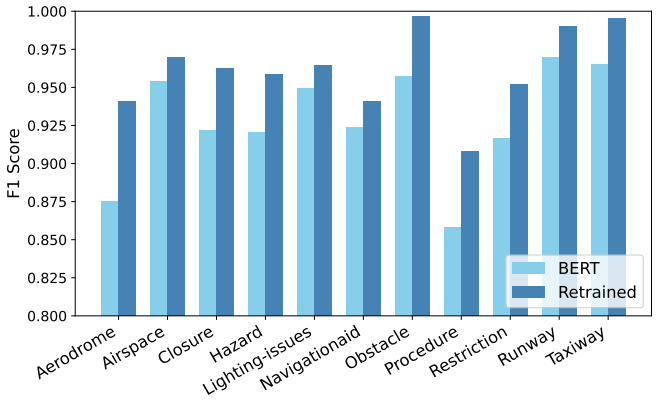


Figure 5. Comparison of per-label F1-scores on the test dataset between the pretrained BERT model and the retrained model using MixMatchNL and focal loss.

Finally, Figure 6 plots the F1-score per label for different NOTAM lengths, measured in words. All the bars with a representative sample size (greater than 5 NOTAMs) have F1-scores above 0.75, with the majority exceeding 0.9 for the retrained model. The BERT model without MixMatch generally obtains lower scores in all slices of the data. This is most accentuated in longer NOTAMs, as can be observed on the bottom right graph, which plots the micro-averaged F1-score. While for short NOTAMs the MixMatch retraining with focal loss provides a minimal advantage, the difference increases progressively with NOTAM length. For NOTAMs with over 20 words the micro-averaged F1-score improves from 0.91 to 0.96. Nonetheless, the retrained model results still show a worse performance for longer messages.

IV. Conclusion

This work introduces a semi-supervised multi-label approach to annotate NOTAMs. For this purpose, we propose a labeling system that we use to partially label a NOTAM dataset. We use the pre-trained LLM BERT as a baseline, which is then specialized on the multi-label classification task through supervised learning using the labeled fraction of the NOTAM dataset as training data. This model is then used to predict labels for the unlabeled NOTAMs, thus producing pseudo-labeled NOTAMs. After that, the model is retrained using the MixMatch algorithm, where the BERT embeddings and the pseudo labels of the unlabeled NOTAMs are mixed with the embeddings and labels of the labeled NOTAMs, thus producing a set of synthetic embedded NOTAMs and labels. The model is then retrained with these data and focal loss to alleviate the issues caused by underrepresented labels.

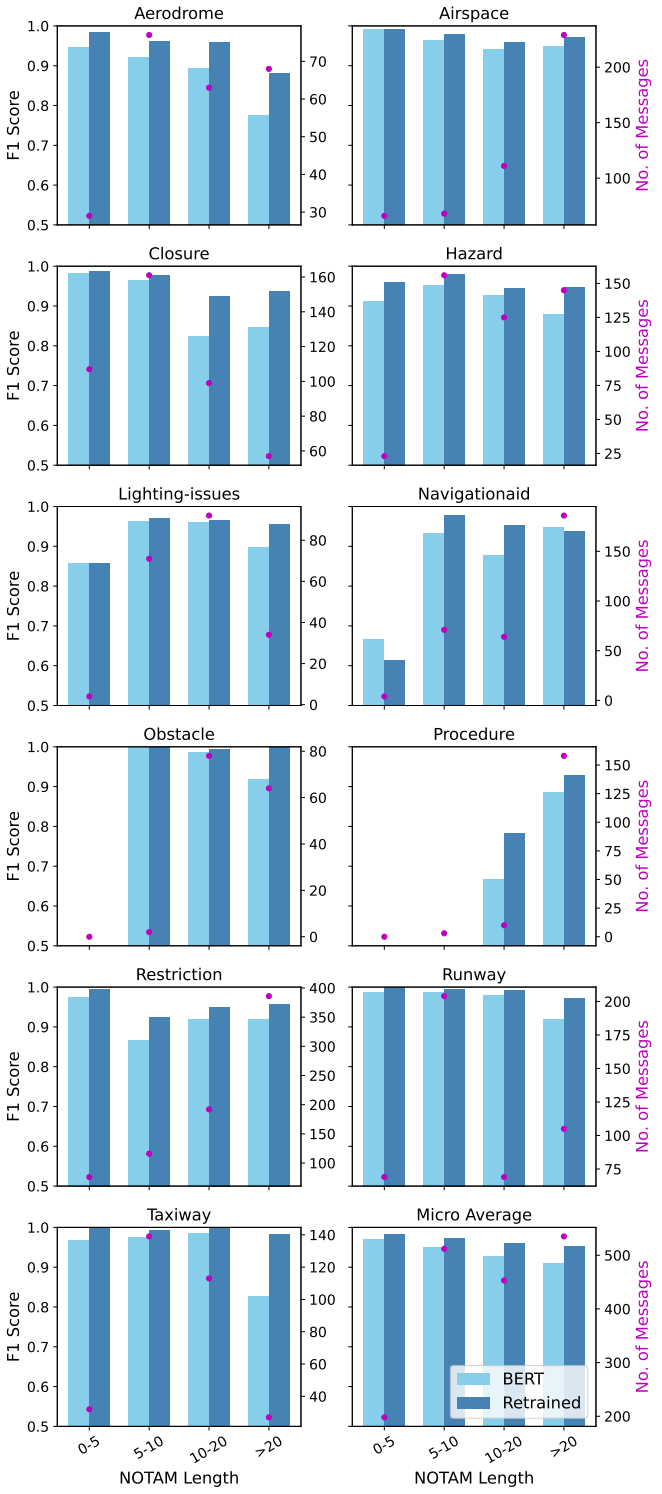


Figure 6. Per-label F1-scores for the BERT model and the retrained model using MixMatchNL and focal loss on the test dataset for different NOTAM lengths.

The results in the test dataset show an improvement in the accuracy metrics of the model when retraining the model with MixMatch and focal loss, with an improvement of the micro F1-score from 0.93 to 0.96. This is also reflected in other metrics for multi-label classification such as the Hamming loss, which is reduced from 0.0294 to 0.0159. The label with the worst F1-score is "procedure", which is most probably due to its definition being less concrete than for the other labels and a lower representation in the data. Despite this, the F1-score on this label also improves when using MixMatch from 0.86 to 0.91. As could be expected, the model performs better in shorter NOTAMs (0 to 5 words) but its accuracy does not reduce dramatically as NOTAM length increases, since its F1-score is still approximately 0.96 when tested on NOTAMs with over 20 words, thus showing that the model is still able to extract the relevant information to select the correct labels even in longer texts.

Based on our experiments, we recommend using a BERT-based architecture initialized from the `bert-base-uncased` checkpoint for multi-label classification of NOTAMs. Fine-tuning with binary cross-entropy loss and a batch size of 4–8 provides stable training dynamics, especially under limited computational resources. A learning rate of 10^{-5} offers a good balance between convergence and generalization. For improved performance, we suggest incorporating a semi-supervised training phase using MixMatch and focal loss. This approach effectively leverages unlabeled data while addressing class imbalance. Specifically, using temperature-scaled pseudo-labels and MixUp interpolation with $\alpha = 0.25$ encourages diversity without introducing excessive noise. Focal loss further enhances learning from minority and hard-to-classify labels. Overall, this two-phase training strategy significantly improves classification metrics and is particularly beneficial for longer and more complex NOTAMs.

For future work, we recommend iterating on the labeling system. As noted in Section III-A, the labels used in this work entail information of diverse nature regarding the NOTAM, some inform about the affected infrastructures (runway, taxiway), while others describe the issue (Hazard, Restriction,...). Hence, although the model obtained good results with the current system, as the number of labels increases to improve the descriptivity of the solution this could become an issue as the output could become too chaotic and unstructured. To overcome this, future work will focus on incorporating hierarchical text classification approaches [20], [21] by grouping the labels in subsets, with a multi-label or single-label classification problem being solved for each of these subsets. The model could then be constrained to limit the output space to coherent combinations of labels, facilitating both the model training process, and the labeling process for the experts.

While the proposed semi-supervised BERT-based model demonstrates strong performance in multi-label classification of NOTAMs, several directions can be explored to further enhance its capabilities. Future work could investigate the use of more advanced transformer architectures such as RoBERTa or Longformer, which may better capture long-range dependencies in longer NOTAMs. Additionally, experimenting with dynamic or adaptive interpolation strategies for MixUp, rather than a fixed α, could improve the integration of labeled and pseudo-labeled data. Incorporating domain-specific pretraining or fine-tuning on aviation-related corpora may also yield more contextualized embeddings. Finally, exploring active learning strategies to selectively label the most informative unlabeled samples could reduce annotation costs while improving model performance.

ACKNOWLEDGMENT

This research is part of the "GenAI Models for ATM" project which has received funding from the SESAR 3 Joint Undertaking (SESAR 3 JU) as part of the SESAR3 ENGAGE2 KTN 1st Call for Catalyst Funding, under grant agreement No. 101114648. The JU receives support from the European Union's Horizon Europe research and innovation program and the SESAR 3 JU members other than the Union. The paper reflects only the view of the authors and that SESAR 3 JU is not responsible for any use that may be made of the information it contains.

The authors also thank Afonso Santinho Faisca, from Lufthansa Systems, for the feedback he gave us regarding the MixMatch BERT classification model and the labeling system.

REFERENCES

[1] International Civil Aviation Organization, "Global campaign on notam improvement (notam2021)," Jun. 2021, accessed: 2024-07-05. [Online]. Available: https://www.icao.int/airnavigation/information-management/Pages/GlobalNOTAMcampaign.aspx

[2] National Transportation Safety Board, "Taxiway overflight air canada flight 759 airbus a320-211 – incident report," National Transportation Safety Board, Tech. Rep. NTSB/AIR-18/01 PB2018-101561, Jul. 2017. [Online]. Available: https://www.ntsb.gov/investigations/accidentreports/reports/air1801.pdf

[3] OPSGROUP, "Bad notams = runway overruns in hamburg," May 2018, accessed: 2024-07-05. [Online]. Available: https://ops.group/blog/bad-notams-runway-overruns-in-hamburg/

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] S. S. Clarke, P. Maynard, J. A. Almeida, S. G. Kumar, S. Rajkumar, A. C. Kemp, and R. Pai, "Natural language processing analysis of notices to airmen for air traffic management optimization," in *AIAA Aviation Forum*, 2021.

[6] B. Mi, Y. Fan, and Y. Sun, "Notam text analysis and classification based on attention mechanism," in *Journal of Physics: Conference Series*, vol. 2171, no. 1. IOP Publishing, 2022, p. 012042.

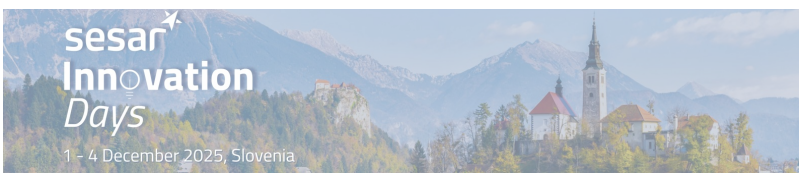
[7] A. Szeto and A. N. Das, "Classification of notices to airmen using natural language processing," in *AIAA SCITECH 2024 Forum*, 2024, p. 2585.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[9] I. Rosa, "Leveraging machine learning and language models to assess the importance of notices to air missions (notams)," Master's thesis, NTNU, 2024.

[10] A. Mogiłło-Dettwiler, "Filtering and sorting of notices to air missions (notams)," Master's thesis, University of Zurich, 2024.

[11] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *international conference on machine learning*. PMLR, 2017, pp. 3780–3788.



[12] H. Zou and Z. Wang, "A semi-supervised short text sentiment classification method based on improved bert model from unlabelled data," *Journal of Big Data*, vol. 10, no. 1, p. 35, 2023.
[13] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," *Artificial intelligence review*, vol. 56, no. 9, pp. 9401–9469, 2023.
[14] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.
[15] J. Coupon, "Notam-classifier (capstone project)," https://github.com/jcoupon/NOTAM-classifier, 2020, accessed: 2025-08-29.
[16] G. Van Rossum, *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.
[17] OPS Group, "Ops group blog," https://ops.group/blog/, 2024, accessed: 2025-05-30.
[18] Federal Aviation Administration, "Icao notam format example," https://www.faa.gov/air_traffic/flight_info/aeronav/notams/media/ICAO_NOTAM_Format_Example.pdf, 2021, accessed: 2025-08-29.
[19] C. McCormick and N. Ryan, "Bert fine-tuning tutorial with pytorch," 2019. [Online]. Available: https://mccormickml.com/2019/07/22/BERT-fine-tuning/
[20] C. N. Silla Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data mining and knowledge discovery*, vol. 22, no. 1, pp. 31–72, 2011.
[21] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.