The Effectiveness of Large Language Models for Textual Analysis in Air Transportation

G. Jarry[a,*], R. Dalmau[b] and P. Very[a]

[a]Aviation Sustainability Unit (ASU) & [b]Engineering Centre of Expertise (ENG)
EUROCONTROL, Brétigny-Sur-Orge, France
{gabriel.jarry,ramon.dalmau-codina,philippe.very}@eurocontrol.int
[*] Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 15, 2024

Keywords: Large language models; air transportation; adverse weather.

1 INTRODUCTION

Large language models (LLMs), such as GPT-3, BERT and T5, have transformed the fields of natural language processing (NLP) and artificial intelligence (AI) since their inception in the mid-2010s. Trained on large amounts of text data, these models excel at producing text that is contextually relevant and grammatically correct, just as humans do.

Their capabilities include answering questions, writing essays, summarising and categorising text, translating languages, and even producing creative content such as poetry and code. The ability of LLMs to understand and create text represents a significant advance over their predecessors, transforming the ability to understand and generate human language.

As a matter of fact, the aviation industry generates a large amount of textual data, such as notices to airmen (NOTAMs), customer feedback, transcriptions of air traffic control (ATC) communications, and incident reports. LLMs can be used to analyse this data, providing comprehensive insights that can improve operational efficiency, safety, and passenger experiences.

This paper focuses on a particular kind of textual data: comments made by flow managers during the implementation of air traffic flow management (ATFM) regulations.

For instance, in a recent regulation applied at Frankfurt Airport, the flow manager that activated the measure included the following remark: *RWY North1 temporary blocked*. The aim of this paper is to examine these remarks and classify them into clusters that encapsulate the most prevalent reasons for ATFM regulations, making use of the capabilities of LLMs. Specifically, we concentrate on regulations caused by weather and which reference location was an aerodrome. This focus is not arbitrary, but rather to facilitate a comparison with the results of Dalmau *et al.* (2023), who used classical machine learning methods to identify the reason behind observed airborne holdings. Such a comparison allows for the evaluation of LLMs' effectiveness in a practical task against a stablished baseline.

2 METHODOLOGY

This section describes the dataset and the first part of the process shown in Fig. 1, which consists of using LLMs to cluster ATFM regulations, specifically those related to weather, based on the textual comments provided by flow managers.

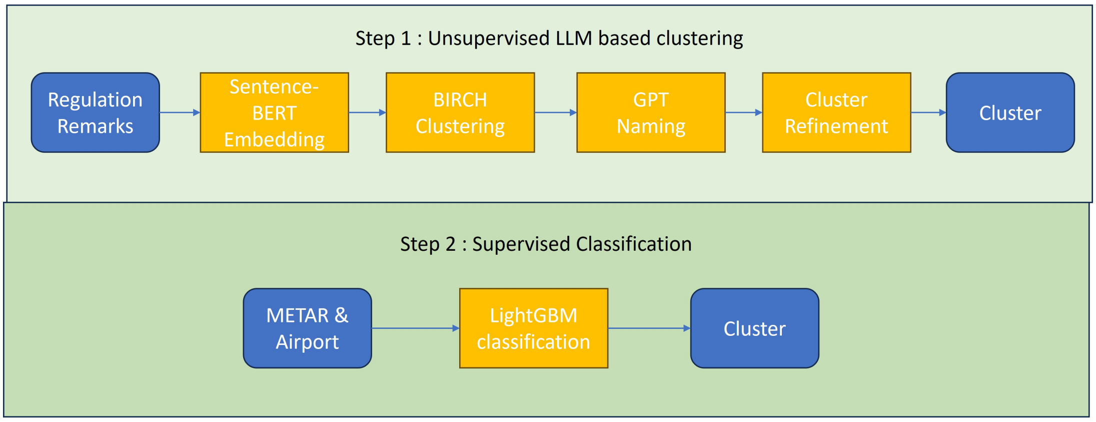

Figure 1 – *Methodology followed to cluster and model weather-related ATFM regulations due to weather from textual comments and weather observations.*

2.1 Dataset

The dataset comprises ATFM regulations for the 45 busiest airports in Europe in 2022, as listed by Wikipedia, with data recorded at 30-minute intervals. Each observation in the dataset, which corresponds to a specific ATFM regulation during a 30-minute interval, is supplemented with weather information from the nearest meteorological aerodrome report (METAR). Additionally, each observation may include textual comments from the flow manager who activated the regulation. These comments provide valuable context for each regulation, despite their inconsistent writing style and use of abbreviations. In Table 1, a subset of these comments is showcased, emphasising their variety and lack of uniformity in the format. In total, the dataset contains 112K weather-related regulations, of which 27K include textual comments.

Table 1 – *Examples of weather regulation remarks*

Textual Description
CBs NOTAM
single RWY OPS due to wind direction
LVP forecast
A/D CLOSED DUE TO STORM DAMAGES
Wind and Rain / after 1230 Aerodrome capacity: Single RWY
WIND DIRECTION
snow clearance NOTAM
Fog (LVP)/ at 10H00 Aerodrome Capacity
LOW VIS AS FROM 2000 AD CAPACITY

2.2 Clustering methodology

The clustering methodology consists of four meticulously scheduled steps with the goal of extracting meaningful patterns from the textual comments of thousands of weather-related ATFM regulations with minimal human intervention.

1. **Sentence embedding generation:** using Sentence-BERT (Reimers & Gurevych, 2019), a tailored variant of the BERT architecture optimised for the generation of semantically rich sentence embeddings, we transformed the textual explanations into a high-dimensional space (i.e., a large vector of numerical values, here 756 dimensions). This step facilitates the comparison of semantic similarities between explanations in the high-dimensional space, thus setting the stage for effective clustering.

2. **First clustering:** we used the Birch algorithm (Zhang *et al.*, 1997) to group the sentence embeddings into initial clusters. The Birch algorithm was chosen for its efficiency in handling large datasets and its ability to produce a manageable number of clusters without sacrificing granularity. This process resulted in 23 clusters. At this point in the process, each cluster was assigned a numerical identifier, but its meaning remained a mystery.

3. **Cluster naming:** to assign meaningful names to the clusters, we used an automated process involving ChatGPT. That is, ChatGPT was given a variety of examples from each cluster and asked to select the most representative name. This approach ensured that the name of each cluster accurately reflected the common thematic elements of its constituent regulations, thereby enhancing interpretability.

4. **Cluster refinement:** we reviewed the automatically generated clusters to ensure coherence and relevance. This included merging overlapping clusters and fine-tuning cluster definitions to better capture the unique characteristics of weather-related disruptions.

3 Supervised learning

The second part of the process shown in Fig. 1 consists of modelling the relationship between the observed weather conditions and the distinct weather-related clusters found during the first step of the process using supervised learning.

The observed weather conditions include numerical features like wind speed, visibility, and ceiling, as well as boolean flags that indicate the presence of specific events like precipitation, thunderstorms, and fog.

The dataset was split with 80% of the observations for training and 20% for testing. Before splitting, the dataset was arranged in chronological order to prevent data leakage. Furthermore, the training set was used for a comprehensive model and hyper-parameter optimisation with 5-fold cross-validation. A variety of classification models were tested, and LightGBM emerged as the best in terms of area under the receiver operating characteristic curve (AUC).

4 RESULTS

Applying the methodology outlined in the preceding section to our dataset yielded the cluster listed in Table 2. These clusters effectively encapsulate the diverse weather conditions influencing airport operations.

Table 2 – *Refined Clusters obtained from weather regulations*

Cluster Id	Cluster description	Samples #
1	Snow and runway conditions	4 450
2	Low visibility and ceiling conditions	7 317
3	Wind	4 008
4	Cumulonimbus and thunderstorms activity	4 888
5	Fog and related conditions	2 319
6	Thunderstorms and adverse weather conditions	1 524
7	Cumulonimbus	2 443

After training the classifier on these clusters, the model's performance on the test set was then assessed using a confusion matrix, which demonstrated its ability to accurately distinguish between different clusters.

The confusion matrix shown in Table 3 reveals interesting patterns and challenges in predicting different weather clusters.

Table 3 – *Confusion Matrix for Weather cluster prediction*

Pred	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
True	%	%	%	%	%	%	%	#
1	92.7	1.0	1.4	1.0	1.6	1.7	0.6	874
2	1.7	70.4	1.6	1.7	15.6	1.3	7.8	1496
3	2.1	0.7	85.9	4.3	0.5	2.4	4.0	1026
4	1.4	1.9	5.5	56.9	2.0	18.1	14.1	817
5	1.3	18.0	0.4	1.5	76.0	1.3	1.5	471
6	2.3	0.6	7.0	23.8	0.6	61.0	4.7	341
7	1.0	8.4	4.7	18.7	1.4	10.7	54.9	486
Total #	885	1216	843	833	639	506	589	5511

Notably, the model accurately predicts ‘Snow and Runway Conditions’ and ‘Wind’ conditions, with success rates of 92.7% and 85.9%, respectively. It also performs well for ‘Low Visibility and Ceiling Conditions’ and ‘Cumulonimbus and Thunderstorms Activity’, with success rates of 70.4% and 76%, respectively. However, certain clusters, such as ‘Fog and related conditions’ and ‘Cumulonimbus’ are more challenging to predict, with success rates of 56.9% and 54.9%, respectively.

5 DISCUSSION

The incorporation of LLMs for text classification and clustering has significantly improved the analysis process, enabling automated and detailed interpretation of the data provided by flow managers. The identification of weather-related disruptions was significantly aided by this approach.

The predictive models formulated through this research have shown a high degree of accuracy in categorising ATFM regulations, and an extended model that includes both weather-related and other types of regulations has shown encouraging results. A comparative evaluation with previous studies underlines the reliability and precision of our methodology, especially in the detail and accuracy of the identified weather-related clusters, while suggesting the need for a model capable of multi-tag classification.

Looking ahead, our future efforts will be directed towards defining a multi-cluster model and investigating its applicability to events beyond those caused by weather, thereby broadening the scope and utility of our approach in the field of air traffic management.

6 REFERENCES

References

Dalmau, Ramon, Very, Philippe, & Jarry, Gabriel. 2023. On the Causes and Environmental Impact of Airborne Holdings at Major European Airports. *Journal of Open Aviation Science*, **1**(2).

Reimers, Nils, & Gurevych, Iryna. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Zhang, Tian, Ramakrishnan, Raghu, & Livny, Miron. 1997. BIRCH: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, **1**, 141–182.