EasyChair Preprint

№ 14614

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

Gabriel Jarry, Philippe Very and Ramon Dalmau

August 30, 2024

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

A case study for categorising weather-related air traffic flow management regulations

Gabriel Jarry, Philippe Very & Ramon Dalmau
EUROCONTROL
Aviation Sustainability Unit (ASU) & Innovation Hub (EIH)
Brétigny-Sur-orge France
{gabriel.jarry, philippe.very, ramon.dalmau}@eurocontrol.int

***Abstract*—This research investigates the use of large language models and machine learning techniques to identify the primary triggers for air traffic flow management regulations. The study focuses on textual remarks made by flow managers who implemented these regulations. The investigation takes a concrete form by using weather-related regulations with the referenced location being an aerodrome. Specifically, a large language model is asked to assign each of these regulations to a specific group, or cluster, based on the remark made by the flow manager, where each cluster represents a particular kind of weather disruption. These clusters then act as labels for the dataset, and each regulation is combined with the weather conditions observed during its implementation. This labelled dataset is then used to train a tree-based classifier using supervised learning. This two-step methodology enables the identification of the most likely precise trigger for each regulation, such as low visibility, snow, strong winds, etc. based solely on observed weather conditions. The clusters identified by the large language model are also compared with those discovered in previous research using self-learning and supervised clustering. Nevertheless, the practical applications of this method go far beyond the classification of weather-related regulations. This approach could be used in post-operational analysis to identify the primary triggers of any type of regulation - not just weather-related. Furthermore, it enables the analysis and classification of other types of text, such as notices to airmen, further broadening its potential use cases. This paper showcases the versatility and broad application of large language models in the field of air transportation.**

***Keywords*—Large language models; air transportation; weather.**

I. 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 [1], 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.

In [1], observed airborne holdings were categorised based on their underlying high-level causes, namely weather conditions or other factors, utilising a self-learning approach. The initial labels used in this process were derived from the reasons for the ATFM regulations in effect at the airport at the time of the holding (if any). Subsequently, the specific low-level causes of the holdings likely caused by weather, such as low visibility, low ceiling, or strong winds were identified by using supervised clustering [2].

The case study in this paper aims to achieve a similar result, but with a twist: using LLMs instead of traditional machine learning. The dataset used in the experiment contains thousands of weather-related ATFM regulations. After using LLMs to analyse flow managers' comments on these regulations, we split this dataset into different clusters, each representing a specific

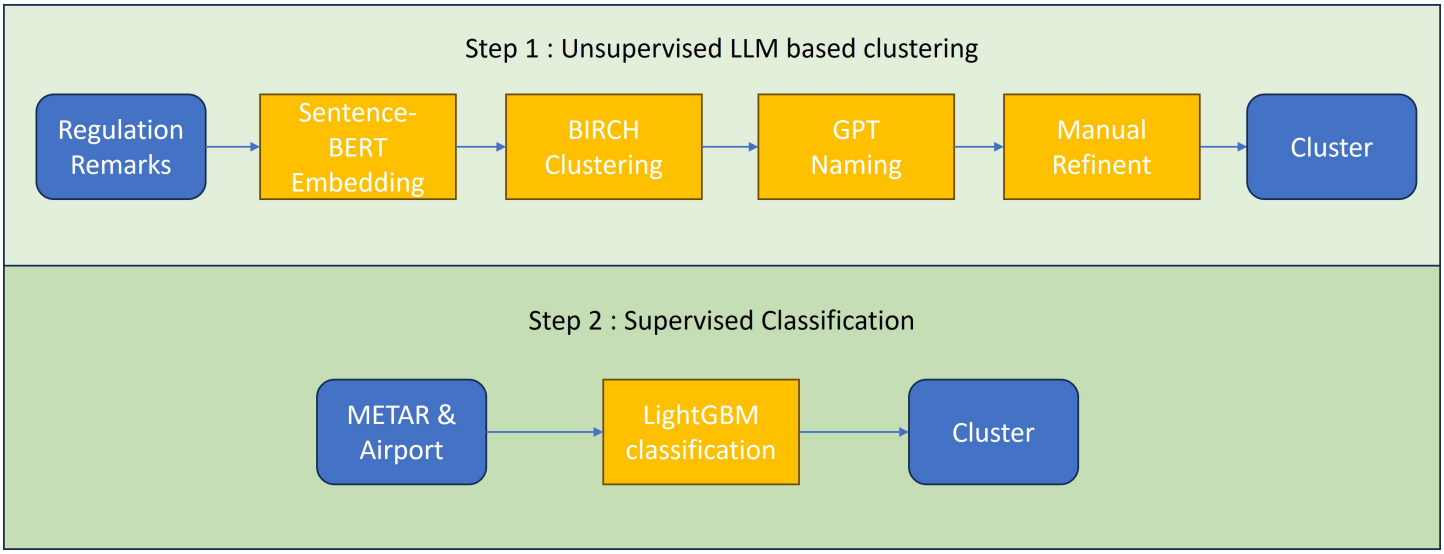

Figure 1. Methodology followed to cluster and model weather-related ATFM regulations due to weather from textual comments and weather observations.

type of weather disruption. We then match each regulation in the dataset to the weather conditions at the time it was implemented, using the clusters to label the dataset.

This enriched dataset, which includes a set of numerical and categorical features from weather observations, as well as a label corresponding to the cluster discovered in the previous step, is then used to train a tree-based classifier using supervised learning. Consequently, the resulting classifier is able to determine the most likely cause of each regulation based solely on the observed weather conditions. As a side effect, this classifier can also determine the cause of any other event (such as a holding or a flight diversion) simply by considering the weather conditions at the time of observation. This capability allows comparison with the results of previous work. Figure 1 illustrates the methodology used in this paper. Sections III and IV will provide more information on the two main steps, respectively.

It is critical to note that the primary goal of this paper is not solely focused on the quantitative results it generates. Instead, the primary focus is on explaining the methodology used and contrasting it with established approaches, such as [1]. The paper focuses on outlining the step-by-step process, the rationale behind it, and how it compares to existing methods. This emphasis highlights the paper's contribution to advancing methodologies in the field of LLMs applied to air transportation textual data, rather than simply presenting experimental results.

II. Literature Review

This section provides a brief overview of embeddings and their relevance in natural language processing, followed by a discussion on the application of LLMs in air transportation.

A. Embbedings and large language models for classifying and clustering text

Recent advances in text classification and clustering have made innovative use of LLMs and deep learning techniques. For example, the CLUSTERLLM approach refines embedders based on LLM feedback, improving text clustering performance [3]. Similarly, another framework uses LLMs to cluster news streams into key events, and then uses temporal analysis and event summarising to improve cluster coherence and significance [4].

The authors of [5] proposed a semi-supervised text clustering approach that incorporates LLMs at various stages, reducing the need for expert feedback while retaining high cluster quality. The framework proposed by [6], on the other hand, combines unsupervised clustering with contrastive learning, simplifying the process and effectively managing high-dimensional data overlap for more effectively clustering of short texts.

On the text classification front, the VGCN-BERT proposed by [7], which combines the local contextual capabilities of BERT with the global vocabulary insights from Graph Convolutional Networks (GCN), demonstrated excellent performance across a variety of datasets. In this context, the comparative study by [8] revealed performance gaps between classical and contextual word embeddings, with CNNs generally outperforming BiLSTM encoders and BERT outperforming ELMo, especially for longer texts. Another comparison shows the superiority of BERT over traditional machine learning methods in NLP tasks across different languages, highlighting the flexibility of this well-known model and the importance of transfer learning [9].

The authors of [10] performed a comprehensive review of text classification algorithms and assessed the effectiveness of

various feature extraction, dimensionality reduction and classification techniques, providing insights into their applicability and practical limitations. Additionally, advances in deep learning-based text classification methods were thoroughly reviewed by [11]. The authors evaluated the performance of over 150 models on tasks such as sentiment analysis and question answering, and proposed future research directions to address the major shortfalls.

In conclusion, the incorporation of LLMs into text clustering and classification tasks represents a significant shift towards more efficient and coherent methodologies. These advancements, combined with deep learning techniques, set an entirely novel benchmark for text analysis, demonstrating the field's rapid evolution and the potential for future innovation.

B. Large language models in air transportation

Recent research has focused on leveraging AI for text classification and clustering in aviation safety analysis, showcasing significant advancements in this domain. For instance, the ASRS-CMFS model, which is built on Transformers (a specific neural network architecture), demonstrates promise in accurately classifying any kind of incident report related to aviation [12].

In the same application, the integration of LLMs like ChatGPT into aviation safety analysis represents a significant step towards automating incident report summaries and identifying human factors [13]. This integration encourages humans and AI to work together to effectively streamline analysis processes.

More and more research emphasises the adaptation of LLMs, particularly for aviation-related tasks, implying a tailored approach to improving text classification and clustering outcomes. For instance, [14] emphasised the importance of domain-specific adaptations by implementing sentence transformers in the aviation domain. The proposed approach improved NLP task performance by pre-training on specific data and fine-tuning.

Similarly, the authors of [15] delved into NLP-based methodologies focusing on aviation safety reports, shedding light on the primary causes of weather-related delays and cancellations. This underscores the relevance of LLMs in uncovering critical insights from textual data for operational enhancements.

Last but not least, the introduction of Aviation-BERT, a model tailored for aviation safety texts, demonstrates the ability of specialised NLP models to improve the analysis of aviation incidents and accidents. Pre-trained with data from prominent aviation databases, this model outperforms its predecessors in understanding the intricacies of aviation narratives, providing a breakthrough tool for safety management [16].

In conclusion, recent research demonstrates the critical role of LLMs in revolutionising text classification and clustering in aviation texts (particularly in the safety domain), paving the way for more effective and practical methodologies.

III. Data and clustering 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.

A. Dataset description

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). A specific set of weather attributes, suitable for machine learning applications (e.g., wind speed, visibility), were extracted from the raw METARs using the `metafora`[1] tool. These attributes will be further discussed in Section IV. The dataset covers the period from January 1st, 2022 to June 1st, 2023.

Each observation in the dataset is annotated with a tag that denotes the broad category of the corresponding regulation. The tag determines whether the regulation was caused by weather (W) or by other causes (O), such as ATC capacity or industrial actions. 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 I, a subset of these comments is showcased, emphasising their variety and lack of uniformity in the format.

In total, the dataset contains 112,875 weather-related regulations, of which 26,999 include textual comments.

TABLE I
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

B. 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 [17], a tailored variant of the BERT architecture

[1] https://github.com/ramondalmau/metafora

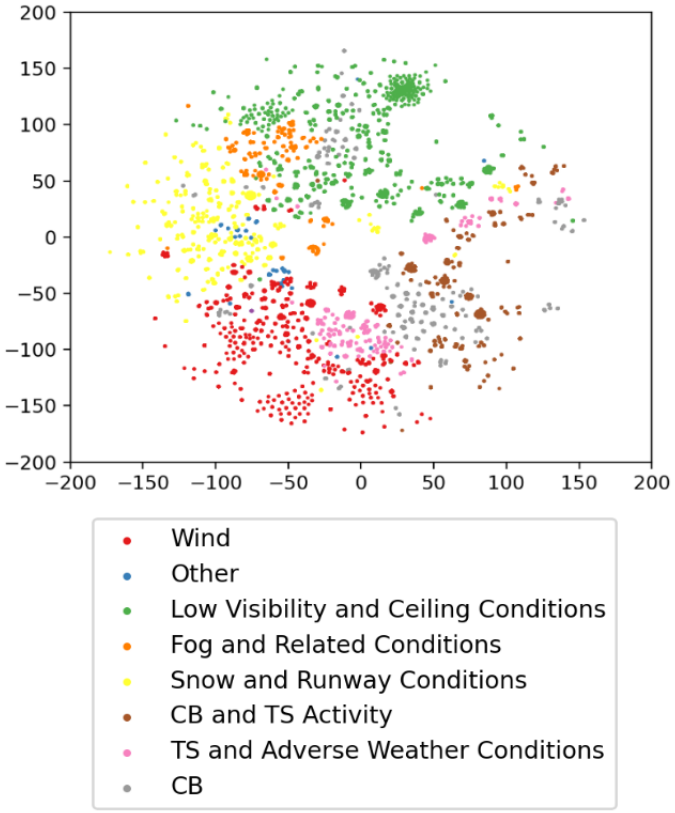

Figure 2. Cluster embedding representation using t-SNE algorithm.

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. To display the embedding, the t-SNE algorithm was used to map the representation into two dimensions. The result is shown in figure 2. We observe a consistent proximity for cumulonimbus and thunderstorm clusters and for fog with low visibility clusters.

2) **First clustering:** we used the Birch clustering algorithm [18] 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. By setting the Birch threshold parameter to 0.8, we achieved a balanced distribution of clusters, resulting in 23 groups. 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. The clusters proposed by ChatGPT are shown in Table II.
4) **Manual refinement:** As a final step, we manually 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. This careful refinement resulted in the set of clusters shown in Table III, which effectively encapsulate the various weather conditions that impact airport operations.

TABLE II
CLUSTERS OBTAINED FROM WEATHER REGULATIONS USING AUTOMATIC CLUSTERING WITH CHATGPT

Cluster name
Snow and Runway Conditions
Thunderstorms and Adverse Weather Conditions
Cumulonimbus and Thunderstorms Activity
Low Visibility and Ceiling Conditions
Fog and Related Conditions
Low Visibility Procedures
Wind
Cumulonimbus
Runway operations
Low Visibility Procedures
Thunderstorms
Aerodrome Closure
Cumulonimbus and Aerodrome Capacity
Stop Bar Unserviceable
Fog and Low Visibility Procedures
Freezing Fog and Fog
Other
ILS CAT II and CAT III
Snow and Freezing Rain
Crosswinds and High Demand
Clustered Cumulonimbus
Forecasted Cumulonimbus and Thunderstorms
Cloud Base
Forecast Hurricane

TABLE III
MANUALLY 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

TABLE IV
DATASET DESCRIPTION.

Features							Label	
Numerical					Boolean			
Name	Mean	Q1	Q2	Q3	Name	Proportion of falses	Class	Occurrences
speed (m/s)	4.0	2.1	3.6	5.1	precipitation	0.87	Unlabelled	41620 (85%)
gust (m/s)	0.6	0.0	0.0	0.0	obscuration	0.94	Weather	3484 (7%)
visibility (m)	9242	9999	9999	9999	thunderstorms	0.98	Other	4051 (8%)
ceiling (m)	2252	1067	3048	3048	snow	0.99		
cover (oktas)	3	0	2	6	clouds	0.92		

IV. SUPERVISED LEARNING

This section describes the second part of the process shown in Fig. 1, which consists of modelling the relationship between the precise reason of ATFM regulations (e.g., visibility or strong winds) and weather observations using supervised learning.

A. Supervised model for weather regulations

Our initial model, which is based on supervised learning, is designed to classify ATFM regulations into the distinct weather-related clusters found during the first step of the process (which are shown in Table III) using weather attributes as well as the name of the airport where the regulation was applied.

The weather attributes 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. Categorical features are dummy encoded, whereas numerical features are standard normalised. Table IV summarises the weather attributes used to train the model. It should be noted that the same set of features were used in our previous study [1].

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). The optimum model had an AUC of 0.947. It included 93 trees, each with a maximum of 20 leaves, and a learning rate of 0.35. 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 V reveals interesting patterns and challenges in predicting different weather clusters. 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.

TABLE V
CONFUSION MATRIX FOR WEATHER CLUSTER PREDICTION

Pred / **True**	**1** %	**2** %	**3** %	**4** %	**5** %	**6** %	**7** %	**Total** #
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

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.

The analysis of misclassifications revealed two primary groups where the model encountered difficulties. These groups are defined by conditions with similar weather events, such as 'Low Visibility and Ceiling Conditions' and 'Fog and Related Conditions' (group 1), and 'Cumulonimbus and Thunderstorms Activity', 'Thunderstorms and Adverse Weather Condition', and 'Cumulonimbus' (group 2). To enhance the model's predictive performance, these closely related clusters were amalgamated into larger clusters. This approach reduced the granularity of the predictions while boosting overall accuracy. Furthermore, to assess the model's overall performance beyond just weather information, an equal number of non-weather-related regulations (e.g., caused by ATC capacity, industrial actions, etc.) were grouped into cluster E, as depicted in Table VI.

B. Supervised model for weather and other regulations

Initially, we focused solely on weather-related clusters. However, we extended the scope of our supervised learning framework to include both weather and non-weather regulations.

This all-encompassing strategy is intended to provide a comprehensive view and lay the groundwork for a more universal

TABLE VI
FINAL REFINED CLUSTERS AFTER THE FIRST SUPERVISED LEARNING STEP

Cluster Id	Cluster description
A	Snow and runway conditions
B	Low visibility, fog and ceiling Conditions
C	Wind
D	Cumulonimbus, thunderstorms and adverse weather
E	Other regulations (not weather)

predictive model. Mirroring the approach taken with the weather-specific model, several classifiers were assessed, with LightGBM once again emerging as the superior choice. The optimal model, identified after a hyper-parameters search, achieved an AUC of 0.982. It comprised 114 trees, each with a maximum of 260 leaves, and a selected learning rate of 0.1

The efficacy of this expanded model was assessed using a confusion matrix, depicted in Table VII. The matrix demonstrates the model's proficiency in accurately categorising ATFM rules into five distinct clusters: A (snow and runway conditions), B (low visibility, fog and ceiling conditions), C (wind), D (cumulonimbus, thunderstorms and adverse weather), and E (other regulations). It is worth noting that clusters C and D exhibit some overlap (8.0% and 5.6%), which is understandable given that wind often accompanies thunderstorms.

TABLE VII
CONFUSION MATRIX FOR REFINED REGULATION CLUSTER PREDICTION

Pred / **True**	**A** %	**B** %	**C** %	**D** %	**E** %	**Total** #
A	92.1	2.1	1.0	2.4	2.4	874
B	1.0	90.3	1.2	5.6	1.9	1967
C	2.4	1.3	83.7	8.0	4.5	817
D	1.6	5.7	5.6	81.9	5.3	1853
E	1.3	2.8	2.6	5.5	87.8	5371
Total #	943	2061	961	2007	4910	10882

V. HOLDING PATTERN USE CASE AND COMPARATIVE ANALYSIS

In this section, we utilise the dataset from our previous research [1], which comprises holding patterns derived from Automatic Dependent Surveillance-Broadcast (ADS-B) trajectory data. This extraction was performed using the 'traffic' library [19]. The dataset is structured in 30-minute intervals of holding patterns, each enhanced with weather observations from the nearest METAR. Additionally, it includes associated fuel consumption calculations performed using OpenAP [20].

We applied to this dataset the final LightGBM supervised classification model and we analyse the difference between the clustering results of our previous research [1], which utilised Shapley values for clustering. The comparison is visually represented in Fig. 3. The left side of the figure displays the refined clusters as derived from the current LLM analysis, while the right side illustrates the clusters identified in our previous research. This graph employs a Sankey diagram to depict the flow and transformation of clusters. The following conclusions can be drawn from Fig. 3:

- Clusters initially representing low visibility and ceiling transition to clusters predominantly representing obscuration and ceiling, indicating a coherent correlation.
- Clusters initially representing thunderstorms and adverse weather conditions bifurcate into distinct clusters focused on clouds, thunderstorms and obscuration.
- Snow and runway conditions are primarily clustered into groups representing snow and obscuration, consistent with expected weather impact scenarios.
- The wind cluster was subject to careful analysis due to its bifurcation into clusters representing wind, ceiling and clouds. Analysis of the sub-clusters reveals a mixture of low ceiling and wind components, suggesting that the clusters may encapsulate multiple causative factors.

The discussion of overlapping clusters is made in section VII.

VI. DISCUSSION

When analysing holding patterns and their associated weather conditions, a significant observation is the overlap of clusters, particularly those related to wind and ceiling conditions. Traditional clustering methods, which tend to classify observations into mutually exclusive groups, may not fully capture the nuanced relationship between such weather variables. This complexity points to the limitations of single-day models and highlights the need for a more sophisticated modelling approach that can account for the multifaceted nature of weather impacts on aviation operations. To address this challenge, a novel modelling framework could be developed in which multiple tags can be assigned to holds and regulations to reflect the multi-factorial influences on ATM decisions. This approach allows for a more granular and accurate representation of the conditions that lead to air traffic holds and regulations.

Secondly, the approach taken for the second refinement after our initial supervised classification provides an alternative method worth considering. Specifically, we could have used the predicted labels as a basis for further analysis, cross-referencing them with the operators remarks to determine a more accurate alignment with the alternative cluster. The complex nature of weather phenomena and their interaction with air traffic management practices introduce a level of uncertainty that our current model may not fully capture. Moreover, our methodology's focus on specific types of disruptions, primarily weather-related, means that it might not account for or predict the full spectrum of factors influencing aircraft holding patterns, such as operational,

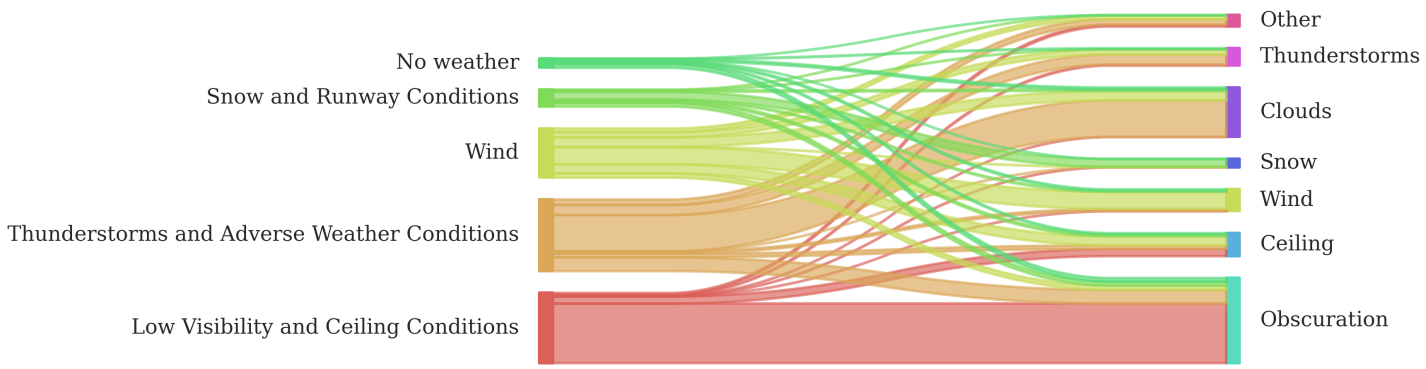

Figure 3. Comparative visualisation of clusters obtain with our current methodology (left) and our previous methodology [1] using Shapley values (right)

or technical issues. Future research directions would benefit from integrating more diverse data sources, expanding the scope to include a broader range of influencing factors to enhance the model's comprehensiveness and applicability.

Finally, relying on LLMs such as ChatGPT to analyse air traffic management texts and find primary clusters presents both opportunities and challenges. While ChatGPT excels at processing and categorising large datasets, it's not an open source model and has a cost if we want to fully automate this part. This limitation could be mitigated by exploring alternative LLMs such as Llama2 or Mistral. Comparing the performance of ChatGPT with these models could help identify the most effective approach for air traffic management applications, aiming at a balance between accuracy, interpretability and resource efficiency.

VII. Conclusions

In this study, we have developed and presented a novel methodology that uses Large Language Models (LLMs) alongside machine learning techniques to infer and identify weather-related causes behind Air Traffic Flow Management (ATFM) regulations using METAR data from major European airports. This research intricately blends the study of ATFM regulation data, reported weather conditions via METAR, and the strategic use of supervised learning algorithms to separate ATFM regulations into specific clusters that reflect different weather and operational factors.

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 ATFM operators. 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.

References

[1] R. Dalmau, P. Very, and G. Jarry, "On the causes and environmental impact of airborne holdings at major european airports," *Journal of Open Aviation Science*, vol. 1, no. 2, 2023.

[2] R. Dalmau and G. Gawinowski, "The effectiveness of supervised clustering for characterising flight diversions due to weather," *Expert Systems with Applications*, vol. 237, p. 121652, Mar. 2024.

[3] Y. Zhang, Z. Wang, and J. Shang, "ClusterLLM: Large Language Models as a Guide for Text Clustering," Nov. 2023. arXiv:2305.14871 [cs].

[4] N. Nakshatri, S. Liu, S. Chen, D. Roth, D. Goldwasser, and D. Hopkins, "Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 4162–4173, Association for Computational Linguistics, Dec. 2023.

[5] V. Viswanathan, K. Gashteovski, C. Lawrence, T. Wu, and G. Neubig, "Large Language Models Enable Few-Shot Clustering," July 2023. arXiv:2307.00524 [cs].

[6] D. Zhang, F. Nan, X. Wei, S.-W. Li, H. Zhu, K. McKeown, R. Nallapati, A. O. Arnold, and B. Xiang, "Supporting Clustering with Contrastive Learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, eds.), (Online), pp. 5419–5430, Association for Computational Linguistics, June 2021.

[7] Z. Lu, P. Du, and J.-Y. Nie, "VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification," *Advances in Information Retrieval*, vol. 12035, pp. 369–382, Mar. 2020.

[8] C. Wang, P. Nulty, and D. Lillis, "A Comparative Study on Word Embeddings in Deep Learning for Text Classification," in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, (Seoul Republic of Korea), pp. 37–46, ACM, Dec. 2020.

[9] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," *Journal of Computational and Cognitive Engineering*, Apr. 2023. arXiv:2005.13012 [cs, stat].

[10] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, p. 150, Apr. 2019. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[11] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning–based Text Classification: A Comprehensive Review," *ACM Computing Surveys*, vol. 54, pp. 62:1–62:40, Apr. 2021.

[12] S. Kierszbaum, T. Klein, and L. Lapasset, "Asrs-cmfs: Using a custom transformer-based model to predict anomalies in aviation incident reports," *Aerospace*, vol. 9, no. 591, 2022.

[13] A. Tikayat Ray, A. P. Bhat, R. T. White, V. M. Nguyen, O. J. Pinon Fischer, and D. N. Mavris, "Examining the potential of generative language models for aviation safety analysis: Case study and insights using the aviation safety reporting system (asrs)," *Aerospace*, vol. 10, no. 9, p. 770, 2023.

[14] L. Wang, J. Chou, D. Rouck, A. Tien, and D. Baumgartner, "Adapting sentence transformers for the aviation domain," in *AIAA SCITECH 2024 Forum*, p. 2702, 2024.

[15] A. Miyamoto, M. V. Bendarkar, and D. N. Mavris, "Natural language processing of aviation safety reports to identify inefficient operational patterns," *Aerospace*, vol. 9, no. 8, p. 450, 2022.

[16] C. Chandra, X. Jing, M. V. Bendarkar, K. Sawant, L. Elias, M. Kirby, and D. N. Mavris, "Aviation-bert: A preliminary aviation-specific natural language model," in *AIAA AVIATION 2023 Forum*, p. 3436, 2023.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[18] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: A new data clustering algorithm and its applications," *Data mining and knowledge discovery*, vol. 1, pp. 141–182, 1997.

[19] X. Olive, "Traffic, a toolbox for processing and analysing air traffic data," *Journal of Open Source Software*, vol. 4, no. 39, pp. 1518, 1–3, 2019.

[20] J. Sun, J. M. Hoekstra, and J. Ellerbroek, "Openap: An open-source aircraft performance model for air transportation studies and simulations," *Aerospace*, vol. 7, no. 8, p. 104, 2020.