RESEARCH ARTICLE

# REVISED Natural language processing for aviation safety: extracting knowledge from publicly-available loss of separation reports [version 2; peer review: 2 approved]

Irene Buselli[1], Luca Oneto[1], Carlo Dambra[1], Christian Verdonk Gallego[2], Miguel García Martínez[2], Anthony Smoker[3], Nnenna Ike[3], Tamara Pejovic[4], Patricia Ruiz Martino[5]

[1]ZenaByte, Genova, Italy
[2]CRIDA, Madrid, Spain
[3]Lund University, Ljungbyhed, Sweden
[4]EUROCONTROL, Brussels, Belgium
[5]ENAIRE, Madrid, Spain

**v2** **First published:** 23 Sep 2021, **1**:110
https://doi.org/10.12688/openreseurope.14040.1

**Latest published:** 18 Feb 2022, **1**:110
https://doi.org/10.12688/openreseurope.14040.2

**Open Peer Review**

**Approval Status** ✓ ✓

| | 1 | 2 |
|---|---|---|
| **version 2**<br>(revision)<br>18 Feb 2022 | ✓<br>view | ✓<br>view |
| | ↑ | ↑ |
| **version 1**<br>23 Sep 2021 | ?<br>view | ?<br>view |

1. **Riccardo Patriarca**, Sapienza University of Rome, Rome, Italy
2. **Arie Adriaensen**, KU Leuven, Leuven, Belgium

Any reports and responses or comments on the article can be found at the end of the article.

## Abstract

Background: The air traffic management (ATM) system has historically coped with a global increase in traffic demand ultimately leading to increased operational complexity.
When dealing with the impact of this increasing complexity on system safety it is crucial to automatically analyse the losses of separation (LoSs) using tools able to extract meaningful and actionable information from safety reports.
Current research in this field mainly exploits natural language processing (NLP) to categorise the reports,with the limitations that the considered categories need to be manually annotated by experts and that general taxonomies are seldom exploited.

Methods: To address the current gaps,authors propose to perform exploratory data analysis on safety reports combining state-of-the-art techniques like topic modelling and clustering and then to develop an algorithm able to extract the Toolkit for ATM Occurrence Investigation (TOKAI) taxonomy factors from the free-text safety reports based on syntactic analysis.
TOKAI is a tool for investigation developed by EUROCONTROL and its taxonomy is intended to become a standard and harmonised approach to future investigations.

Results: Leveraging on the LoS events reported in the public databases of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo and the United Kingdom Airprox

Board,authors show how their proposal is able to automatically extract meaningful and actionable information from safety reports,other than to classify their content according to the TOKAI taxonomy.
The quality of the approach is also indirectly validated by checking the connection between the identified factors and the main contributor of the incidents.

Conclusions: Authors' results are a promising first step toward the full automation of a general analysis of LoS reports supported by results on real-world data coming from two different sources.
In the future,authors' proposal could be extended to other taxonomies or tailored to identify factors to be included in the safety taxonomies.

**Keywords**
ATM, Safety, Resilience, Natural Language Processing, Losses of Separation, Safety Reports, TOKAI

This article is included in the Societal Challenges gateway.

**Corresponding authors:** Irene Buselli (irene.buselli@zenabyte.com), Luca Oneto (luca.oneto@gmail.com)

**Author roles: Buselli I**: Conceptualization, Data Curation, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Oneto L**: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Dambra C**: Funding Acquisition, Investigation, Project Administration, Supervision, Validation; **Verdonk Gallego C**: Conceptualization, Data Curation, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Review & Editing; **García Martínez M**: Conceptualization, Data Curation, Project Administration, Supervision, Writing – Review & Editing; **Smoker A**: Conceptualization, Project Administration, Supervision, Validation, Writing – Review & Editing; **Ike N**: Conceptualization, Validation, Writing – Review & Editing; **Pejovic T**: Conceptualization, Funding Acquisition, Methodology, Supervision, Validation, Writing – Review & Editing; **Ruiz Martino P**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 892542 (project FARO)
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**How to cite this article:** Buselli I, Oneto L, Dambra C *et al.* **Natural language processing for aviation safety: extracting knowledge from publicly-available loss of separation reports [version 2; peer review: 2 approved]** Open Research Europe 2022, **1**:110 https://doi.org/10.12688/openreseurope.14040.2

**First published:** 23 Sep 2021, **1**:110 https://doi.org/10.12688/openreseurope.14040.1

### REVISED Amendments from Version 1

The authors would like to thank both the reviewers for their thorough comments, which helped to improve the quality of the work.

The current version of the paper includes the following changes in response to those constructive comments:

- a couple of new paragraphs has been added in the "Discussion and Conclusions" presenting and discussing in more detail the limitations of the work;

- additional details have been added in the "Methods" section to better describe and justify the selection of methodologies and parameters;

- some specifications and clarifications have been added (mainly in the "Introduction" and "Scope of the work" sections) to better explain the relation between the scope of the work, research data and modern safety (and resilience) principles, and, in particular, between the use of a neutralized taxonomy and the exploitation of reports containing non-standardised and non-neutralised language;

- Figure 6 has been corrected (it had been accidentally substituted with a copy of Figure 7 in the previous version) and a legenda has been added to Figure 1;

- some minor editorial changes have been made in order to improve the readability and intelligibility of the article in response to reviewers' comments.

**Any further responses from the reviewers can be found at the end of the article**

## 1 Plain language summary

Nowadays, the need for automation and digitisation in the field of aviation safety is becoming crucial. In particular, this work focuses on the automated analysis of safety reports (i.e., reports describing incidents or other safety events) through different natural language processing techniques. The application of these techniques on a series of Spanish and UK reports enabled the identification of the main common topics (e.g., excessive workload), the automatic grouping of similar incidents (e.g., all the incidents originated from pilots' unfulfillment of procedures and regulations), and the extraction of the most recurrent factors (e.g., the factor representing perception problems) according to a standard taxonomy (i.e., the Toolkit for ATM Occurrence Investigation).

## 2 Introduction

The air traffic management (ATM) system has historically coped with a globally increasing traffic demand. This growing demand and increase in the amount of flights, together with the changing nature of human work, the dynamics of interactions between humans and technologies, and the way those interactions propagate at micro-meso-macro level[1], is leading to increased operational complexity[2]. One consequence of this is that new safety events are emerging and they are gradually becoming of a more complex and uncommon nature than those of yesteryear[3]. As such, attempts to understand and prevent these safety events require more detailed knowledge of underlying system dynamics. According to the Single European Sky (SES) European ATM Masterplan[4], the increased complexity of the ATM system should be absorbed by increased deployment of automation solutions in order to achieve a more efficient and safe traffic management.

The FARO project — saFety And Resilience guidelines for aviatiOn — focuses on the problem of dealing with the impact that an increasingly complex environment has on the system safety. FARO is an exploratory research project, part of the SESAR – Single European Sky ATM Research and Development programme. In particular, this paper reports the first steps of the project, which respond to the objective of capitalising the extant knowledge of safety by exploring the field of systematic extraction of information through data-driven techniques. In this work the focal research subject is a specific manifestation of ATM safety, the loss of separation (LoS), and in particular the analysis of the LoS safety reports produced by states' Civil Aviation Authorities and Air Navigation Service Providers (ANSPs) after investigation.

The reports considered in this study are specific to Spanish and UK airspaces, and are collected in the public databases of, respectively, the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA). and the UK Airprox Board (UKAB).

In general, safety reports are extremely valuable sources of data to learn from past incidents, as well as to identify new threats to safety and ways to avoid them[5]. However, manual analysis of these reports is complex and requires considerable resources. Each safety report is composed mainly of free text in natural language, which makes it difficult for automated tools to process them. This work exploits natural language processing (NLP) towards partial automation in the analysis of safety reports.

In the last two decades, the application of NLP to safety reports has been increasingly explored, but research has mainly focused on developing models and algorithms to categorise incident reports[6–9]. All of these works rely on an initial set of labels and training data consisting of safety reports previously labelled by domain experts. The biggest limitation of this approach is its lack of generality: to generate a new set of labels and training data, substantial resources and effort would be needed. In this context, the importance of introducing common sets of labels — i.e., common taxonomies — became evident. On one hand, tools like the Toolkit for ATM Occurrence Investigation (TOKAI) have been developed to generate structured safety data[10], and the data collected have proved to be extremely useful for quantitative analysis[11]; on the other hand, some applications of NLP techniques have focused on the categorisation of the safety reports according to taxonomy factors[12]. However, the categorisation approach shows another limitation: while it is a good way to automatise a task performed by domain experts, this approach does not allow the discovery of unknown patterns or further knowledge. To partially overcome this limitation, unsupervised techniques like topic modelling[5,13–15] and similarity clustering[5,16] are now being explored.

To address current gaps in the literature, in this work the authors propose a twofold approach:

- An unsupervised phase: an exploratory data analysis (EDA) is performed on safety reports combining state-of-the-art techniques like topic modelling and clustering;
- A partially supervised phase: for the first time, an algorithm able to extract TOKAI taxonomy factors from the free-text safety reports is developed, based on syntactic analysis.

Both these phases focus on the mining of free text contained in LoS reports in order to identify — via topic modelling and clustering — and categorise — via the TOKAI-taxonomy-extraction algorithm — common patterns of behaviour. In particular, topic modelling and clustering act in an unsupervised fashion, enabling the detection of possibly unknown recurrent behaviours or conditions during LoS events, while the application of syntactic analysis allows the association of predetermined patterns of behaviour to TOKAI taxonomy factors (e.g., perception, conformance to procedures, or memory). An analysis of the importance of combining unsupervised approaches and taxonomy exploitation as well as the main limitation of both methodologies can be found in 17. The choice of the TOKAI taxonomy for the second phase is based on three main reasons.

First, the TOKAI taxonomy makes a significant shift from traditional causal taxonomies based on negative perspectives (i.e., describing errors or failures) by its use of neutralised language: TOKAI factors are neither negatively nor positively oriented, so they can ideally explain both ordinary operational situations and safety occurrences[11]. The second reason is of a more practical nature: the structure of TOKAI taxonomy is particularly suited to allow aggregation at different levels of detail, given its multi-level hierarchical structure. Lastly, the TOKAI taxonomy is intended to become a standard and harmonised approach to future investigations, allowing ANSPs to share lessons from ATM occurrences[11]. As such, extracting the same factors from past reports may be useful to partially align the past analyses with the future ones — even if it should be borne in mind that reports written with completely different logics and conceptual philosophies can be hardly comparable even when their content is reshaped according to the same taxonomy.

More in general, it is worth keeping in mind that both the unsupervised discovery of unknown patterns and the supervised extraction of taxonomy factors can only be as rich as the data contained in the free text of the reports. The impact of this matter on this work is discussed more thoroughly at the end of the paper, after presenting all the results.

The rest of the paper is organised as follows. Section 3 details the scope of this work. Section 4 describes the available data used to test the methodologies (presented in Section 5). Section 6 reports on the results from the application of the proposed methodologies on the available data. Finally, Section 7 concludes the paper.

## 3 Scope of the work

The scope of this work is to facilitate the extraction of meaningful and actionable information from recent (i.e., between 2017 and 2019) CEANITA and UKAB LoS reports, and, in particular, to automatically identify recurrent behaviours and common precursors. More specifically, a twofold approach was applied:

- First, an EDA was performed in order to get general insights into the LoS phenomena (see Section 3.1);
- Then, an algorithm able to extract selected TOKAI taxonomy factors from the free text of CEANITA reports was developed, based on syntactic analysis (see Section 3.2).

### 3.1 Exploratory data analysis

The exploratory data analysis was conducted in two stages.

In the first stage, the most recurrent topics in the corpus of both CEANITA and UKAB LoS reports were identified exploiting topic modelling[18]. Topic modelling is an unsupervised NLP technique able to automatise the extraction of the most recurrent topics and compute their prevalence in each report. This technique enables a high-level analysis of the content of each report, which can therefore be described through numerical features and possibly compared in a scalable way, without the need to read and understand them one by one.

In the second stage, a cluster analysis[19] is applied to group similar LoS events in terms of the various themes or safety areas contained in the reports (i.e., topics prevalence, main causes of the incident, safety barriers, etc.).

### 3.2 Automatic extraction of TOKAI taxonomy factors

Each CEANITA report concludes with a free-text description of the main actions performed by air traffic controllers (ATCos) and pilots, summarising the dynamic of the incident. Analogously, in a sample of UKAB reports, the final assessment of cause is listed in free text and includes the main contributory – both to the incident and to its resolution – factors based on pilots' and controllers' actions. This information is crucial to understand the dynamics at play in the LoS. Therefore, the automatic extraction and classification of these behaviours can be of paramount importance. In particular, labelling these behaviours according to a standard taxonomy — in our case the TOKAI one — can enable the application of quantitative-analysis techniques[11] on information extracted from various reports and/or repositories.

Syntactic analysis is a branch of NLP which focuses on determining the grammatical structure of a sentence. State-of-the-art tools for syntactic analysis[20] are able to identify base-form verbs (e.g., the group of verbs related to TOKAI factor A.1: "detect", "identify", "see', "hear", etc.), as well as to retrieve information about their role in the sentence: their form (e.g., active or passive) and their subject. Thus, each action can be potentially associated to a TOKAI taxonomy factor, whilst maintaining information about who performed the action.

After extracting these taxonomy factors it is possible to estimate the occurrences of each factor in the corpus of reports. Consequently, relying on this information, a sort of sanity check can be performed to validate this syntactic-analysis approach: a simple Machine Learning model is developed to predict whether the main responsibility of the incident is ascribed to the ATCo, pilot, or both (which is an information reported in each report as a result of the investigation). In fact, the more reliable the extracted information is, the more reasonable it is to assume it should be predictive of the main contributor(s) of the incident. Whilst this is not a direct validation of the model, the outputs are indicative of the reliability of the algorithm.

For the sake of completeness, it is worth noticing that the UKAB and CEANITA reports used as the base material are not written and constructed using the TOKAI taxonomy. This taxonomy offers great analytical benefits – including the granularity with which it is able to categorise events as factors as well as its use of neutralised language, which is broadly consistent with contemporary approaches to safety – but it has to be borne in mind that the algorithm simply classifies the information included in the reports and, as the reports are not written in a standardised way, the algorithm is in turn not able to standardise them, but only to reshape their content.

## 4 Data description

For the scope of this study, two publicly-available data sources were exploited: CEANITA reports (see Section 4.1) and UKAB reports (see Section 4.2).

### 4.1 CEANITA LoS reports

The considered CEANITA LoS reports consist of 89 safety reports, written in Spanish and published by Spanish Safety Aviation Agency (AESA), which is the Spanish Civil Aviation Authority, under the commission of CEANITA, covering safety-related occurrences in the Spanish airspace between January 2018 and July 2019.

The initial sections of these reports are written in fixed formulas or tabular format. This enables the direct extraction of some categorical or numerical variables through an automated search for keywords or table margins, and therefore the computation of some basic descriptive statistics such as:

- the *ICAO risk category*: 9% of the occurrences are classified as A, while the majority is classified as B (55.1%) and C (31.5%), and only 3.4% as D and 1.1% as E (classes assigned according to the ICAO classification[21,22]);
- the *main causes*: the most frequent ones in the corpus are wrong clearance (52%), deviation from procedures (22%), wrong or no resolution (17%–15%), coordination problems (17%), and late or no detection (15%–16%) — note that multiple causes are possible;
- the *airspace class*: most of the reported incidents happened in class C, D (40% each), and A (11%), while only 6% in G and 3% in E (classes assigned according to the ICAO classification[23]);
- the *pilots and ATCo contribution*: pilots' contribution is classified as direct in 36% of the cases, as indirect in 15%, and as none in 49%. ATCo contribution is, instead, direct in the majority of cases (72%), indirect in 9% of the incidents, and none in 19%.

The remaining part of each report is written as a free text and divided in the following sections:

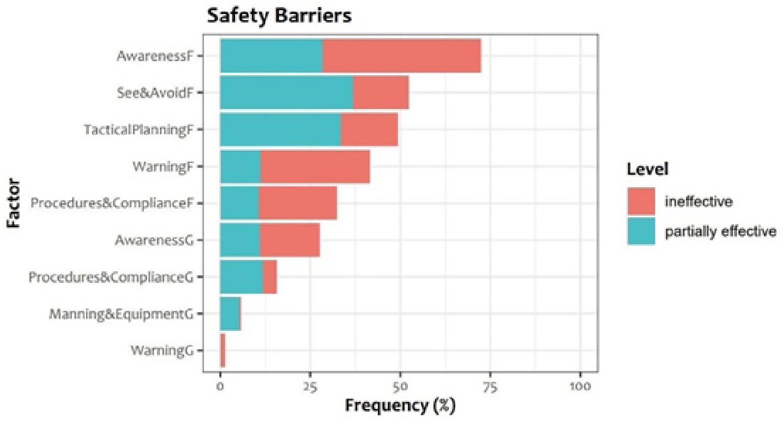
- *Initial situation*: in this section the initial situation (i.e., the initial location and condition of the aircraft involved in the LoS) is described with text and images.
- *Communications and radar tracks*: in this section the communications of interest between ATCos and pilots are summarised.
- *Extract from received reports*: this section is a summary of the different reports received by the commission of investigation (i.e., from the pilot, the co-pilot, the executive controller, and other involved agents, if any).
- *Conclusions*: this section describes the conclusions of the investigation, summarising the dynamic of the LoS based on the main actions performed by the involved human actors.

### 4.2 UKAB LoS reports

The considered UKAB LoS reports consist of 549 safety reports, written in English and published by UKAB, covering safety-related occurrences in the UK airspace between January 2017 and December 2019.

Similarly to the CEANITA reports, some sections of these UKAB reports are written in fixed formulas or tabular format, so that some information can be directly extracted, such as:

- the *ICAO risk category*: 9.3% of the events are classified as A, 26.8% as B, 48.5% as C and 2.2% as D, while 13.3% are classified as E, which means a non-negligible part of the reports describe situations that were not actual incidents;
- the *airspace class*: most of the reported incidents happened in class G (89.1%), while 7.3% in class D, 2.2% in class A and 1.3% in class C;
- the *safety barriers*: their distribution in the considered reports can be seen in Figure 1 and a detailed explanation of each single barrier is available from AIRPROX.



Legenda

| Barrier | Abbreviation |
|---|---|
| GROUND ELEMENTS | |
| Regulations, Processes, Procedures and Compliance | Procedures&ComplianceG |
| Manning & Equipment | Manning&EquipmentG |
| Situational Awareness of the Confliction & Action | AwarenessG |
| Electronic Warning System Operation and Compliance | WarningG |
| FLIGHT ELEMENTS | |
| Regulations, Processes, Procedures and Compliance | Procedures&ComplianceF |
| Tactical Planning and Execution | TacticalPlanningF |
| Situational Awareness of the Conflicting Aircraft & Action | AwarenessF |
| Electronic Warning System Operation and Compliance | WarningF |
| See & Avoid | See&AvoidF |

**Figure 1. Frequency of ineffectiveness and partial ineffectiveness of Safety Barriers in UKAB reports.**

The remaining part of each report is written as a free text and divided in the following sections:

- *Information reported to UKAB*: this section summarises the different reports received by the Airprox Board, not only from the pilot, the executive controller, and the other directly involved agents, but also from UKAB Secretariat and experts commenting on the event;
- *Board's discussion*: this section describes the Board's discussion and its conclusions, motivating each consideration based on the information available;
- *Assessment of cause and risk*: this section briefly summarises the main causes, contributory or interesting factors, and the effectiveness of Safety Barriers as identified by the Board. Safety Barriers are described in Figure 1 together with the distribution of their effectiveness.

## 5 Methods

This section presents the theoretical framework of the methods exploited to achieve the scope of the work (see Section 3) leveraging the data described in Section 4. Four main techniques were exploited: topic modelling (Section 5.1), clustering analysis (Section 5.2), syntactic analysis (Section 5.3), and data-driven predictive models (Section 5.4).

### 5.1 Topic modelling

Topic modelling is an unsupervised NLP technique designed for the first time by David Blei and John Lafferty[18], and largely used in the transportation domain[24,25]. The idea behind topic modelling is to represent a corpus of documents in terms of a certain number of topics, identified in a completely unsupervised fashion, based only on how the words are distributed in the documents. For these characteristics, this method is particularly suited to outline the main themes in a collection of documents. The statistical intuition behind topic modelling can be summarised in three points:

- A document can be defined as a set of words/n-grams.
- A document contains different topics according to a certain distribution.
- A topic can be in turn defined through a certain distribution of words/n-grams prevalence.

Thus, by observing the frequencies of words/n-grams in a collection of documents, it is possible to estimate the two underlying distributions fitting the observed frequencies. In particular, the most widely used technique for topic modelling, the latent Dirichlet allocation (LDA), is characterised by the theoretical assumption that these distributions can be derived based on the Dirichlet probability distribution[5,13]. This framework also generates a topic-word matrix in which each topic is represented through weights associated to each word/n-grams. This information can be used to interpret the (otherwise unlabelled) topics.

In this work, for both the use cases, n-grams with n>2 were discarded as a number of preliminary analyses (performed to understand the relevance and usefulness of different n-grams to describe the reports' content, both per se and in the topic-modelling framework) revealed that the role of n-grams with n>2 was substantially negligible per se and source of additional noise.

Both the LDA models were developed using the textmineR library version 3.0.5 from R development environment version 4.0.3. The number of iterations for the Gibbs sampler to run was set to 500 and the burn in was set to 180 (they were set according to a mixture of standards assumptions and convergence assessments by looking at the likelihood graphs produced by the models), while every 10 iterations of the sampler alpha was set to be optimised and the likelihood to be re-computed. The initial alpha, beta, and especially the number *k* of topics to be generated was tuned by testing different options and evaluating the results in terms of probabilistic coherence and R-squared (in particular, the initial alpha was finally set to 0.1 and beta to 0.05 for both the models, while for the CEANITA use case $k = 27$ and for the UKAB one $k = 60$).

The final number of meaningful and significant topics to be considered was manually identified to be 12 for the CEANITA reports and 23 for the UKAB reports, according to the experts of the field involved in this work (i.e., from CRIDA, Lund University and ENAIRE). Some topics were discarded as too similar between each other, some topics because they made sense on a lexical point of view (e.g., commonly used idioms or standard sentence formulations) but did not convey any meaningful information, others because they did not make much sense or were not very coherent. This selection was performed in general agreement, suggesting a certain robustness and reproducibility of this evaluation.

## 5.2 Clustering analysis

Clustering analysis[19] is a technique used to group data according to a certain definition of similarity. Many clustering methods exist. In particular, hierarchical clustering is one of the most largely exploited ones[26]. In this context, the agglomerative hierarchical clustering, as opposed to the divisive one, has been shown to be the most effective[26], and this is one of the reasons why it is the one used in this work. The idea behind the agglomerative hierarchical clustering is the following. Initially, each point in the dataset is separate from the other and considered as an individual cluster. Then, each cluster is merged with other clusters based on their mutual distance, where the definition of distance is chosen according to how well it describes the concept of (dis)similarity in the considered data (e.g., correlation-based distances are often used in gene expression data analysis). Keeping track of each step of the process, the clusters are thus merged until all the data converge to a single cluster. Finally, the user selects the best number of clusters based on the knowledge of the subject, or the intra-cluster variability, or other particular statistical metrics[27]. The simplicity of the underlying idea and the high interpretability of the results is another reason why the authors considered it suitable for the work. However, the choice of agglomerative hierarchical clustering was not only due to a priori knowledge, but it was also confirmed by the actual comparison of the application of different algorithms on the two use cases of interest.

In this case, data were merged according to Ward's minimum variance criterion (i.e., the distance between objects is proportional to the squared Euclidean distance, which is the standard for most applications), using the basic stats library version 0.1.0 from R development environment version 4.0.3, after normalising all the numerical variables (and after mapping the qualitative variables of interest into numerical features, i.e., the main cause and the contribution for the CEANITA reports and the safety barriers for the UKAB ones). The choice of the number of clusters was made considering both experts' knowledge of the field and statistical metrics, in particular the dendrogram and the scree plot.

## 5.3 Syntactic analysis

Syntactic analysis deals with the problem of analysing a string in natural language to identify the syntactic and grammatical structure of each sentence. In this work, syntactic analysis is performed using the UDPipe library version 0.8.6 from R development environment version 4.0.3, a state-of-the-art open-source library which automatically generates sentence segmentation, tokenisation, part-of-speech tagging, lemmatisation, and dependency parsing. Models are provided for 50 languages. An example of the output of the UDPipe library can be found in Table 1. A detailed explanation of the tool can be found in 20.

## 5.4 Data-driven models (for validation of Algorithm 1)

Data-driven predictive models are based on the idea of learning relations between inputs (e.g., taxonomy factors prevalence extracted by Algorithm 1) and outputs (e.g., LoS direct contribution reported in the reports) through a series of examples (i.e., historical data). This will serve as validation of the quality of the taxonomy factors prevalence extracted by the proposed Algorithm 1.

In this context, support vector machines (SVMs)[28] represent state-of-the-art solutions for many real-world applications[29,30] in the framework of (shallow) machine learning algorithms. Even if, currently, deep learning approaches[31] were shown to outperform shallow learning models in many tasks (e.g., vision and speech recognition), they require very large amounts of data to be trained, which were not available for this research. As for the previously described techniques, the choice of SVMs over other data-driven methods was due both to a priori knowledge and experimental results.

**Table 1. Example of syntactic analysis with UDPipe for the sentence "The radar controllers did not issue timely traffic information".** The meaning of "Part of Speech" and "Dependency" elements is standard[a].

| Sentence | Lemma | Part of Speech | Dependency |
|---|---|---|---|
| The | the | det | det |
| radar | radar | noun | compound |
| controllers | controller | noun | nsubj |
| did | do | aux | aux |
| not | not | part | advmod |
| issue | issue | verb | root |
| timely | timely | adj | amod |
| traffic | traffic | noun | compound |
| information | information | noun | obj |

[a] https://universaldependencies.org/u/dep/all.html

SVMs are the most effective algorithms in the family of Kernel Methods[28] (i.e., methods exploiting the "kernel trick" to extend linear techniques to the solution of nonlinear problems). SVMs have a series of hyperparameters—the kernel, which is often fixed to be the Gaussian one[32], the kernel hyperparameter, and the complexity hyperparameter — which deeply influence their performance and need to be tuned during the model-selection phase[33].

However, data-driven predictive models need not only to be tuned (by finding the optimal hyperparameters) but also to be evaluated in terms of their performance in a rigorous statistical way. Model selection and error estimation are meant to deal exactly with this problem[33]. Resampling techniques like k-fold cross validation and non-parametric bootstrap are between the most commonly exploited solutions, since they are proved to work well in many situations[33]. The idea behind these techniques is simple: the original dataset is re-sampled once or more, without replacement, to build three independent datasets called learning, validation, and test set. The learning set is exploited to train the model. The validation set is exploited to find the optimal hyperparameters (i.e., the ones that lead to the optimal performance). The test set is exploited to estimate the performance of the final model: in this way, the test is independent from both the learning and the validation.

Performance measures strongly depend on the task to be solved. In this case, dealing with classification problems, accuracy and confusion matrix are the most widely used metrics[34].

In particular, in this work two SVM models with Gaussian kernel were trained on each collection of reports (i.e., one for ATCos and one for pilots, as better explained in Section 6), performing accurate model selection (the kernel and the complexity hyperparameters were searched in $\{10^{-4.0}, 10^{-3.5}, \bullet \bullet \bullet , 10^{3.0}\}$) based on both accuracy and balancing of the confusion matrices.

## 6 Results

This section shows how the methods presented in Section 5 have been applied to achieve the scope of the work (see Section 3) demonstrating the effectiveness of the proposed approach on both the sets of data described in Section 4. Specifically, Section 6.1 presents the results of EDA, summarising the main outcomes produced by topic modelling and clustering analysis, while Section 6.2 presents the results of the syntactic analysis approach used to connect the reports with the TOKAI taxonomy, also validating the quality of the methodology.

### 6.1 Exploratory data analysis

This section shows how topic modelling (see Section 5.1) can be used to extract the main topics from the 89 CEANITA reports and the 549 UKAB reports and how clustering analysis is able to group the LoS events in a meaningful way, in both the sets of data considered.

***6.1.1 Topic modelling on CEANITA reports***. The application of LDA for topic modelling on CEANITA reports led to the identification of 12 main topics. These 12 topics can be defined by lists of words and bigrams, to which the FARO experts have associated representative labels (see Table 2). The selection of the topics was performed through both automated procedures (i.e., relying on coherence metrics) and more handcrafted adjustments (i.e., consulting the FARO experts, which filtered the topics and retained the most meaningful and coherent ones according to their knowledge of the domain).

Topic modelling results enable the description of the reports at a higher resolution then simple descriptive statistics. For instance, while the main causes of the incident are identified just by looking at variables described in Section 4.1 (e.g., if a wrong ATCo clearance was responsible), topic modelling also provides additional information (e.g., if the ATCo's behaviour that led to the wrong clearance might have been affected by an excessive workload or an emergency situation). Indeed, this technique also outputs the probability of finding a certain topic—namely, the topic's prevalence — in each report, thus generating for every document a set of numerical features quantitatively describing its content. Figure 2 shows the average prevalence of each topic over the CEANITA reports. Observing Figure 2 it can be seen that exogenous factors like fire-extinguishing emergencies or adverse-weather problems are quite rare (only about 10% of the incidents contain one of these topics), while workload is present in almost 40% of the reports.

***6.1.2 Topic modelling on UKAB reports***. The application of LDA for topic modelling on UKAB reports led to the identification of 20 main topics, whose description in terms of words and bigrams in Table 3. Figure 3 shows the average prevalence of each topic over the UKAB reports. There is some significant heterogeneity in prevalence: for instance, communication, late sighting, and downwind leg are quite frequent (around 30%) while topics about parachuting, weather or training are quite rare (around 2%).

***6.1.3 Clustering analysis on CEANITA reports***. In order to identify the relations between the topics' prevalence and the other contextual information (i.e., the main causes of the LoS and the pilots' and ATCo's contribution to the incident), a further analysis was then conducted by applying clustering analysis (see Section 5.2). For this purpose, for each CEANITA report, a feature set was created, composed of the prevalence of the 12 topics, the main causes, and the level of pilots' and ATCo's contribution to the LoS. Hierarchical clustering with Ward distance was then applied on the resulting dataset. After visualising different statistical metrics through dendrograms and screeplots (i.e., the two most common methods for cluster selection) and jointly consulting the FARO experts, 8 different clusters were identified. As shown in Figure 4, they strongly differ in size and normalised frequency of the features. Indeed, looking at Figure 4 some observations arise:

- Two very small subgroups (Clusters 6 and 8) are identified as particularly different from the others. In

**Table 2. Words and bigrams of the 12 topics extracted with LDA from CEANITA reports, together with the representative label associated to each topic by FARO's experts (English translation from Spanish).**

| Words/Bigrams | | | | | | | Topic |
|---|---|---|---|---|---|---|---|
| helicopter | drop | water | fires | extinguishing | coordination | drop area | fire |
| load | work | high | alone | workload | instructions | previous | workload |
| departure | to take off | aircraft climb | runway | to take off aircraft | rate | they are | departure |
| wind | tail | down-wind | leg | wind leg | right tail | runway | traffic circuit |
| weather | adverse | adverse weather | detours | meteorologic conditions | due to weather | thunderstorm | weather |
| runway | go around | go | around | to take off | to land | aircraft established | go around |
| sectors | sector aircraft | frequency sector | high | coordination | transfer | limit | sector operations |
| answer | received | finally | decided | they saw | communication | visual contact | communication |
| clearance | course descent | aircraft to descend | descent rate | sector to descend | aircraft to maintain | rate | descent |
| received | coordinating | confirming | to confirm receipt | maintaining formation | sector informs | receipt | coordination |
| alert | early | early alert | activation function | activation | function | alert function | alert |
| military | military formation | formation | military aircraft | defence | air defence | main centre | military |

particular, Cluster 8 is composed of two LoS events where the main topic is "fire" (indeed, they are the reports referred to Llutxent fire in summer 2018); Cluster 6 instead contains the four incidents caused by level bust (and, as expected, according to the heatmap, the main contribution was from the pilots and the other main conclusion was "deviation from procedures").

- The the highest frequency of ATCo contribution and an interesting high prevalence of "descent" topic largest cluster (Cluster 5) is mainly composed of wrong-clearance and late-detection incidents, with clearly.
- Cluster 4 contains incidents mainly caused by "wrong resolution" of the ATCo, with high prevalence of topics related to go-around, departure, and weather.
- Cluster 7 is composed of incidents caused mainly by transfer or coordination problems. The most frequent topics here are "sector operations" and "military".
- Incidents in Cluster 3 are essentially due to Pilots' errors, in particular to airspace infringement and unfulfillment of the Visual Flight Rules (VFR).
- Cluster 2 is characterised by incidents due to Pilots' deviations from procedures, especially in the landing phase (see topic "traffic circuit").
- Cluster 1 is composed of incidents due to ATCo inability to both detect and resolve the LoS. This cluster is interestingly characterised by high values of the topic "alert";
- Interestingly some topics (e.g., workload and communication) are almost homogeneously distributed in all the clusters, without peaks in their relative-frequency values.

***6.1.4 Clustering analysis on UKAB reports***. An analogous analysis was conducted on UKAB reports: the feature set in this use case is composed of the prevalence of the 20 topics and the safety barriers. Also in this corpus 8 clusters were identified, differing in size and features (see Figure 5). Figure 5 shows that:

- Also in this use case, two relatively small clusters (Cluster 6 and Cluster 7) are easily identifiable as sort of "outliers": one is essentially described by the absolute prevalence of the topic "paragliders" and the other one by the topic "military jets", without other particularly evident features.

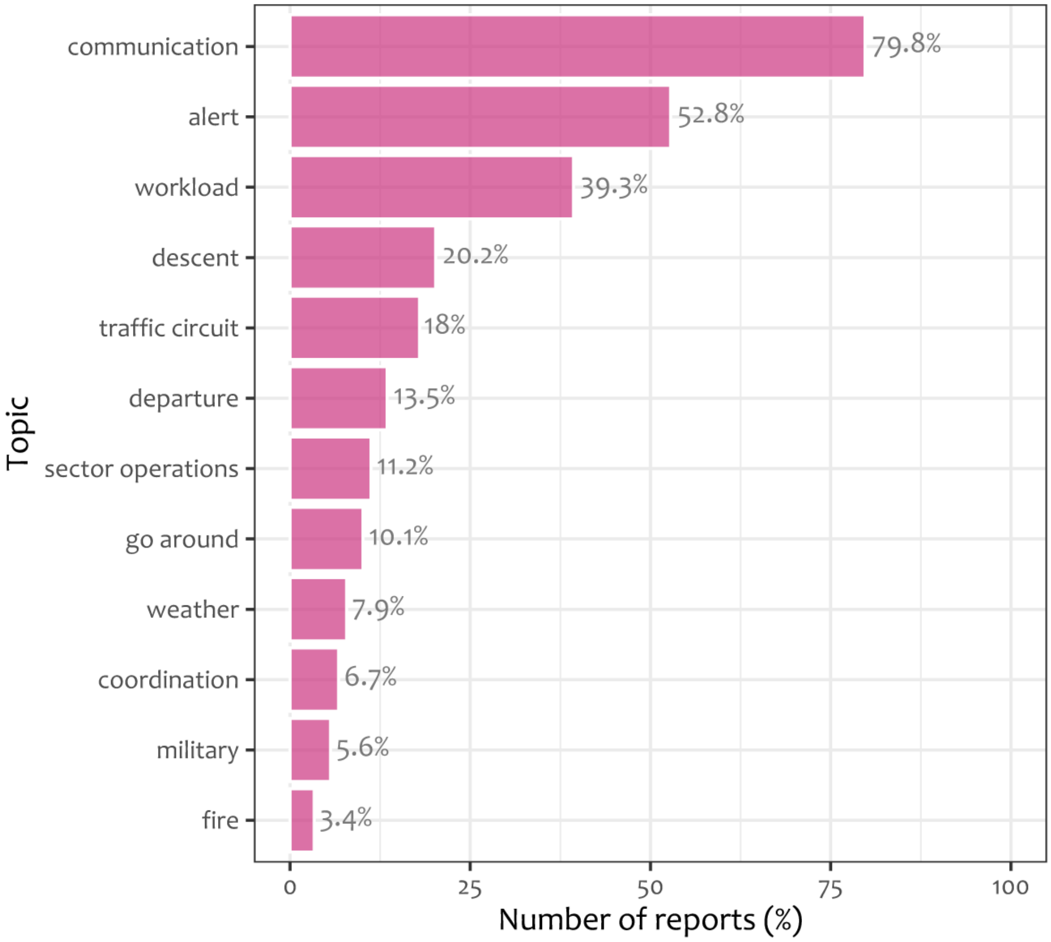


**Figure 2. Average prevalence of each of the 12 topics of Table 2 over CEANITA reports.**

- Cluster 8 is interestingly characterised by three main topics/factors: weather, monitoring and electronic warning system at ground level, which may be correlated in some way.

- Cluster 1, even being the largest one, seems to reunite the incidents where the responsibility is rarely attributed to the pilot: most of the Safety Barriers related to flight level have lower levels than the average, while the darkest colours on the heatmap correspond to concern, communication, and workload, highlighting a significant human component.

- Cluster 2 is mainly formed by incidents happened during landings, with a slight correlation with procedures compliance, tactical planning and execution ascribed to pilots.

- Cluster 3 mostly contains late-sighting incidents, correlated with factors like lack of situational awareness of the pilot and electronic warning problems; interestingly, this cluster is the one with higher prevalence of the glider topic.

- Cluster 4 is formed by low-altitude incidents: the heatmap highlights both departure, descent, overhead join, and parachuting topics; the main causal factor seems to be Manning & Equipment at ground level.

- Cluster 5 is again about late-sighting incidents, with particular focus on two topics: transponder and manoeuvres.

In order to sum up, at the end of this section reporting EDA results on both CEANITA and UKAB reports, it is interesting to notice that the identification of topics and clusters in the two corpora reveal both clear differences (e.g., various topics found in UKAB reports – like “paraglider”, “glider”, or “parachuting” – did not emerge in the CEANITA ones, and even for some common topics – like “workload” or “descent” – there are in fact differences in frequencies) and strong similarities (e.g., topics like “communication” or “weather” have very similar prevalence in the two corpora, and many of the clusters identified follow a similar logic). While similarities are somehow easily expected, there may be various reasons behind the differences, which can be summarised in these two points:

**Table 3. Words and bigrams of the 20 topics extracted with LDA from UKAB reports, together with the representative label associated to each topic by FARO's experts.**

| Words/Bigrams | | | | | | | Topic |
|---|---|---|---|---|---|---|---|
| overhead | join | joining | deadside | overhead join | crosswind | circuit | overhead join |
| student | instructor | student pilot | training | solo | hand | control | training |
| advised | requested | acknowledged | asked | received | inbound | passed pilot | communication |
| fast | military | jet | high | fast jet | range | manoeuvres | military jet |
| survey | manoeuvring | company | operations | conducting | aerobatics | manoeuvres | manoeuvres |
| runway | go around | go | around | to take off | to land | aircraft established | go around |
| sectors | sector aircraft | frequency sector | high | coordination | transfer | limit | sector operations |
| answer | received | finally | decided | they saw | communication | visual contact | communication |
| site | gliding | winch | glider | launch | active | sites | glider |
| departure | departing | climb | climbing | airborne | departed | depart | departure |
| converging | sighting pilot | required give | late sighting | pilot required | considered converging | converging pilot | late sighting |
| descent | descend | descending | altitude | descended | feet | vertical | descent |
| trainee | ojti | handover | clearance | instruction | cleared | training | training |
| helicopter | site | helicopters | helicopter pilot | wing | landing | lifting | helicopter |
| service | altitude | cloud | weather | receiving | altitude ft | condition | weather |
| transponder | primary | primary contact | twin | selected | serviceable | equipped | transponder |
| approach | final | runway | instructed | landing | final approach | leg | final approach |
| monitor | traffic information | required monitor | warning | definite risk | monitor flight | definite | monitoring |
| para | drop | parachuting | dropping | parachute | site | para dropping | parachuting |
| busy | workload | high | working | handover | controlling | inbound | workload |
| paraglider | paramotor | flypast | paragliders | paraglider pilot | paramotor pilot | paraglider pilots | paraglider |
| concerned | normal | concerned proximity | standards | pilot concerned | safety standards | pertained | concern |
| circuit | downwind | visual circuit | leg | pattern | ahead | circuit traffic | downwind leg |

- A difference in the context and type of the reported events: in particular, as described in Section 4, the vast majority of the incidents reported in the UKAB sample happened in class G while the incidents in the CEANITA sample are mostly from airspace classes C, D, and A. The different nature of the incidents reported implies great differences in flight rules and in the role of air traffic control, which may justify differences in the dynamics of the LoS and, consequently, in the reports' content;

- A difference in the way incidents (even similar) are described: UKAB and CEANITA reports are clearly different in terms of reporting logic and culture, and unsupervised NLP techniques do not extract necessarily what was important in the incident but

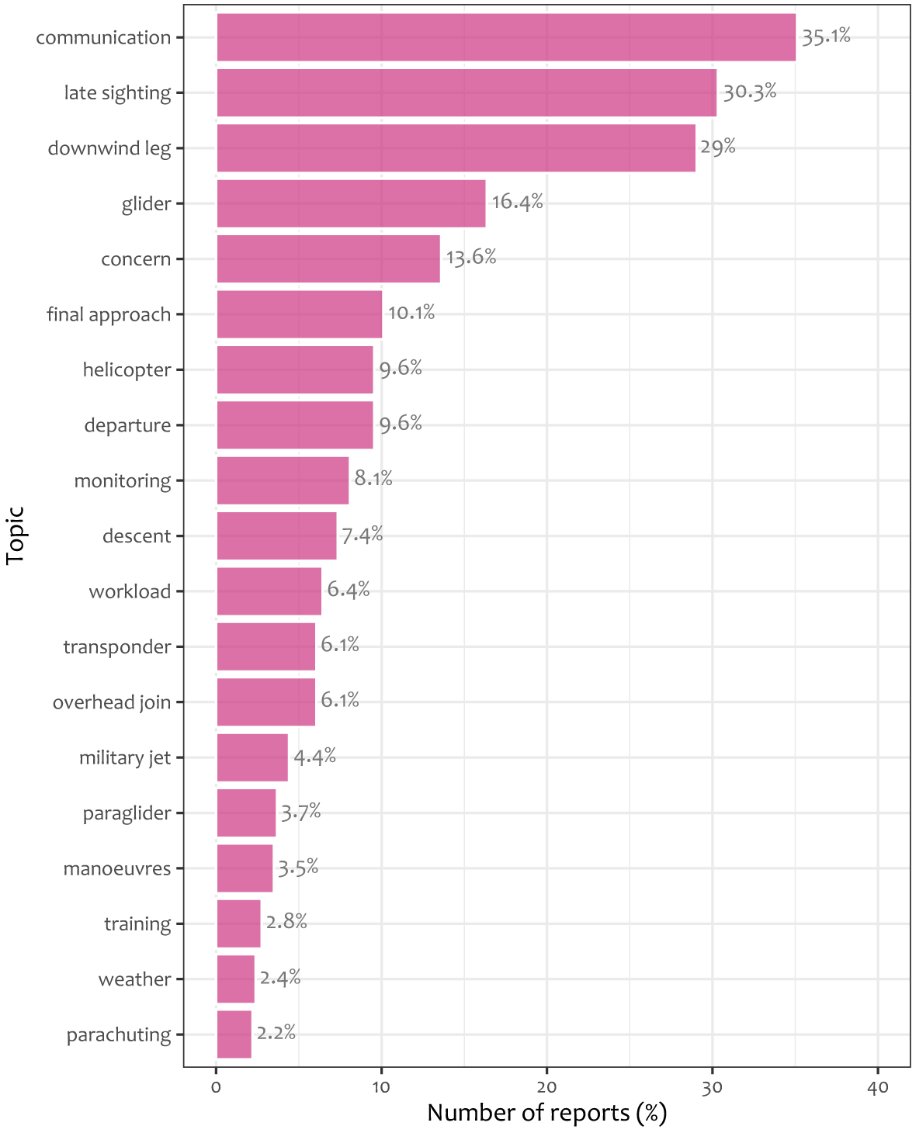


**Figure 3. Average prevalence of each of the 20 topics of Table 3 over UKAB reports.**

what the authors of the text considered important when writing the report.

### 6.2 Automatic extraction of TOKAI taxonomy factors

While topic modelling enabled a higher-resolution insight into the reports with respect to the simple descriptive statistics, knowledge expressed by topics can still be vague and potentially misleading (e.g., while the topic "workload" can intuitively be assumed to appear only when workload was high, "communication" for example can suggest very different scenarios, ranging from lack of communication to perfect communication). The exploitation of syntactic analysis represents an attempt to dig deeper and extract even more precise information.

Syntactic analysis (see Section 5.3) is a powerful tool to identify the text structure and meaning, and, in particular, in this work it enabled the association of each considered report to the corresponding TOKAI taxonomy factors. Specifically, for the purposes of this research, only Part A of the TOKAI taxonomy was exploited (i.e., the one related to the personnel), since the actions reported in the paragraphs of interest from both the CEANITA and the UKAB reports are usually more related

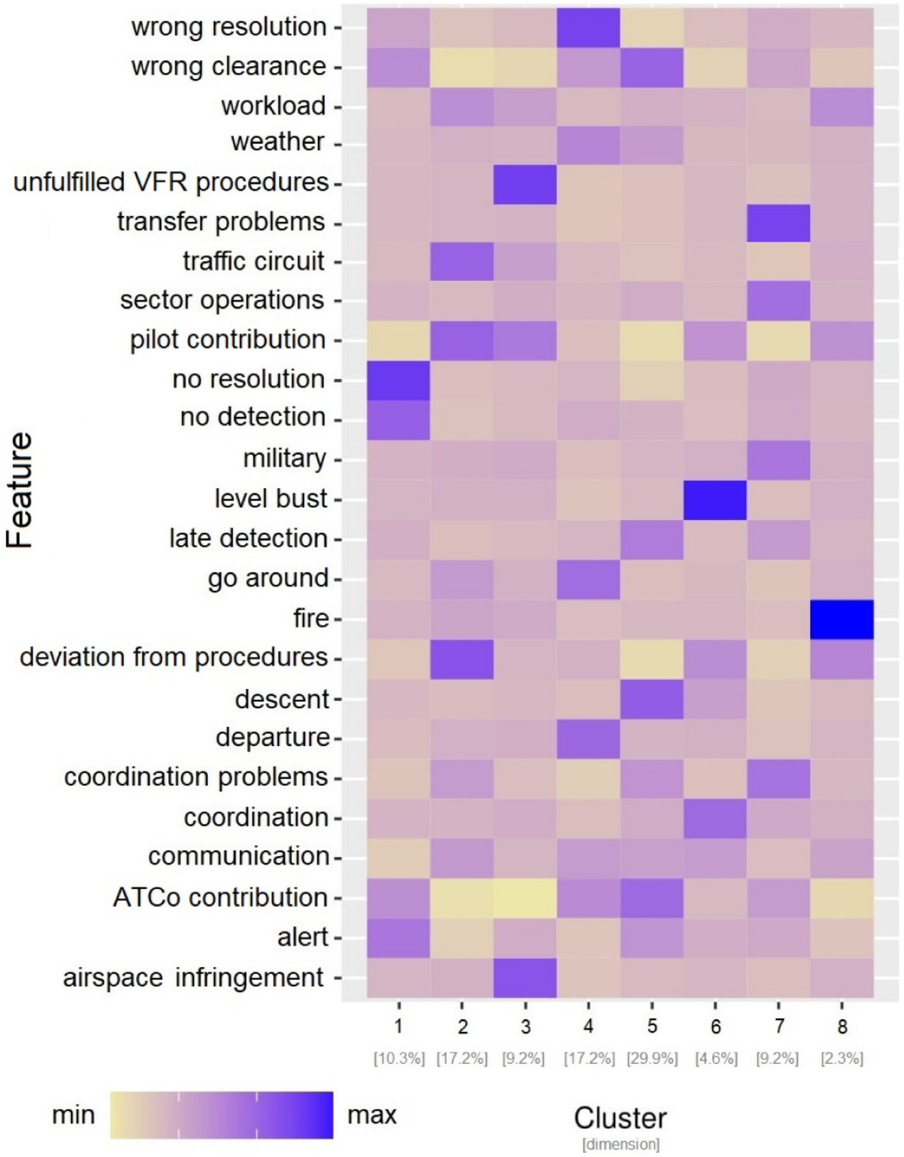


**Figure 4. Characterisation of the 8 clusters in CEANITA reports through the size and the normalised relative frequency of each feature.**

to this subject. To have a clearer picture of the considered taxonomy factors, Table 4 reports Part A of the TOKAI taxonomy together with factors' specifications[11] and examples of sentences associated to them through the algorithm developed by the authors.

The algorithm to link the report text to the TOKAI taxonomy factors can now be illustrated in detail (see Algorithm 1).

***6.2.1 Outcomes on CEANITA reports***. The proposed algorithm (Algorithm 1) has then been applied to the conclusive section of each CEANITA report. This portion of text summarises the dynamics of the LoS based on the main actions of pilots and ATCos. There follows an example from report 067/18 (translated from Spanish to English):

> "According to what stated by the executive controller of ACC Barcelona Sector CCC, different conflicts

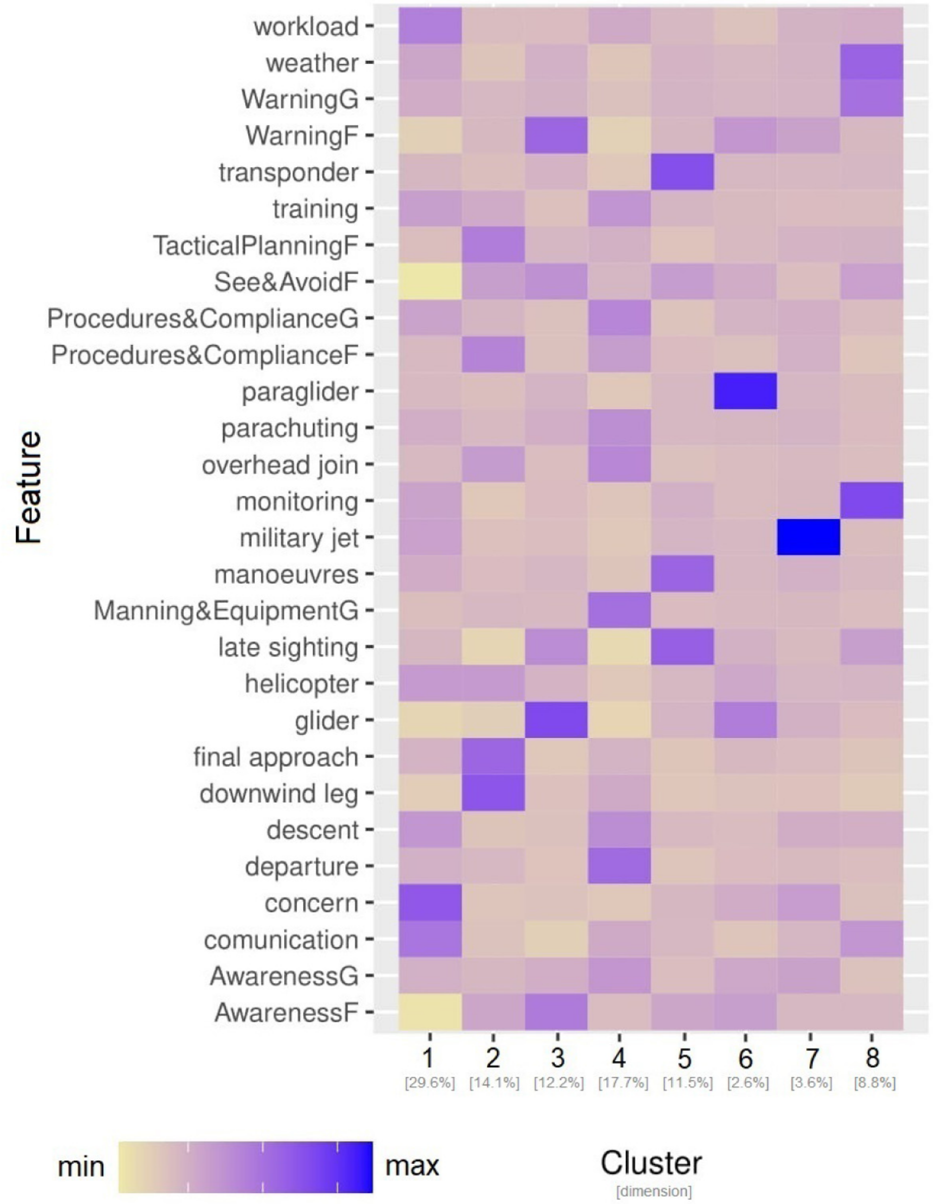


**Figure 5. Characterisation of the 8 clusters in UKAB reports through the size and the normalised relative frequency of each feature.**

were going on: due to this, he did not plan the descent of Aircraft 1, which ended up in conflict with the trajectory of Aircraft 2. When Aircraft 1 asked for descent, he did not check if the traffic around Aircraft 1 was in potential conflict with it and authorised it to descend at FL310. This produced the loss of separation between Aircraft1 and Aircraft 2. The planning controller of Sector CCC immediately informed the executive controller of the conflict, but this did not prevent the airprox.

On the other hand, Sector CCC provided incomplete traffic information to Aircraft 2."

**Table 4. Part A of the TOKAI taxonomy factors: specifications and examples of sentences associated to the taxonomy by the developed tool.**

| Factor | Specifications | Example |
|---|---|---|
| A-1. Perception | See - identification; See - detection; Hear - identification; Hear - detection; Perceive visual information - accuracy; Perceive auditory information - accuracy. | Sector CAO authorised aircraft 1 without detecting aircraft 2. |
| A-2. Memory | Remember to monitor or check; Remember to act; Remember previous actions; Recall information from working memory; Recall information from long-term memory. | Aircraft 2 was authorised by the Sector, not remembering presence of Aircraft 1. |
| A-3. Decision | Judge/Project; Decide/Plan. | APP LEMG planned the approximation sequence incorrectly. |
| A-4. Action | Select/Position manually; Convey/Record information. | Aircraft 1 did not communicate its position correctly. |
| A-5. Conformance | Deliberate or malicious act; Individual conformance with rules or procedures; Team conformance with rules or procedures. | Aircraft 2 did not comply with the instruction. |

**Algorithm 1: Algorithm to link the report text to the TOKAI taxonomy factors exploiting Syntactic Analysis**

**Input: 1.** The sequences of verbs/actions in the base form for each factor (e.g., for factor A-1, the list "see", "identify", "detect", "hear", etc.). This sequences can be created directly by human operators, which can be supported by automatic tools. Possibly, two sequences can be created for each factor, a positive and a negative one (e.g., for A-2, "remember" is in the positive sequence, while "forget" in the negative one).

**2.** The text of the conclusive section of the report of interest.

**Output:** For each of the factors (i.e., A-1, A-2, etc. in Table 4) and for each subject (e.g., pilot or controller) the number of positive and negative occurrences.

The text of the report is processed via UDPipe (see Section 5.3 and Table 2 as reference);

In the UDPipe output (i.e., the result of lemmatisation, part-of-speech tagging, and dependency parsing) we search, for each of the factors, the verbs in factor's lists (both for the positive and negative lists);

For each of the identified verb, the subject is retrieved, also taking into account passive forms where the subject is the agent;

A check for negative forms or adverbs (e.g., "incorrectly") is performed in the identified sentence to cope with the inversion of meaning (i.e., positive verbs become negative if a negative form or adverb is present);

The application of Algorithm 1 on these sections of CEANITA reports produced a rich output (i.e., multiple TOKAI factors were found in most of the papers). To have a more comprehensive look at the results, factors' subjects can be grouped into flight elements (i.e., aircraft, pilot, etc.) and ground elements (i.e., controller, sector, etc.); it is then possible to estimate for each CEANITA report how the five factors are distributed between the subjects, both in terms of positive and negative occurrences. Please note that in this section the words "negative" and "positive" should be intended in a purely linguistic sense (i.e., "he did not perform an action" is negative and "he performed an action" is positive, independently of the positive or negative impact or evaluation of that action).

Figure 6 shows the global distribution (i.e., the sum over the different reports) of negative occurrences of each TOKAI-taxonomy factor by group of subjects. Figure 6 suggests that the main omissions — namely, the actions reported in negative form in the text — for the flight subjects are classified as factor A-4 (i.e., problems with action) and A-5 (i.e., problems with conformance with rules), while for the ground subjects they are again mostly classified as factor A-4 and, to a lesser extent, as factor A-1 (i.e., problems with perception). Interestingly, further analysing the data, it is possible to discover that almost all the problems with factor A-4 are relative to (lack of) conveyance of information, for both flight and ground elements.

To assess the reliability of the proposed syntactic-analysis-based algorithm (Algorithm 1) on CEANITA reports, an indirect validation was performed: indeed, it could not be validated in the standard way since there is no ground truth i.e., the correct classification in terms of TOKAI taxonomy does not exist for the considered reports. Thus, a simple predictive model was developed to predict the main contribution (ATCo or pilots) in a LoS, based on the extracted number of positive and negative occurrences of each taxonomy factor (i.e., the output of Algorithm 1). The idea behind this validation is that a good performance of this predictive model would indicate that the extracted information is reasonably accurate, since TOKAI taxonomy factors should well describe the ATCo's and pilots' contribution to the event. Specifically, for each LoS, the goal was to predict:

- the pilots' contribution, i.e., classified as direct or not;
- the ATCos' contribution, i.e., classified as direct or not;

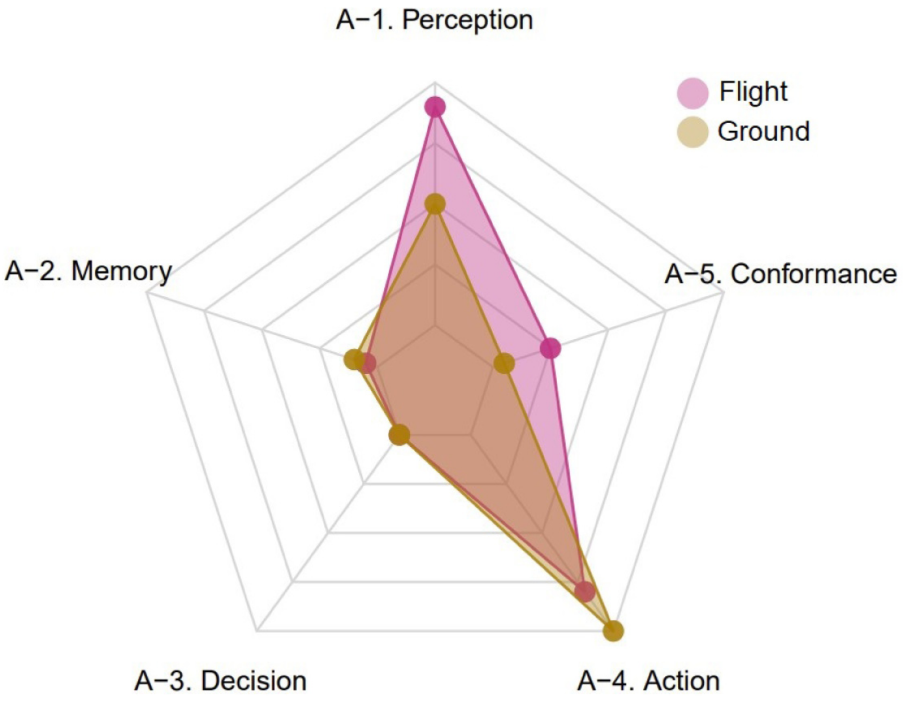


**Figure 6. Global distribution of negative occurrences of each TOKAI-taxonomy factor by group of subjects (CEANITA reports).**

based on:

- the number of positive and negative occurrences of each taxonomy factor (the outputs of Algorithm 1);
- the differences in prevalence between flight and ground subjects for each taxonomy factor;
- the airspace class (in fact, similar behaviours of ATCo and pilots can lead to different contribution assessments in different airspace classes, due to different regulations).

Note that ATCo and pilots can be both indicated in the reports as directly responsible to the incident.

Table 5 and Table 6 report the confusion matrices of the developed predictive models, developed according to what described in Section 5.4).

Confusion matrices in Table 5 and Table 6 appear reasonably balanced, especially considering that the classes are highly unbalanced. The global accuracy of the prediction is ≈83% for pilots contribution and ≈85% for ATCo contribution. Therefore, it can be stated that:

- the proposed approach is able to automatically link each CEANITA report to the TOKAI taxonomy factors exploiting syntactic analysis;

**Table 5. Confusion matrices (%) on the dummy predictive problem of estimating pilots' direct contribution based on outputs of Algorithm 1) via SVM to validate Algorithm 1 on CEANITA reports.**

| | | **Pred.** | |
|---|---|---|---|
| | | **No** | **Yes** |
| **Truth** | **No** | 51.6±0.1 | 12.4±0.1 |
| | **Yes** | 4.5±0.3 | 31.5±0.3 |

**Table 6. Confusion matrices (%) on the dummy predictive problem of estimating ATCo's direct contribution based on outputs of Algorithm 1) via SVM to validate Algorithm 1 on CEANITA reports.**

| | | **Pred.** | |
|---|---|---|---|
| | | **No** | **Yes** |
| **Truth** | **No** | 25.8±0.2 | 3.4±0.2 |
| | **Yes** | 11.2±0.2 | 59.6±0.2 |

- the indirect validation performed through a dummy prediction problem showed promising performance supporting the quality of the proposed approach;
- as a side result of this indirect validation, the extracted link between CEANITA reports and TOKAI taxonomy appears to be a good proxy of the contribution assessment.

***6.2.2 Outcomes on UKAB reports***. Algorithm 1 has been further applied to the conclusive sections of some UKAB reports. Indeed, as the structure of UKAB reports has evolved during the years, only a subsample of them (in particular, the reports written in 2017 and a small part of the 2018 ones — the complete list is available in the output dataset[35]) contains this free-text summary of the contributory factors.

An example of the analysed text from report 2017002 reads:

> "The radar controllers did not issue timely Traffic Information. The Tac Right controller's workload was such that he was distracted and did not sufficiently monitor the F15s. The F15 crews were not aware that AARA8 was active."

Despite the fact that the considered text in UKAB reports is quite different from the one in CEANITA reports, the output of Algorithm 1 is again of great interest. In particular, Figure 7 shows the global distribution of negative occurrences of each TOKAI-taxonomy factor in the selected UKAB reports by group of subjects. Figure 7 suggests that, similarly to the CEANITA use case, the most frequent factor is factor A-4, for both pilots and ATCos, while factors A-2 and A-3 are the least mentioned and factor A-5 is mostly associated with the flight subjects. Nevertheless, there is a huge difference in the prevalence of factor A-1: indeed, in UKAB reports problems with perception seem to be reported much more often, in particular for flight subjects. The comparison between Figure 6 and Figure 7 reveals another interesting detail: the shape of the yellow polygon — corresponding to the ground subjects — is almost identical in both the graphs, while the pink one differs essentially for the peak in A-1. This seems in line with what emerged in the data description (Section 4), as the two samples of reports are very different in terms of contribution assessment: in the CEANITA corpus, the majority of the LoS events is associated to ATCo's contribution, while, in the UKAB sample, the safety barriers suggest pilots' actions are assessed as contributory in the vast majority of the cases and, in particular, they are mostly classified as Situational Awareness and See & Avoid problems, which may indeed be related with A-1 factor in the TOKAI taxonomy.

Analogously to the CEANITA use case, to assess the reliability of the proposed Syntactic-Analysis-based algorithm (Algorithm 1) on UKAB reports a simple predictive model was developed. Since UKAB reports do not contain the same neat indication of ATCos' or pilots' contribution, two similar variables considered:

- the presence of flight safety barriers, roughly corresponding to pilots' contribution;
- the presence of ground safety barriers, roughly corresponding to ATCos' contribution;

according to their definition in Figure 1. Similarly to the CEANITA scenario, the predictors were based on the Algorithm 1 output and on the airspace class.

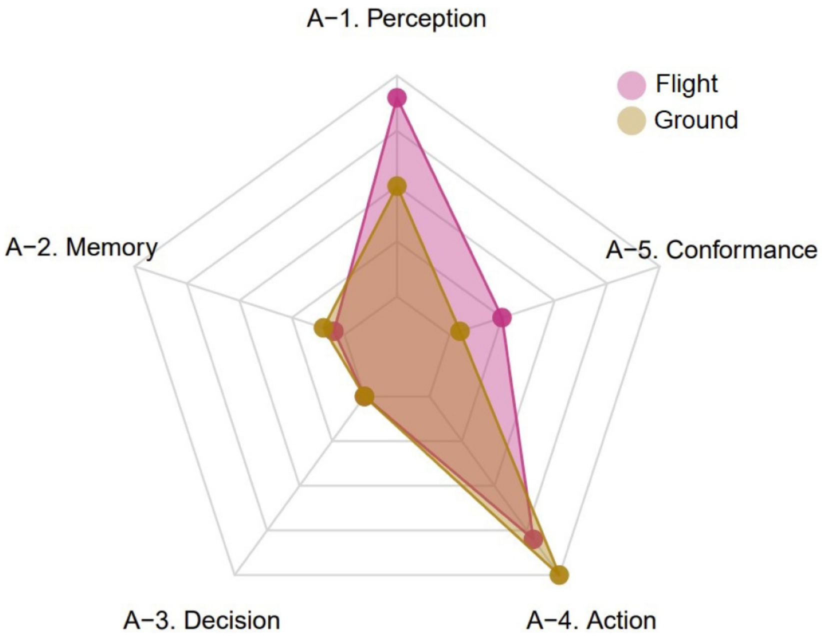


**Figure 7. Global distribution of negative occurrences of each TOKAI-taxonomy factor by group of subjects (UKAB reports).**

Table 7 and Table 8 report the confusion matrices of the predictive models trained on the 127 UKAB reports (developed analogously to the CEANITA use case).

Also in this case, despite the classes being highly unbalanced, the confusion matrices appear quite balanced. The global accuracy of the prediction is ≈80% for pilots' contribution and ≈76% for ATCos' contribution.

When looking at the accuracy of CEANITA and UKAB validation models, it is fundamental not to consider these numbers completely comparable, i.e., the lower accuracy of the UKAB model is not necessarily associated with a lower accuracy in the outcome of Algorithm 1 on the UKAB sample. Indeed, by reading some of the reports, it is evident that, while the CEANITA conclusive text is strictly associated with the final contribution assessment, the text considered in the UKAB reports does not focus exactly on the same aspects evaluated in the Safety Barriers assessment. Therefore, the fact that the TOKAI factors extracted from the UKAB reports are less predictive than those ones extracted from the CEANITA reports in terms of pilots' and ATCos' contribution might be due to the different settings of the two dummy predictive problems exploited for validation purposes.

## 7 Discussion and conclusions

The objective of this work was to facilitate the extraction of meaningful and actionable information from LoS reports and, in particular, to identify recurrent behaviours and precursors. Therefore, the authors proposed an approach based on (i) an EDA and (ii) an automatic classification of extracted knowledge considering a state-of-the-art safety taxonomy (the TOKAI one). The approach was tested on the LoS events reported in the CEANITA and UKAB public databases.

**Table 7. Confusion matrices (%) on the dummy predictive problem of estimating pilots' direct contribution based on outputs of Algorithm 1) via SVM to validate Algorithm 1 on UKAB reports.**

| | | Pred. | |
|---|---|---|---|
| | | **No** | **Yes** |
| **Truth** | **No** | 9.1±0.5 | 9.1±0.5 |
| | **Yes** | 10.6±0.5 | 71.2±0.5 |

**Table 8. Confusion matrices (%) on the dummy predictive problem of estimating ATCo's direct contribution based on outputs of Algorithm 1) via SVM to validate Algorithm 1 on UKAB reports.**

| | | Pred. | |
|---|---|---|---|
| | | **No** | **Yes** |
| **Truth** | **No** | 15.2±0.3 | 12.1±0.3 |
| | **Yes** | 12.1±0.3 | 60.6±0.3 |

For EDA purposes, unsupervised NLP techniques were applied to identify latent topics. In addition, this exploration was complemented with a clustering analysis, which facilitated the identification and grouping of similar incidents. Results demonstrated the capacity of these techniques to effectively identify meaningful topics and group together incidents, finding eight different clusters, which were assessed as valid by domain experts. For the automatic extraction of the safety factors and their classification according to the TOKAI taxonomy, the authors leveraged syntactic analysis. This is a pioneering work in the field, and the results showed an understanding of the potential that these methods bring to safety analysis, also trying to keep in mind a resilience engineering perspective. Indeed, the classification of actions according to the TOKAI taxonomy (TOKAI factors are neither negatively nor positively oriented) goes in the direction of reframing of human behaviour not as a sequence of errors that lead to an undesired outcome (i.e., only pointing out where people went wrong), but as emergent from the system, arising as a function of complex interactions. The results of this classification were validated by demonstrating the strong connection between the factors identified and the main contributor to the incident.

Therefore, it can be said that the main objective of the work has been reached and the applicability of the approach has been proven on two very different samples. However, one of the major strengths of this work (i.e., the fact that information can be automatically extracted from different reports with different languages and narratives, independently on the context that generated them) somehow coincides with its biggest limitation: the proposed NLP tools rely only on the text they analyse, so that two different reports of the same exact incident would possibly generate two different outcomes. This means that, in essence, the factors that are identified as significant by these automated tools are not necessarily the ones with the most significant role in the considered incidents, but only the ones with the most significant role according to who wrote the reports. This limitation is nonnegligible as it is largely acknowledged that investigation reports are far from being standardised: not only are they strongly dependent on the ATM expertise, operational competences, training, backgrounds, and culture of the reporting organisations, but also inter-rater reliability issues appear to be significant, even when the reference background and taxonomy are aligned[36]. Furthermore, as a consequence, these NLP tools inherit part of the reports' safety culture in the process of identifying relevant information, making it difficult to maintain a resilience engineering perspective in the analysis.

Nevertheless, this feature paves the way for even more interesting applications of the proposed approach, including for instance the development of diagnostic tools to identify reports' narrative issues (e.g., the presence of expressions of blame culture or the absence of expected factors/topics in a reports' databases), the comparison between the reporting characteristics of different operators (e.g., pilots and ATCos), or the analysis of how reporting philosophy evolved in a certain period of time.In the future, these techniques could also be extended to other taxonomies or tailored to identify factors which should be included in the safety taxonomies, together with hidden sources of resilient performance (e.g., when not fulfilling a procedure resulted actually opportune[37]), based on their presence on the reports, and could help facilitating the analysis pointed out in 38.

## Data availability

### Underlying data

The reports considered in this study are collected in public databases.

The considered CEANITA reports are those classified as AIRPROX, ranging from 003_18 to 071_19. The reports are available at: https://www.seguridadaerea.gob.es/es/ambitos/gestion-de-la-seguridad-operacional/ceanita#Informes%20Definitivos. The considered UKAB reports those from 2017001 to 2019335. These are available at: https://www.airproxboard.org.uk/Reports-and-analysis/Monthly-summaries/Monthly-Airprox-reviews/.

Zenodo: irene-buselli/ORE2021_14040: ORE2021_14040 v1.0[35].

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

## References

1. Bergström J, Dekker SWA: **Bridging the Macro and the Micro by Considering the Meso: Reflections on the Fractal Nature of Resilience.** *Ecol Soc.* 2014; **19**(4): 22.
**Publisher Full Text**
2. SESAR Joint Undertaking: **European ATM master plan - executive view, 2015 edition.** 2015.
**Reference Source**
3. Tulechki N: **Natural language processing of incident and accident reports: application to risk management in civil aviation**. PhD thesis, Université Toulouse le Mirail-Toulouse II, 2015.
**Reference Source**
4. SESAR Joint Undertaking: **European ATM master plan - executive view, 2020 edition.** 2020.
**Reference Source**
5. Tanguy L, Tulechki N, Urieli A, *et al.*: **Natural language processing for aviation safety reports: From classification to interactive analysis.** *Comput Ind.* 2016; **78**: 80–95.
**Publisher Full Text**
6. Oza N, Castle JP, Stutz J: **Classification of aeronautics system health and safety documents.** *IEEE Trans Syst Man Cybern C Appl Rev.* 2009; **39**(6): 670–680.
**Publisher Full Text**
7. Switzer J, Khan L, Muhaya FB: **Subjectivity classification and analysis of the ASRS corpus**. In: *IEEE Int Conf Inf Reuse Integr*. 2011.
**Publisher Full Text**
8. Wolfe S: **Wordplay: an examination of semantic approaches to classify safety reports**. In *AIAA Infotech@Aerospace*. 2007.
**Publisher Full Text**
9. Persing I, Ng V: **Semi-supervised cause identification from aviation safety reports**. In *Joint Conference of the AnnualMeeting of the ACL and the International Joint Conference on Natural Language*. 2009; **2**: 843–851.
**Publisher Full Text**
10. Patriarca R, Cioponea R, Di Gravio G, *et al.*: **Managing Safety Data: the TOKAI Experience for the Air Navigation Service Providers.** *Transportation Research Procedia.* 2018; **35**: 148–157.
**Publisher Full Text**
11. Patriarca R, Di Gravio G, Cioponea R, *et al.*: **Safety intelligence: Incremental proactive risk management for holistic aviation safety performance.** *Safety Science.* 2019; **118**: 551–567.
**Publisher Full Text**
12. Ananyan S, Goodfellow M: **Example application of PolyAnalyst with IATA STEADES data**. 2004.
**Reference Source**
13. Kuhn KD: **Using structural topic modeling to identify latent topics and trends in aviation incident reports.** *Transp Res Part C Emerg Technol.* 2018; **87**: 105–122.
**Publisher Full Text**
14. Irwin WJ, Robinson SD, Belt SM: **Visualization of large-scale narrative data describing human error.** *Hum Factors.* 2017; **59**(4): 520–534.
**PubMed Abstract** | **Publisher Full Text**
15. Robinson SD: **Temporal topic modeling applied to aviation safety reports: A subject matter expert review.** *Safety Science.* 2019; **116**: 275–286.
**Publisher Full Text**
16. Sjöblom O: **Data mining in promoting aviation safety management**. In *International Conference on Well-Being in the Information Society*. 2014; **450**: 186–193.
**Publisher Full Text**
17. Robinson SD, Irwin WJ, Kelly TK, *et al.*: **Application of machine learning to mapping primary causal factors in self reported safety narratives.** *Safety Science.* 2015; **75**: 118–129.
**Publisher Full Text**
18. Blei DM, Lafferty JD: **Topic models**. *Text mining: classification, clustering, and applications*. 2009; **10**(71): 34.
**Reference Source**
19. Duran BS, Odell PL: **Cluster analysis: a survey**. Springer Science & Business Media, 2013.
**Reference Source**
20. Straka M, Straková J: **Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe**. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. 2017.
**Publisher Full Text**
21. International Civil Aviation Organization: **Doc 4444: Air traffic management**. 2016.
**Reference Source**
22. UK Airprox Board: **Uk airprox board (ukab) factsheet**. 2016.
**Reference Source**
23. International Civil Aviation Organization: **Annex 11: Air traffic services**. 2001.
**Reference Source**
24. Das S, Dixon K, Sun X, *et al.*: **Trends in transportation research: Exploring content analysis in topics.** *Transp Res Rec.* 2017; **2614**(1): 27–38.
**Publisher Full Text**
25. Sun L, Yin Y: **Discovering themes and trends in transportation research using topic modeling.** *Transp Res Part C Emerg Technol.* 2017; **77**: 49–66.
**Publisher Full Text**
26. Murtagh F, Contreras P: **Algorithms for hierarchical clustering: an overview.** *wiley Interdiscip Rev Data Min Knowl Discov.* 2012; **2**(1): 86–97.
**Publisher Full Text**
27. Mirkin B: **Choosing the number of clusters.** *wiley Interdiscip Rev Data Min Knowl Discov.* 2011; **1**(3): 252–260.
**Publisher Full Text**
28. Shawe-Taylor J, Cristianini N: **Kernel methods for pattern analysis.** Cambridge university press, 2004.
**Publisher Full Text**

29. Fernández-Delgado M, Cernadas E, Barro S, *et al.*: **Do we need hundreds of classifiers to solve real world classification problems?** *J Mach Learn Res.* 2014; **15**(1): 3133–3181.
**Reference Source**

30. Wainberg M, Alipanahi B, Frey BJ: **Are random forests truly the best classifiers?** *J Mach Learn Res.* 2016; **17**(1): 3837–3841.
**Reference Source**

31. Goodfellow I, Bengio Y, Courville A: **Deep learning.** MIT press Cambridge, 2016.
**Reference Source**

32. Keerthi SS, Lin CJ: **Asymptotic behaviors of support vector machines with gaussian kernel.** *Neural Comput.* 2003; **15**(7): 1667–1689.
**PubMed Abstract** | **Publisher Full Text**

33. Oneto L: **Model Selection and Error Estimation in a Nutshell.** Springer, 2020.
**Publisher Full Text**

34. Shalev-Shwartz S, Ben-David S: **Understanding machine learning: From theory to algorithms.** Cambridge university press, 2014.
**Publisher Full Text**

35. Oneto L, Buselli I: **irene-buselli/ORE2021_14040: ORE2021_14040 v1.0.** (v1.0). [Data set]. *Zenodo.* 2021.
**http://www.doi.org/10.5281/zenodo.5503831**

36. Olsen NS, Shorrock ST: **Evaluation of the HFACS-ADF safety classification system: inter-coder consensus and intra-coder consistency.** *Accid Anal Prev.* 2010; **42**(2): 437–444.
**PubMed Abstract** | **Publisher Full Text**

37. NTSB: **Aircraft accident report NTSB/AAR-19/03**. 2018.
**Reference Source**

38. CANSO: **Incidents investigation toolbox**. 2021.
**Reference Source**

# Open Peer Review

## Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 15 March 2022

https://doi.org/10.21956/openreseurope.15560.r28603

✓ **Arie Adriaensen**
Centre for Industrial Management/Traffic and Infrastructure, KU Leuven, Leuven, Belgium

I thank the authors for their answers to the review and for the revision of the manuscript. I am positive about the resulting additions and clarifications. Limitations of the study are now better clarified. At the same time the methodologic considerations have better been explained by additional specifications about the choice of parameters. This leads to a better defence of the final choices in the methodology.

I am looking forward to the application of the proposed approach to more reporting data. As the authors propose, there can also be benefit in identifying forthcoming taxonomy extensions or yet-to-be-identified factors for future safety taxonomies.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Socio-technical modelling; Human-Robot interaction; Resilience Engineering; Human Factors; Aviation Safety

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 21 Mar 2022

**Luca Oneto**, ZenaByte, Genova, Italy

Dear Arie Adriaensen, thank you very much for appreciating our effort toward the improvement of our work. We are already trying to find a way to get more data and test our proposal.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 14 March 2022

https://doi.org/10.21956/openreseurope.15560.r28604

✓

**Riccardo Patriarca**
Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Rome, Italy

I truly thank the authors for their critical reflections and reconsiderations of their manuscript in light of my comments. I appreciate your efforts, and critical appraisal of the study.

My overall impression on the revised version is positive. I agree with the frank observations by the authors about the preliminary nature of this research, and with their comment to the revised version: "It is fair to say that the NLP work package tried not to disregard the RE philosophy and to deal with complexity as much as possible, but found it often difficult to conjugate with the reporting philosophy of the data sources."

I do believe the steps suggested in this work may motivate an extension of the repertoire of available approaches to enhance aviation safety investigations and I do look forward to future research in this area.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Aeronautical engineering, air traffic management, system safety, safety risk management, resilience engineering, socio-technical systems, accident analysis, aviation safety

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 14 Mar 2022

**Luca Oneto**, ZenaByte, Genova, Italy

Dear Riccardo Patriarca, thank you very much for appreciating our effort toward the improvement of our work. We are already thinking about possible future evolutions that we hope can further improve and empower aviation safety investigations.

***Competing Interests:*** No competing interests were disclosed.

## Version 1

Reviewer Report 15 October 2021

https://doi.org/10.21956/openreseurope.15131.r27674

?

**Arie Adriaensen**
Centre for Industrial Management/Traffic and Infrastructure, KU Leuven, Leuven, Belgium

Dear authors and editorial team,

First of all I would like to complement the authors on the work they have performed, which shows a proposal for a novel and systematic methodology to produce meaningful interpretations out of large quantities of available reporting data. I agree that there is a general consensus within the aviation and academic community, that ironically the success of aviation reporting from the last decades also carries the responsibility of more meaningful data processing. Hence, the research topic is relevant.

**Concerns before Indexing**
My main concern is on the relation between scope, research data and how interrelatedness might partially jeopardize the research objectives stated by the authors.

The FARO project stands for saFety And Resilience guidelines for aviation and the manuscript's rationale also refers to resilience principles. The use of resilience is aligned with the fact that, as the authors state, "today's safety events are gradually becoming of a more complex and uncommon nature than those of yesteryear"; "growing amount of flights has ultimately led to increased operational complexity", and; "increased complexity of the ATM system should be absorbed by increased deployment of automation solutions in order to achieve a more efficient and safe traffic management".

I agree with the authors that in order to overcome the limitation of the taxonomic approach currently used for the processing of aviation reporting, unsupervised techniques like topic modelling and similarity clustering derived from NLP can provide a great way forward. Nevertheless, the discovery of unknown patterns or further knowledge can only be as rich as the data contained in the free text of the reports. This limitation in data collection from incident investigation is typically referred to as the What-You-Look-For-Is-What-You-Find (WYLFIWYF) – principle (Lundberg 2009; Le Coze 2013). It should be considered that what Pilots and ATCos report is also the social product of the safety causation views traditionally used in the aviation community. So far, pilots are largely trained to think about safety in terms of Swiss Cheese-based & HFACS-based causation models, which are essentially expressions of safety-barrier thinking. These can be valuable models and concepts, but do not belong to the resilience engineering paradigm.
The pilot community concepts do not necessarily produce data that are aligned with more recently developed causation models in which safety is seen as a control problem (See Leveson, 2009) or as the emergent product of functional resonance between different actions and agents (Hollnagel,

2012) to just name a few causation views that would generate different reporting data. Likewise, the understanding from system dynamics due to increased deployment of automation and increased complexity of the ATM system might not typically be covered by reports from pilots and ATCos. This is a limitation of every reporting system where typically front-end operators are expected to report, but just mentioned here to highlight the limitation of (and the impact from) this specific data source to reach a specific research objective.

EUROCONTROL has been at the forefront of incentivising resilience-principles for many years. It could be expected that this has influenced ATCos to report differently than the pilot community, although this outside of the scope of this study.

The arguments above are the reason why in my opinion the authors might not have reached the full objectives expressed at the beginning of the manuscript, it is still fair to say that a partial objective is reached by having increased the learning potential from existing reporting data. I suggest he authors to adjust the objectives in line with the chosen methodology. It is essentially only the process of data interpretation, which is covered by the methodology, without scrutinizing how that data could already have been influenced by the way the reporting system is constructed. Although this limitation would be true for any interpretation of reporting data, it is important in the light of reaching an increased understanding of resilience, increased operational complexity and increased deployment of automation solutions.

The chosen data is readily available and accessible, which is very helpful for readers and reviewers.

**METHODOLOGY**
The authors have done a great job in complementing different data processing techniques to reach a multi-perspective analysis of a fixed set of reports.

The created algorithm reveals a possible limitation of specific safety causation models, which previously has been addressed in the data gathering. While it is certainly preferable to differentiate occurrences between positive and negative occurrences of each taxonomy factor as the authors have done, certain causation views would even go further and say that some occurrences are neither positive or negative, but only produce positive or negative effects depending on the surrounding elements in the system. The functional resonance principle as described by Hollnagel (2012) is such an example. This is not to say, that this principle necessarily needs to be used because it essentially dismisses a taxonomic approach, but it does highlight the possibility that certain words might also be neutral or contextual. Whereas workload intuitively is only used when workload was high, communication for example can be very descriptive and produce systemic effects that are harder to capture. Are there any limitations that appeared during the clustering techniques used, which were not captured or misinterpreted by NLP techniques?

In the light of methodological choices, it would be beneficial if the authors could defend their choice of single words and bigrams in the clustering analysis. It is not clear if the authors considered other options like for example proximity search clustering, or other additional techniques like exclusion criteria or weighing for topological modelling.

It is interesting to see that the words, bigrams and resulting topics from the cluster analysis reveal clear differences and similarities between the CEANITA and UKAB reports. Have the researchers hypothesised why these differences appear? Is It believed to be due to a difference in the reporting culture, or in the actual causal relations that triggered he reports? More importantly, could at any point anything be learned from the difference between the CEANITA and UKAB outputs to improve the NLP learning?

My choice to only assign a partial pass for the statistical analysis and its interpretation is only about the fact that more options could be described or the chosen options could be more clearly defended.

**ADDITIONAL REMARK**

It seems that the polygons for Figure 6 (CEANITA reports) and Figure 7 (UKAB reports) are identical. Although the manuscript announced that the polygons are nearly identical for ground subjects, it also announces and essential difference (a peak in A1) for flight subjects. Could the authors verify if one figure has accidentally been used for both regions?

**References**

1. Hollnagel E: FRAM, the Functional Resonance Analysis Method: Modelling Complex Socio-technical Systems. *Routledge*. 2012.
2. Le Coze J: What have we learned about learning from accidents? Post-disasters reflections. *Safety Science*. 2013; **51** (1): 441-453 Publisher Full Text
3. Leveson N: A new accident model for engineering safer systems. *Safety Science*. 2004; **42** (4): 237-270 Publisher Full Text
4. Lundberg J, Rollenhagen C, Hollnagel E: What-You-Look-For-Is-What-You-Find – The consequences of underlying accident models in eight accident investigation manuals. *Safety Science*. 2009; **47** (10): 1297-1311 Publisher Full Text

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and does the work have academic merit?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Socio-technical modelling; Human-Robot interaction; Resilience Engineering; Human Factors; Aviation Safety

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 24 Jan 2022

**Luca Oneto**, ZenaByte, Genova, Italy

**Dear Arie Adriansen, we thank you for your valuable comments that allowed us to improve the manuscript. What follows is the pointwise response to your comments with the corresponding changes in the new version of the manuscript. (Reviewer comments in italics)**

*Dear authors and editorial team, First of all I would like to complement the authors on the work they have performed, which shows a proposal for a novel and systematic methodology to produce meaningful interpretations out of large quantities of available reporting data. I agree that there is a general consensus within the aviation and academic community, that ironically the success of aviation reporting from the last decades also carries the responsibility of more meaningful data processing. Hence, the research topic is relevant.*

**We thank the reviewer for their positive comments. The following is a pointwise response to your comments with the corresponding changes in the new version of the manuscript.**

*My main concern is on the relation between scope, research data and how interrelatedness might partially jeopardize the research objectives stated by the authors.*
*The FARO project stands for saFety And Resilience guidelines for aviation and the manuscript's rationale also refers to resilience principles. The use of resilience is aligned with the fact that, as the authors state, "today's safety events are gradually becoming of a more complex and uncommon nature than those of yesteryear"; "growing amount of flights has ultimately led to increased operational complexity", and; "increased complexity of the ATM system should be absorbed by increased deployment of automation solutions in order to achieve a more efficient and safe traffic management". I agree with the authors that in order to overcome the limitation of the taxonomic approach currently used for the processing of aviation reporting, unsupervised techniques like topic modelling and similarity clustering derived from NLP can provide a great way forward. Nevertheless, the discovery of unknown patterns or further knowledge can only be as rich as the data contained in the free text of the reports. This limitation in data collection from incident investigation is typically referred to as the What-You-Look-For-Is-What-You-Find (WYLFIWYF) – principle (Lundberg 2009; Le Coze 2013). It should be considered that what Pilots and ATCos report is also the social product of the safety causation views traditionally used in the aviation community. So far, pilots are largely trained to think about safety in terms of Swiss Cheese-based & HFACS-based causation models, which are essentially expressions of safety-barrier thinking. These can be valuable models and concepts, but do not belong to the resilience engineering paradigm. The pilot community concepts do not necessarily produce data that are*

*aligned with more recently developed causation models in which safety is seen as a control problem (See Leveson, 2009) or as the emergent product of functional resonance between different actions and agents (Hollnagel, 2012) to just name a few causation views that would generate different reporting data. Likewise, the understanding from system dynamics due to increased deployment of automation and increased complexity of the ATM system might not typically be covered by reports from pilots and ATCos. This is a limitation of every reporting system where typically front-end operators are expected to report, but just mentioned here to highlight the limitation of (and the impact from) this specific data source to reach a specific research objective. EUROCONTROL has been at the forefront of incentivising resilience-principles for many years. It could be expected that this has influenced ATCos to report differently than the pilot community, although this outside of the scope of this study. The arguments above are the reason why in my opinion the authors might not have reached the full objectives expressed at the beginning of the manuscript, it is still fair to say that a partial objective is reached by having increased the learning potential from existing reporting data. I suggest the authors to adjust the objectives in line with the chosen methodology. It is essentially only the process of data interpretation, which is covered by the methodology, without scrutinizing how that data could already have been influenced by the way the reporting system is constructed. Although this limitation would be true for any interpretation of reporting data, it is important in the light of reaching an increased understanding of resilience, increased operational complexity and increased deployment of automation solutions.*

**Author Response: We concur with these considerations. We acknowledge these limitations and concur with the fact that they were indeed not explicit enough on the paper. Action Undertaken: This subject has now been mentioned in the Introduction, discussed further when commenting in the Results (end of Section 6.1) and then more thoroughly discussed in the Discussion and Conclusions (Section 7)**.

*The chosen data is readily available and accessible, which is very helpful for readers and reviewers. The authors have done a great job in complementing different data processing techniques to reach a multi-perspective analysis of a fixed set of reports.*

**Author Response: We thank the reviewer for his positive comments.**

*The algorithm reveals a possible limitation of specific safety causation models, which previously has been addressed in the data gathering. While it is certainly preferable to differentiate occurrences between positive and negative occurrences of each taxonomy factor as the authors have done, certain causation views would even go further and say that some occurrences are neither positive or negative, but only produce positive or negative effects depending on the surrounding elements in the system. The functional resonance principle as described by Hollnagel (2012) is such an example. This is not to say that this principle necessarily needs to be used because it essentially dismisses a taxonomic approach, but it does highlight the possibility that certain words might also be neutral or contextual.*

**Author Response: Given that the matter raised by the reviewer is clear and relevant, this is just to clarify that the classification of taxonomy-factor occurrences as positive and negative has to be seen, and is intended, in a purely “grammatical” sense: “he did not comply with procedures” is classified as negative and “he followed the instruction” as positive accordingly to the verb form. However, this is not to any extent an evaluation or judgement on the positive or negative value of the action or of its effects: for instance, not fulfilling a procedure may result in a positive outcome (as**

**pointed out in the Conclusions) and in general any evaluation of a behaviour strictly depends on the context [1]. This does not completely solve the problem as, of course, when interpreting a graph of this kind the natural tendency is to make the grammatical meaning coincide with the "effective" one, but it has to be borne in mind that this may indeed not be the case. Action Undertaken: This clarification has now been added in the paper (Section 6.2.1)**

*Whereas workload intuitively is only used when workload was high, communication for example can be very descriptive and produce systemic effects that are harder to capture. Are there any limitations that appeared during the clustering techniques used, which were not captured or misinterpreted by NLP techniques?*
**Author Response: This limitation is acknowledged and the syntactic analysis part (to link reported actions and TOKAI factors) was intended exactly to try and overcome this limitation: while "communication" is neutral, identifying TOKAI factors like "the pilot did not communicate the position" or "the ATCo correctly provided traffic information" gives a higher level of information. Of course, this raises the previously mentioned problem about how to consider the impact of negative/positive sentence in a proper way, as already pointed out. Action Undertaken: A short paragraph about this has been added at the beginning of Section 6.2.**

*In the light of methodological choices, it would be beneficial if the authors could defend their choice of single words and bigrams in the clustering analysis. It is not clear if the authors considered other options like for example proximity search clustering, or other additional techniques like exclusion criteria or weighing for topological modelling.*
**Author Response: A number of preliminary analyses were performed to understand the relevance and usefulness of different n-grams to describe the reports' content (both per se and in the topic-modelling framework), and in both CEANITA and UKAB use cases the role of n-grams with n>2 resulted substantially negligible per se and source of additional noise. As the reviewer suggests, we considered other (often more complex) alternatives to the proposed one, but the increasing complexity was not balanced by actual differences in the results, so we tried to keep the approach as simple as possible in order to easily allow for inspection and modification. Action Undertaken: This clarification has now been added in Section 5.1.**

*It is interesting to see that the words, bigrams and resulting topics from the cluster analysis reveal clear differences and similarities between the CEANITA and UKAB reports. Have the researchers hypothesised why these differences appear? Is It believed to be due to a difference in the reporting culture, or in the actual causal relations that triggered he reports? More importantly, could at any point anything be learned from the difference between the CEANITA and UKAB outputs to improve the NLP learning?*
**Author Response: We agree with the fact that both similarities and differences between CEANITA and UKAB reports are interesting to observe. Unfortunately, it is not easy to identify when the differences are due to the reporting culture (and to the different nature of the philosophy and processes of the two different investigation actors) or to the actual dynamic of the incident, nevertheless we agree it is appropriate to add some considerations about it. Action Undertaken: Some considerations about that have been added at the end of Section 6.1.**

*My choice to only assign a partial pass for the statistical analysis and its interpretation is only about the fact that more options could be described or the chosen options could be more clearly defended.*
**Author Response: We  recognise and acknowledge  the reviewer's point. Action Undertaken: Some considerations about the reasons behind the choices of the methodologies (and more some additional specifications about the choice of parameters) were added in Section 5.1, 5.2, 5.4.**

*It seems that the polygons for Figure 6 (CEANITA reports)  and Figure 7 (UKAB reports) are identical. Although the manuscript announced that the polygons are nearly identical for ground subjects, it also announces and essential difference (a peak in A1) for flight subjects. Could the authors verify if one figure has accidentally been used for both regions?*
**Author Response: We thank the reviewer for noticing it, there was indeed a problem (Figure 7 had been accidentally duplicated and used for both). Action Undertaken: Figure 6 has been substituted with the corrected figure.**

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 11 October 2021

https://doi.org/10.21956/openreseurope.15131.r27673

?

**Riccardo Patriarca**
Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Rome, Italy

Dear authors, Dear editorial team,

I thank you for giving me the opportunity to review this manuscript. I believe the presented topic is timely and the methodologies suggested by the authors are definitely at the pace with modern machine learning developments. Nonetheless, I have noticed several shortcomings from an epistemological point of view, besides observations and comments linked to the methodological development itself.

I am documenting here my main concerns to support a critical revision of the proposed manuscript.

**MAJOR CONCERN**
*"authors show how their proposal is able to automatically extract meaningful and actionable information from safety reports and to classify them according to the TOKAI taxonomy. The quality of the approach is also indirectly validated by checking the connection between the identified factors and*

*the main contributor of the incidents."*

About this statement, I am here reflecting on the actual possibility to reach this target. More specifically, the two mentioned database for reporting (and please note that this comment would apply to any set of organizations) might have different ATM expertise, or operational competences, different training, different backgrounds, different organizational culture even. While there is a common interest in standardising the investigation reports, it is largely acknowledged in the practitioners community that we are not there yet. From the report we see, the authors have suggested they want to follow Resilience Engineering principles (cf. the principle of “equivalence of successes and failures”), but is the data source they are using actually able to support that view?

How did the authors confirm the investigations they are using have been actually conducted in light of systemic (Resilience Engineering- driven) principles? Did they prove CEANTIA and UKABE (but again, I believe the same issue would apply to many other organizations) reports have been written following this logic? Are the reports actually capturing the complexity of reality? Since it does not seem the authors had any verification/validation in this sense, there is a risk that the analysis would just superimpose explanatory factors (coming from TOKAI) that actually are reliant on a different conceptual philosophy. It would have been different if the narrative comes from the same source where TOKAI factors come, at least to guarantee uniformity. Otherwise, this classification could even diverge into the side-effects typical of behaviourism.

With respect to aviation, one could reflect upon the inter-consensus issues documented even in the case the exact same taxonomy is used (in this case HFACS):
Nikki S. Olsen, Steven T. Shorrock, Evaluation of the HFACS-ADF safety classification system: Inter-coder consensus and intra-coder consistency, Accident Analysis & Prevention, Volume 42, Issue 2, 2010, Pages 437-444, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2009.09.005

These issues have been discussed in different domains, see for example:
Jonas Wrigstad, Johan Bergström, Pelle Gustafson, One event, three investigations: The reproduction of a safety norm, Safety Science, Volume 96, 2017, Pages 75-83, ISSN 0925-7535, https://doi.org/10.1016/j.ssci.2017.03.009.

I believe that’s a fundamental issue with the paper, which jeopardises the integrity of the results and conclusions. I believe the authors did not document properly these aspects and the related limitations.

*"Each CEANITA report concludes with a free-text description of the actions performed by air traffic controllers (ATCos) and pilots that contributed to the incident. Analogously, in a sample of UKAB reports, the final assessment of cause is listed in free text and includes the main contributory factors based on pilots’ and controllers’ actions. [...] the main causes: the most frequent ones in the corpus are wrong clearance (52%), deviation from procedures (22%), wrong or no resolution (17%–15%), [...] this section briefly summarises the main causes, contributory factors and the effectiveness of Safety Barriers as identified by the Board."*

Indeed, from these statements, it seems the reports document only the actions that “contributed” to the incident. Similarly, the identification of causes seems to be misaligned with Resilience Engineering principles. This is also proved by the main causes listed in page 4, which recalls a reductionist simplified understanding not aligned with Resilience Engineering. How much the authors believe the listed causes are socially constructed and how much are they really representative of the messy reality, trade-offs, goal conflicts operators face in everyday work? Again in TOKAI investigations narratives are paired with contributing/mitigating explanatory factors in a neutralised language. It seems risky to just look at one side of the story.

Similarly, the barrier-based approach documented in Figure 1 (ether effective or ineffective) recalls a traditional understanding and modelling of reality (bimodal, in contrast with the principles mentioned above), and it is unclear how the authors integrate it with modern safety thinking (and Resilience Engineering mentioned in the paper). The authors seem to make extensive usage of these elements (as documented in 5.2).
I feel Table 1 is superfluous and could have been presented jointly as an extended legend of Figure 1.

*"This section shows how the methods presented in Section 5 have been applied to achieve the scope of the work (see Section 3) demonstrating the effectiveness of the proposed approach on both the sets of data described in Section 4."*

Based on my previous concerns, I invite the authors to reflect whether they have been actually able to demonstrate their initial objective, or the scope/claims of the paper need to be restructured/resized to cope with what has been /can actually be achieved.

Why did the authors select only LoS? I agree they are particularly relevant events, but was it an opportunistic choice (i.e. numerosity)? Was there any additional justification?

**METHODOLOGY**
*"a simple model is developed".*
What do the authors mean here? It seems to be too generic.

I appreciate the fact that authors are using and citing mainly open-source resources for NLP analyses, however I do believe there should be a higher level of granularity for the respective detailed methodological aspects.

The technical details and values for the usage of LDA need to be justified. Did the authors perform any sensitivity analysis (number of iterations, burn in, alpha, etc.)? Even more, in 6.1.1., why only word and bigram? How did the authors systematically decide not to use any other n-gram?

Similarly, the way the authors describe the identification of the k topic is too generic. What kind of testing of different options and evaluation has been performed?

Again, how did the authors scale down the topics to 12 and 23 (from respectively 27 and 60)? Is this a signal of the inaccuracies of the underlying dataset (see previous concerns) which should have prevented from subsequent analyses? How did the authors validate the final number of clusters (experts involvement is mentioned in section 6.1.1, but how many experts, which backgrounds, how did the authors measure consensus)? Considering this manual refinement, to which extent, is this approach reproducible (resources to be involved, criteria for validation, consensus, etc.).

Similar comments apply for missing details on the Clustering description (distance criterion, quantitative criteria for clusters identification, etc.) and for the SVM paragraph (which seems to be too generic and not contextualized in the specific problem/result).

An overall revision is needed to add the required contribution, and remove duplicated statements (e.g. on the sample size, on the organizations involved etc.)

**IMPRECISE STATEMENTS**
*"TOKAI is a general taxonomy developed by EUROCONTROL."*
This statement is not correct. TOKAI (and its operating version e-TOKAI) includes a taxonomy, but it is a tool for investigation.

The authors discuss the increasing demand of air traffic as a critical aspect for complexity. I would argue that it is not only the growth in demand that increases complexity. It is also about the changing nature of human work, the dynamics of interactions between humans and technologies, the way those interactions propagate at micro-meso-macro level.

*"FARO is about "the impact that an increasingly complex environment has on the system safety"."*
However, the authors miss the opportunity to define what is complexity in the context of their research and how its definition shaped the project investigation methods and results. Are the methods actually able to capture such complexity?

*"Note that ATCo and pilots can be both directly responsible to the incident."*
Again, if the authors are grounding their work on Resilience Engineering principles, I wonder if this is the correct wording or might be misinterpreted as oriented towards blame culture.

**FINAL OBSERVATIONS**
I do firmly believe that we would benefit from machine learning analyses in the safety management domain. I welcome the idea proposed by the authors to use modern techniques within safety investigations, but at the same time I recommend them to be cautious in language used, and the way they document what can be actually done or how these results configure within modern and well-established safety paradigms.

**References**

1. Olsen NS, Shorrock ST: Evaluation of the HFACS-ADF safety classification system: inter-coder consensus and intra-coder consistency.*Accid Anal Prev*. 2010; **42** (2): 437-44 PubMed Abstract | Publisher Full Text
2. Wrigstad J, Bergström J, Gustafson P: One event, three investigations: The reproduction of a safety norm. *Safety Science*. 2017; **96**: 75-83 Publisher Full Text

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and does the work have academic merit?**
Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
No

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Aeronautical engineering, air traffic management, system safety, safety risk management, resilience engineering, socio-technical systems, accident analysis, aviation safety

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 24 Jan 2022

**Luca Oneto**, ZenaByte, Genova, Italy

**Dear Riccardo Patriarca, we thank you for your valuable comments that allowed us to improve the manuscript. What follows is the pointwise response to your comments with the corresponding changes in the new version of the manuscript. (Reviewer comments in italics)**

*Dear authors, Dear editorial team, I thank you for giving me the opportunity to review this manuscript. I believe the presented topic is timely and the methodologies suggested by the authors are definitely at the pace with modern machine learning developments.*
**We thank you for your positive comments.**

*Nonetheless, I have noticed several shortcomings from an epistemological point of view, besides observations and comments linked to the methodological development itself. I am documenting here my main concerns to support a critical revision of the proposed manuscript.*
*"authors show how their proposal is able to automatically extract meaningful and actionable information from safety reports and to classify them according to the TOKAI taxonomy. The quality of the approach is also indirectly validated by checking the connection between the identified factors and the main contributor of the incidents." About this statement, I am here reflecting on the actual possibility to reach this target. More specifically, the two mentioned databases for reporting (and please note that this comment would apply to any set of organizations) might have different ATM expertise, or operational competences, different training, different backgrounds, different organizational culture even. While there is a common interest in standardising the investigation reports, it is largely acknowledged in the practitioners community that we are not there yet. From the report we see, the authors have suggested they*

*want to follow Resilience Engineering principles (cf. the principle of "equivalence of successes and failures"), but is the data source they are using actually able to support that view? How did the authors confirm the investigations they are using have been actually conducted in light of systemic (Resilience Engineering- driven) principles? Did they prove CEANTIA and UKAB (but again, I believe the same issue would apply to many other organizations) reports have been written following this logic? Are the reports actually capturing the complexity of reality? Since it does not seem the authors had any verification/validation in this sense, there is a risk that the analysis would just superimpose explanatory factors (coming from TOKAI) that actually are reliant on a different conceptual philosophy. It would have been different if the narrative comes from the same source where TOKAI factors come, at least to guarantee uniformity. Otherwise, this classification could even diverge into the side-effects typical of behaviourism.*

**Author Response: A first, probably misleading issue, in the quoted statement is that we did indeed not actually classify THEM (i.e., the reports) but we did classify IT (i.e., the information in the reports). Even taking into account this correction, we understand and share the reviewer's concern: since our algorithm classifies the information included in the reports, and the reports are not written in a standardised way (and especially not accordingly to the TOKAI logic), we have to bear in mind that our algorithm is in turn not able to standardise them. We can only extract and reshape the information that is already in the reports, inheriting all the reports' limitations. Our intention is to show how it is possible to develop a tool to extract taxonomy factors from the text. Then, we should make it more clear in the paper that the extracted factors will just describe what is written in the report, nothing more than that: we don't claim to have a magical tool able to make reports written with completely different logics immediately comparable and standardised. We believe that this tool remains potentially useful despite these limitations. Indeed, it may even become a diagnostic tool to identify reports' narrative issues (e.g., we may discover that the fact that some factors are not mentioned in the CEANITA or UKAB reports is actually a problem and investigation should focus more on those aspect) or a tool to understand how reporting philosophy evolves in a period of time. We also agree that the data sources may not be suitable to fully adopt a Resilience Engineering view: neither of the two sources can be indeed proven to systematically capture complexity, even if they are quite progressive and enlightened (UKAB in particular), but they cannot be labelled as consistent with a RE perspective. This point is acknowledged and we understand the corresponding claims made in the paper may actually be misleading. What we intended is just that, given the reports as they are, the extraction, processing and visualisation of the results is intended to take into account a RE perspective (e.g., when choosing the taxonomy, the choice of the TOKAI one was made also in that light) as much as possible. Actions Undertaken: The quoted statement has been slightly modified (see the abstract). Statements about RE have been mitigated (see Introduction and Conclusions). A paragraph was added (see Section 3.2) to make both our intentions and the approach's limitations explicit. A paragraph was added in the Conclusions to discuss these limitations and possible ways to manage them in the future.**

With respect to aviation, one could reflect upon the inter-consensus issues documented even in the case the exact same taxonomy is used (in this case HFACS): Nikki S. Olsen, Steven T. Shorrock, Evaluation of the HFACS-ADF safety classification system: Inter-coder

consensus and intra-coder consistency, Accident Analysis & Prevention, Volume 42, Issue 2, 2010, Pages 437-444, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2009.09.005 These issues have been discussed in different domains, see for example: Jonas Wrigstad, Johan Bergström, Pelle Gustafson, One event, three investigations: The reproduction of a safety norm, Safety Science, Volume 96, 2017, Pages 75-83, ISSN 0925-7535, https://doi.org/10.1016/j.ssci.2017.03.009. I believe that’s a fundamental issue with the paper, which jeopardises the integrity of the results and conclusions. I believe the authors did not document properly these aspects and the related limitations.
**Author Response:** **We concur with this comment, and of its significance. Thank you for introducing the references, that at least one we are familiar with.** **Actions Undertaken:** **Some considerations about inter-rater reliability have been added in the Conclusions when discussing limitations.**

"Each CEANITA report concludes with a free-text description of the actions performed by air traffic controllers (ATCos) and pilots that contributed to the incident. Analogously, in a sample of UKAB reports, the final assessment of cause is listed in free text and includes the main contributory factors based on pilots’ and controllers’ actions. [...] the main causes: the most frequent ones in the corpus are wrong clearance (52%), deviation from procedures (22%), wrong or no resolution (17%–15%), [...] this section briefly summarises the main causes, contributory factors and the effectiveness of Safety Barriers as identified by the Board." Indeed, from these statements, it seems the reports document only the actions that “contributed” to the incident. Similarly, the identification of causes seems to be misaligned with Resilience Engineering principles. This is also proved by the main causes listed in page 4, which recalls a reductionist simplified understanding not aligned with Resilience Engineering. How much the authors believe the listed causes are socially constructed and how much are they really representative of the messy reality, trade-offs, goal conflicts operators face in everyday work? Again in TOKAI investigations narratives are paired with contributing/mitigating explanatory factors in a neutralised language. It seems risky to just look at one side of the story. Similarly, the barrier-based approach documented in Figure 1 (ether effective or ineffective) recalls a traditional understanding and modelling of reality (bimodal, in contrast with the principles mentioned above), and it is unclear how the authors integrate it with modern safety thinking (and Resilience Engineering mentioned in the paper). The authors seem to make extensive usage of these elements (as documented in 5.2).
**Author Response:** **As previously mentioned, we agree that the reports analysed are often not aligned with RE principles, implying a series of limitations to the work. However, here a clarification is necessary: both UKAB and CEANITA reports are composed of different sections (see Section 4) and in particular: 1) The assessment of Main causes (for CEANITA) and Barriers (for UKAB) represents a particularly synthetic and simplified piece of information, which undoubtedly implies a reductionist view. This information is mainly considered in the EDA, before the application of more insightful analyses, and one of the main objectives of our work is indeed to reach a higher-level insight about the LoSs in order to better capture the complexity of reality and not just look at one side of the story. 2) The different free-text sections, in particular the conclusions (CEANITA) and final assessment (UKAB), are instead of a slightly different nature: especially in the CEANITA case, they summarise the dynamics of the incident in a more extensive and richer way, also outlining actions**

**not directly contributing to the incident (which was probably written in an unclear way in the paragraph the reviewer quotes). The TOKAI algorithm and all the other NLP models were applied to these free-text sections, and indeed the RE perspective is mainly referred to this part of the work (with all the already mentioned limitations, already mitigated in the new version of the paper). Actions Undertaken: Section 3.2 has been edited to better explain the content of the free-text parts.**

*I feel Table 1 is superfluous and could have been presented jointly as an extended legend of Figure 1.*
**Author Response: We agree. Actions Undertaken: Corrected (see Figure 1).**

*"This section shows how the methods presented in Section 5 have been applied to achieve the scope of the work (see Section 3) demonstrating the effectiveness of the proposed approach on both the sets of data described in Section 4." Based on my previous concerns, I invite the authors to reflect whether they have been actually able to demonstrate their initial objective, or the scope/claims of the paper need to be restructured/resized to cope with what has been /can actually be achieved.* **Author Response: We agree. Actions Undertaken: An extensive discussion of the work's limitations has been added to the Conclusions. Some statements about the objectives have been slightly mitigated in the Introduction.**

*Why did the authors select only LoS? I agree they are particularly relevant events, but was it an opportunistic choice (i.e. numerosity)? Was there any additional justification?*
**Author Response: The FARO project selected SMIs as a large part of the consortium have specific experience on this: on one side, ENAIRE and EUROCONTROL advised on operational matters, on the other side, two of the universities involved have conducted extensive research into the dynamics and mathematical formulations of trajectory prediction and probabilistic conflict modelling. It was a natural choice given this experience and the composition of the consortium.**

**"***a simple model is developed". What do the authors mean here? It seems to be too generic.*
**Author Response: While we agree that that sentence was a bit too generic (now it has been corrected to "a simple Machine Learning model"), further details about that model are already extensively included in the following sections ("In particular, a SVM with Gaussian kernel trained with the 89 CEANITA reports was used, performing accurate model selection - the kernel and the complexity hyperparameters were searched in {10-4, 10-3.5 ,…, 103}…") and the methodological foundations of SVM are described in Section 5. Actions Undertaken: "a simple model" has been changed to "a simple Machine Learning model".**

*I appreciate the fact that authors are using and citing mainly open-source resources for NLP analyses, however I do believe there should be a higher level of granularity for the respective detailed methodological aspects. The technical details and values for the usage of LDA need to be justified. Did the authors perform any sensitivity analysis (number of iterations, burn in, alpha, etc.)? Even more, in 6.1.1., why only word and bigram? How did the authors systematically decide not to use any other n-gram?*
**Author Response: The Number of iterations and burn in were set by assessing convergence through the likelihood graphs produced by the models (i.e., numbers**

**large enough to not see any more changes in the graph). Initial alpha and beta have been tuned together with k (even if default values turned out to be appropriate for both initial alpha and beta). A number of preliminary analyses were performed to understand the relevance and usefulness of different n-grams to describe the reports' content  (both per se and in the topic-modelling framework), and in both CEANITA and UKAB use cases the role of n-grams with n>2 resulted substantially negligible per se and source of additional noise. <u>Actions Undertaken:</u> All these details have now been included in Section 5.1.**

*Similarly, the way the authors describe the identification of the k topic is too generic. What kind of testing of different options and evaluation has been performed? Again, how did the authors scale down the topics to 12 and 23 (from respectively 27 and 60)? Is this a signal of the inaccuracies of the underlying dataset (see previous concerns) which should have prevented from subsequent analyses? How did the authors validate the final number of clusters (experts involvement is mentioned in section 6.1.1, but how many experts, which backgrounds, how did the authors measure consensus)? Considering this manual refinement, to which extent, is this approach reproducible (resources to be involved, criteria for validation, consensus, etc.).*
**<u>Author Response:</u> The number k of topics was tuned accordingly to coherence and R2 metrics, searching the parameter in {10,11,...,99,100}. The final number of topics was manually chosen according to experts' opinion (mainly from CRIDA, Lund University, and ENAIRE): some topics were discarded as too similar between each other, some topics because they made sense on a lexical point of view (e.g., commonly used idioms or standard sentence formulations) but did not convey any meaningful information, others because they did not make much sense or were not very coherent. To our knowledge, especially w.r.t. the R implementation we used, it is quite common to reduce the number of topics to identify the most interesting ones, and we do not consider it a signal of datasets' inaccuracy. Since the reviewer mentions section 6.1.1 we assume he is referring to topics instead of clusters. When mentioning experts' opinions, we refer mainly to those from CRIDA, Lund University, and ENAIRE, in particular involving researchers in the aviation safety and resilience fields, navigation managers and also a former ATCo. Each expert made a separate personal selection of the topics before discussing together the final one, but interestingly the single selections were not particularly different between each other, and there was a general agreement on the final selection. We can assume that the approach, even if not exactly reproducible due to the variability of human evaluation, could be replicated without massive differences. <u>Actions Undertaken:</u> All these details have now been included in Section 5.1.**

*Similar comments apply for missing details on the Clustering description (distance criterion, quantitative criteria for clusters identification, etc.) and for the SVM paragraph (which seems to be too generic and not contextualized in the specific problem/result).*
**<u>Author Response:</u> As reported in Section 5.2, we applied Ward's minimum variance criterion and the criteria for cluster selection were the dendrogram, the scree plot (with distance between clusters as metrics). These two methods reduced the number of clusters to a couple of options, which were manually evaluated by the aforementioned domain experts. The SVM paragraph was more generic as we decided to better specify the details of the models in the following sections when describing**

**experimental results. This was due to the fact that we found it necessary to provide some context and explain which model was applied in each different case. Details about that model were extensively included in Section 6.2.1 ("In particular, a SVM with Gaussian kernel trained with the 89 CEANITA reports was used, performing accurate model selection - the kernel and the complexity hyperparameters were searched in {10-4, 10-3.5 ,…, 103}…"). However, thanks to the reviewer's comment, we reflected on the opportunity to move these details in Section 5 to improve the quality and the clarity of the paper. Actions Undertaken: Details about SVM have been included in Section 5.**

*An overall revision is needed to add the required contribution, and remove duplicated statements (e.g. on the sample size, on the organizations involved etc.)*
**Author Response: We agree (only note that the repetition of sample size was sometimes due to avoid confusion since for UKAB reports the sample sizes used for the EDA and the TOKAI algorithm is not exactly the same). Actions Undertaken: An overall revision has been performed.**

IMPRECISE STATEMENTS "TOKAI is a general taxonomy developed by EUROCONTROL." This statement is not correct. TOKAI (and its operating version e-TOKAI) includes a taxonomy, but it is a tool for investigation.
**Author Response: We thank the reviewer for noticing this with this comment. Actions Undertaken: The statement has been corrected.**

*The authors discuss the increasing demand of air traffic as a critical aspect for complexity. I would argue that it is not only the growth in demand that increases complexity. It is also about the changing nature of human work, the dynamics of interactions between humans and technologies, the way those interactions propagate at micro-meso-macro level.*
**Author Response: Yes, we concur. This position is also central in other FARO work packages. Actions Undertaken: Introduction has been modified to include these considerations.**

*FARO is about "the impact that an increasingly complex environment has on the system safety"." However, the authors miss the opportunity to define what is complexity in the context of their research and how its definition shaped the project investigation methods and results. Are the methods actually able to capture such complexity?*
**Author Response: In principle we agree. A clarification about FARO is however necessary: FARO is an exploratory research project adopting a methodology that embraces orthodox safety in one work package (WP4) and a thorough study of resilient performance in a second work package (WP5). These two different perspectives of a social-technical system are contributor a third work package (WP6) which takes these two different perspectives and synthesises these to explore a view of ATM system performance. This paper reports only another – preliminary – work package (WP3) that was principally the one exploring the use of NLP on reports with the objective of capitalising extant safety knowledge. In essence whilst FARO does adopt resilience engineering precepts and principles and does focus on the impact of complexity, it is not in the specific area of interest of the task that is reported in the paper. It is fair to say that the NLP work package tried not to disregard the RE**

**philosophy and to deal with complexity as much as possible, but found it often difficult to conjugate with the reporting philosophy of the data sources. To sum up: FARO indeed focuses on complexity but mainly in other parts of the project, for which this piece of research was just a starting point. Actions Undertaken: A couple of statements summarising the preliminary nature of this research have been added in the Introduction.**

**“***Note that ATCo and pilots can be both directly responsible to the incident." Again, if the authors are grounding their work on Resilience Engineering principles, I wonder if this is the correct wording or might be misinterpreted as oriented towards blame culture*.
**Author Response: We agree that the statement may be misleading. This was purely a description of what is written in the reports (i.e., there is a section about pilots/ATCos responsibilities where someone is indicated as directly responsible, possibly both ATCos and pilots) which, as already discussed, contain different parts misaligned with RE principles. Actions Undertaken: “ATCo and pilots can be both directly responsible” has been changed to “ATCo and pilots can be both indicated in the reports as directly responsible”.**

***I*** *do firmly believe that we would benefit from machine learning analyses in the safety management domain. I welcome the idea proposed by the authors to use modern techniques within safety investigations, but at the same time I recommend them to be cautious in language used, and the way they document what can be actually done or how these results configure within modern and well-established safety paradigms.*
**We thank the reviewer for his comments which helped us enhance the quality of the paper by noticing limitations and inaccuracies.**

***Competing Interests:*** No competing interests were disclosed.