APPLICATION OF NATURAL LANGUAGE PROCESSING FOR AIRCRAFT DEFECT TRACKING IN MAINTENANCE OPERATIONS

Michael J. Scott[1], Oliver Kirkpatrick[1], Wim Verhagen[1], Vlado Kekoc[2], Bob Teunisse[3], Jenny Zhang[1], Haytham Fayek[1] & Pier Marzocca[1]

[1]RMIT University, Melbourne, Victoria, Australia
[2]Defence Science and Technology Group, Fishermans Bend, Victoria, Australia
[3]Capability Acquisition and Sustainment Group, Melbourne, Victoria, Australia

Abstract

Accurate defect identification and tracking of aircraft components and systems is key to maintaining fleet operational readiness and minimising sustainment costs. Maintainer reports offer a rich data source, with lead indicators of defects able to be found within unstructured natural language text fields. However, maintenance textual reports in present form require manual and often difficult analysis, which limits their use for proactive maintenance and reliability checks. The application of Natural Language Processing (NLP) and Machine Learning (ML) approaches offers a solution to exploit previously under-utilised maintenance reports and complementary sources such as pilot reports. NLP coupled with ML techniques can be used to detect the emergence of recurrent defects and could supplement maintainers in making data-driven decisions. This can enable advanced part ordering and resource planning, helping to reduce aircraft downtime by mitigating effects of supplier lead times. This paper introduces a framework for application of NLP towards technical reports and demonstrates a proof-of-concept version of this NLP framework using a set of maintenance and pilot reports applicable to a mixed fleet of aircraft types operated over a two year period. Analysis of the processed outputs enables identification of early lead indicators that have potential to prevent failures, if actioned early by maintainers. End-user feedback highlights the potential value of the proposed approach and implementation, while several avenues for subsequent research are identified.

Keywords: Aircraft, Fleet Management, Machine Learning, NLP, Predictive Maintenance.

1. Introduction

AIRCRAFT fleet management is continually competing against time and cost, as task compliance is a non-negotiable safety and regulatory constant, which can inherently render maintenance efforts reactive rather than proactive in practice [1]. Consider a maintainer encountering a defect on an aircraft, such as an actuator or corrosion problem. In such instances the maintainer is limited in capacity to optimise their actions, as there is typically no way to readily:

- compare how severe and wide-spread the problem is across the fleet;
- understand recurrent history over a longer timeline for individual aircraft;
- track this defect over subsequent bases or shifts; and
- support human intuition and connections between pilot and maintainer reports.

Various types of technical reports, in particular maintenance and pilot reports, contain observations which may aid to achieve the functions mentioned above. In addition, a maintainer may want to correlate maintenance reports with geographic information, signal response data from flight data recorders, sensor data (*e.g.*, from aircraft monitoring systems) or another unscheduled event [2].

However, at present inspecting technical reports, let alone combining findings with other sources of data, remain time-consuming tasks, requiring data science expertise and gathering disparate information from varying sources. Consequently, frameworks capable of merging and analysing various data types will be essential in identifying connections that isolated investigations typically overlook [3].

Developing big data analytics approaches for fleet management is possible with the large amount of reporting required in operations and mandated by regulators in maintenance and reliability programs. For the purposes of improving operational readiness, the application of Natural Language Processing (NLP) presents an opportunity to automatically extract insights from maintenance and pilot reports. NLP has the capability to extract meaningful information from textual reports, enriching engineering and maintenance analysis towards achieving improved operational efficiency and platform readiness, as demonstrated by MaintNet, a collaborative open-source library for predictive maintenance language resources [4]. While Commercial-off-the-shelf options are available [5], providing operators NLP capability for recurrent defect management, these require frequent data exchange to their propriety backend with limited explainability of underlying processes and require large amounts of structured data from operators, including a high level of attention in clearing spurious results.

NLP has been addressed in prior research as a means to enhance aircraft maintenance and safety reporting. Research in this area has focused on using NLP to identify operational inefficiencies in aviation safety reports, providing crucial insights for improving safety protocols [6]. A systematic review of the field indicates a broad exploration of methodologies and applications, suggesting promising future directions for NLP in aviation safety [7]. Additionally, NLP application in analysing occurrence reports within safety critical industries is pivotal for informing research and identifying common challenges [8]. This technology supports safety data analysis by aiding in the categorisation of reports, although it faces challenges like the need for manual annotation and variability in labelling [9, 10]. Overall, NLP integration into aircraft maintenance and safety reporting is instrumental in advancing safety management systems, despite some existing limitations.

Recent literature [11, 12, 13, 14, 15, 16, 17] highlights the application of NLP approaches in maintenance contexts, including aviation. In a comprehensive review, [12] highlight challenges for NLP in industrial maintenance, including development of interpretable models and evaluation in real industrial settings. Prior to this, the authors discuss classifications and applications of NLP approaches in industrial maintenance. In terms of classifications, word-embedding approaches are contrasted to other categories such as transformer-based methods. The latter are increasingly popular in NLP applications, including in aerospace as for instance shown in the work by [15], but have drawbacks in terms of explainability as well as verification and validation, as transformer-based methods are black-box in nature, are computationally expensive to fine-tune and the outputs can be opaque, posing challenges in interpretability and transparency. In the context of aircraft maintenance analysis and decision-making it is critical that end-users can trace the workflow steps and clearly understand outputs.

In contrast, other authors in the aerospace maintenance domain [11, 14] have applied word-embedding methods such as Term Frequency - Inverse Document Frequency (TF-IDF), Word2Vec and others to analyse aircraft maintenance data. The appeal of these methods are that they are suitable (to a degree) to automatically evaluate large amounts of maintenance textual data which would otherwise be prohibitive to manually analyse. Word-embedding methods are sufficiently flexible to adapt to domain-specific jargon, which poses more of a barrier for the adoption of other competing, off-the-shelf NLP methods. While the aforementioned initiatives have shown great promise, a notable gap in the current state of the art is the lack of subsequent models and tools to facilitate decision making. While a few papers venture into classification or prediction, the next step - making use of these outputs - is typically not addressed. Coupled with the aforementioned issue of model interpretability as well as a relative lack of industrial applications, several major gaps are currently present in the state of the art, preventing the step towards tools of impact in industrial practice.

This paper aims to move beyond the current state of the art by addressing the aforementioned gaps in the literature, focusing primarily on the extension of NLP techniques towards explainable and actionable data for decision support. To achieve this, an NLP framework is introduced which combines the use of well-established, explainable NLP techniques with the additional post-processing step of risk rating displayed in a dashboard. The dashboard leverages a risk rating approach and allows for data-driven decision support, as further set out in Section 2. Section 3 covers the application of the NLP framework to a representative dataset, which is first characterised in Section 3.1 before results and expert validation are discussed in Section 3.2. Finally, concluding the paper with a discussion on end-user feedback in Section 4 and remarks on future work required to develop a minimum viable product (MVP) for operational use.

2. Methodology: NLP and Machine Learning

This section details the developed NLP framework. It comprises a linear workflow organised into three distinct phases, being i) data pre-processing; ii) vectorisation and machine learning; and iii) risk rating with maintainer analysis. Figure 1 illustrates the sequential processes of the NLP framework, highlighting the main methods utilised to process raw unstructured text from pilot and maintenance unserviceable (U/S) reports. It should be noted that several maintenance reports are generated for compliance and registering work order completions for nominal checks as part of maintenance programs, however U/S reports are the focus for this approach as they contain details of anomalies.

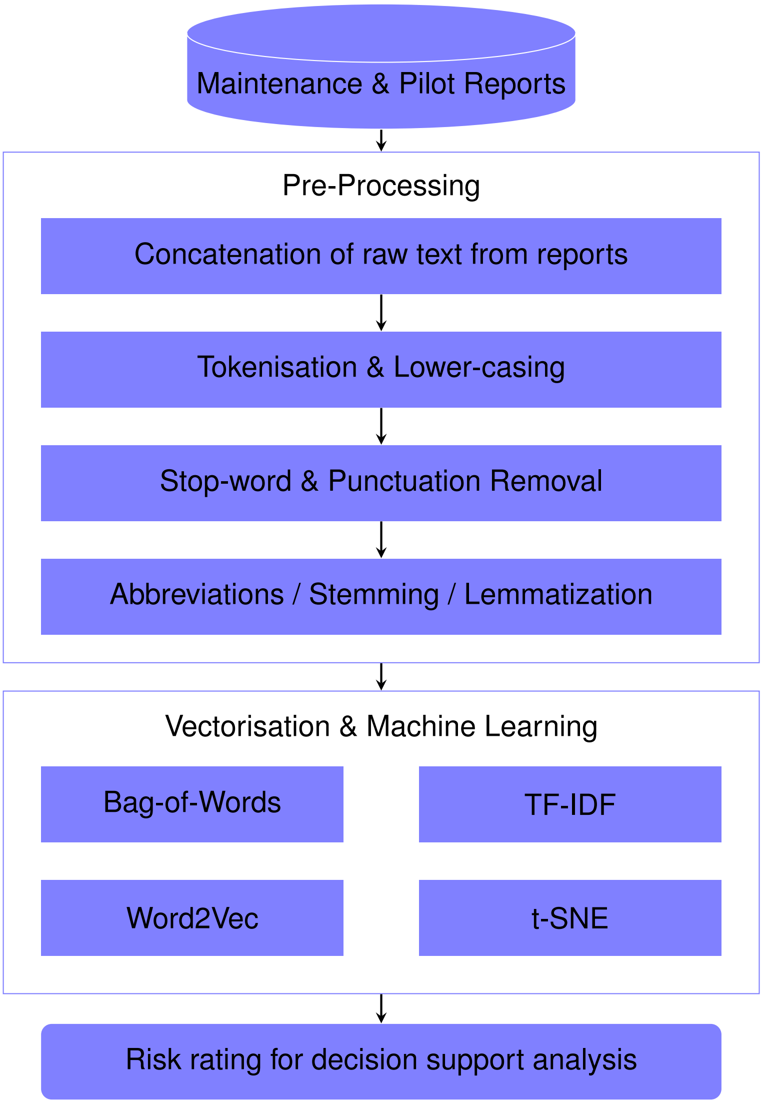

Figure 1 – Example flowchart of the process for NLP and analysis of maintenance & pilot reports

Initially, data acquisition is performed, followed by a pre-processing step where text data undergoes normalisation procedures (*e.g.*, lower casing), removing noise (*i.e.*, terms such as "the", punctuation, etc.) and standardising formats. This is followed by stages that include tokenization, where text is split into tokens, and parsing, which arranges tokens into a structured format suitable for computational analysis. The framework incorporates machine learning algorithms, detailed as a separate module, to perform semantic analysis, extracting meaning from text based on contextual cues. The main methods used on the processed text are Term Frequency - Inverse Document Frequency (TF-IDF) and Word2Vec vectorization, converting the qualitative data into an embedding.

Subsequent to vectorization, the second layer of the framework commences by filtering key terms, for example, "Actuator" and "Hydraulic". The framework data processing component analyses the frequency of referenced maintenance actions to establish a risk rating for each report. This risk rating is determined through a combination of algorithmic assessment and a one-off manual verification against a predefined key term taxonomy, assigning a risk rating for various maintenance action keywords. The keywords are categorised into three risk levels: low (1), medium (2), and high (3), corresponding to the classification for prioritising maintenance actions and managing the workflow effectively by triaging potential risks. For instance, actions like 'check' and 'examine' are considered low risk, whereas 'install' and 'replace' are high risk and may require more immediate attention or additional resourcing and parts, which is based on the principles of MSG-3 decision logic [18].

Lastly, the final output is an interactive dashboard that provides a synthesised dataset for in-depth analysis, thereby aiding fleet maintainers in making informed data-driven decisions. End-users such as maintainers or engineers can then make fleet management decisions based on these new insights. This may include proactive parts ordering and optimisation of resourcing.

The following subsections details the two foundational methods integral to the NLP framework, specifically TF-IDF and Word2Vec. These methods are essential in the text analysis by transforming textual data into numerical representations that capture semantic meaning. By understanding the principles and applications of TF-IDF and Word2Vec, it is possible to gain deeper insights into the nuances of what the NLP framework is performing in the background and derivation and implications for supplementing maintainer decision-making.

2.1 Term Frequency-Inverse Document Frequency (TF-IDF) for Term Extraction

In NLP, extracting relevant and unique terms from a corpus of text documents is crucial for various applications such as text summarisation and topic modelling, in this case identifying trends through quantification of terms. One widely-used technique for this purpose is TF-IDF. TF-IDF is a statistical measure used to evaluate the importance of a word in a report (document) relative to a collection of reports (corpus).

The TF-IDF value of a term t in a document d within a corpus D is calculated as follows:

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D) \tag{1}$$

where $\text{TF}(t,d)$ represents the term frequency of t in d, and $\text{IDF}(t,D)$ represents the inverse document frequency of t across the corpus D.

2.1.1 Term Frequency (TF)

The term frequency $\text{TF}(t,d)$ is defined as the number of times term t appears in document d. It can be normalised by the total number of terms in the document to avoid bias towards longer documents. The normalised term frequency is given by:

$$\text{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2}$$

where $f_{t,d}$ is the raw count of term t in document d.

2.1.2 Inverse Document Frequency (IDF)

The inverse document frequency IDF(t,D) measures the rarity of a term across the corpus. It is defined as:

$$\text{IDF}(t,D) = \log\left(\frac{N}{|d \in D : t \in d|}\right) \tag{3}$$

where N is the total number of documents in the corpus, and $|d \in D : t \in d|$ is the number of documents containing term t. The logarithm is used to dampen the effect of terms that appear in many documents.

2.1.3 Application of TF-IDF in Report Analysis

To extract terms that are both relevant to individual reports and unique across the collection of reports, TF-IDF measure is applied. By calculating the TF-IDF score for each term in each report, it enables identification of terms that are significant within a specific report, while being distinctive when compared to the entire collection. This process helps in highlighting keywords that encapsulate the unique content of each report, but importantly extracting out significant problematic components and sourcing the key corrective maintenance action terms.
The implementation of TF-IDF involves the following steps:

1. Preprocessing the text data by tokenizing, removing stop words, and stemming or lemmatising.
2. Calculating the term frequency for each term in each document.
3. Computing the inverse document frequency for each term across the corpus.
4. Multiplying the term frequency by the inverse document frequency to obtain the TF-IDF score.
5. Selecting the top n terms with the highest TF-IDF scores for each document to represent the most relevant and unique terms.

This methodology effectively captures the salient features of each maintenance report, providing a robust basis for further NLP tasks such as clustering, classification and semantic analysis, which leads into the following section.

2.2 Word2Vec and Cosine Similarity

A well-established technique in NLP is Word2Vec [19] that maps words into continuous vector spaces. Word2Vec models use neural networks to capture semantic relationships between words by training on large corpora of text, which can be on specific datasets ($e.g.$, pilot and maintenance reports).

Word2Vec creates dense vector representations (embeddings) of words, so that words with similar meanings are positioned close to each other in the vector space. This is achieved through two primary model architectures: Continuous Bag of Words (CBOW) and Skip-gram. The CBOW model predicts the target word from the surrounding context words, while the Skip-gram model predicts the context words given a target word.

An important metric in the underlying NLP Word2Vec method is cosine similarity, which effectively quantifies the similarity, or cosine of the angle between word vectors, providing a metric for semantic relationships between words. The cosine similarity equation, as shown in Equation 4, is defined as a measure of the cosine of the angle between two vectors projected in a multi-dimensional space.

$$\text{Cosine Similarity}(A,B) = cos(\theta) = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

where A and B are vectors of the terms in a corpus, while θ represents the angle between vectors A and B. The absolute of the vectors are taken to calculate the respective Euclidean norms.

Terms are made relative with each other, syntactic and semantic word similarities, based on proximity in the corpus forming an pseudo ontology. In this Equation 4, cosine similarity is calculated as the cosine of the angle θ between vectors A and B. The result will still be a value between -1 and 1, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates opposite. The inclusion of θ enables explicit consideration of the angle between the vectors when measuring their similarity. This is illustrated in the following Figure 2 showing terms "rudder", "lamp" and "seat" relative to term "actuator".

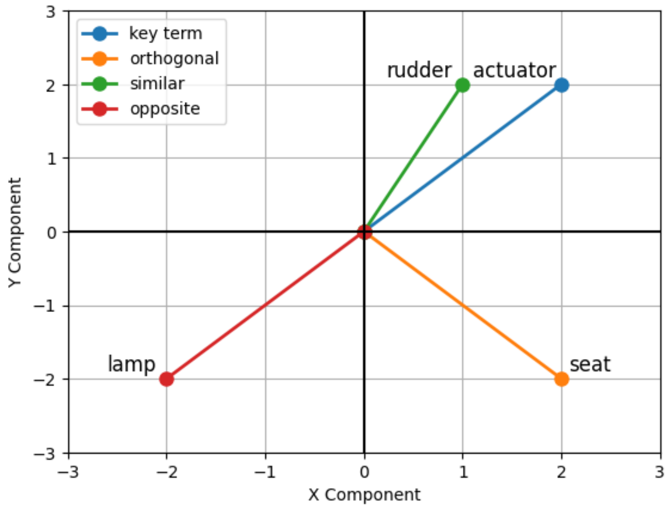

Figure 2 – Illustrative example of Cosine Similarity between terms - showing "rudder" and "actuator" are most similar, with a lower cosine angle than "seat" and "lamp".

A trained Word2Vec model combined with the cosine similarity metric helps to rapidly connect pilot and maintenance reports and understand the connections between reports that would be manually difficult to identify with a large dataset for a fleet of aircraft.

3. Results

3.1 Dataset characteristics

The dataset being used in the study is a surrogate dataset representative of typical aircraft maintenance reports and based upon real-life industrial data. This surrogate dataset comprises approximately 120,000 pilot and maintenance reports to evaluate trends and identify lead indicators within the unstructured text field. It should be noted that structured data such as time and aircraft serial number are used in the analysis for tracking purposes.

Comparisons can be made between other studies and open-source datasets available, for example MaintNet [4], NASA Aviation Safety Reporting System (ASRS) [1] and Federal Aviation Administration (FAA) Safety reports from the Service Difficulty Reporting System (SDRS) [2], the surrogate dataset is aligned well to our purposes for a varied yet challenging dataset.

Comparing the four datasets, from a random sample of 1,000 reports, key text metrics indicate the variability and complexity of the language within aeronautical reports and the suitability for NLP model training. The following metrics are used in the comparison, as shown in Table 1: lexical diversity, sentiment subjectivity, syntactic complexity, difficult words count, lexicon count, and sentence count. Shown in Figure 3 are the normalised text metrics, which clearly indicate that the ASRS dataset is better suited for NLP model training due to its high mean values. For simpler, more straightforward

[1] https://asrs.arc.nasa.gov
[2] https://sdrs.faa.gov

Table 1 – Comparison of mean text metrics across different aircraft textual datasets.

Label	Lexical Diversity	Sentiment Subjectivity	Syntactic Complexity	Difficult Words	Lexicon Count	Sentence Count
ASRS	0.55	9.31	22.31	36.90	254.32	15.72
MaintNet	0.84	0.37	1.03	4.14	10.57	1.93
SDRS	0.83	4.26	1.37	7.09	33.25	2.96
Surrogate	0.93	0.88	0.25	4.20	15.24	1.60

text, the MaintNet dataset is more suitable, while SDRS provides a balance of complexity and diversity. Lastly, the Surrogate dataset is ideal for varied vocabulary with simpler sentence structures, which is more representative of maintenance reports, rather than safety reports. However, for model training the ASRS stands out given the more comprehensive language and lexicon count.

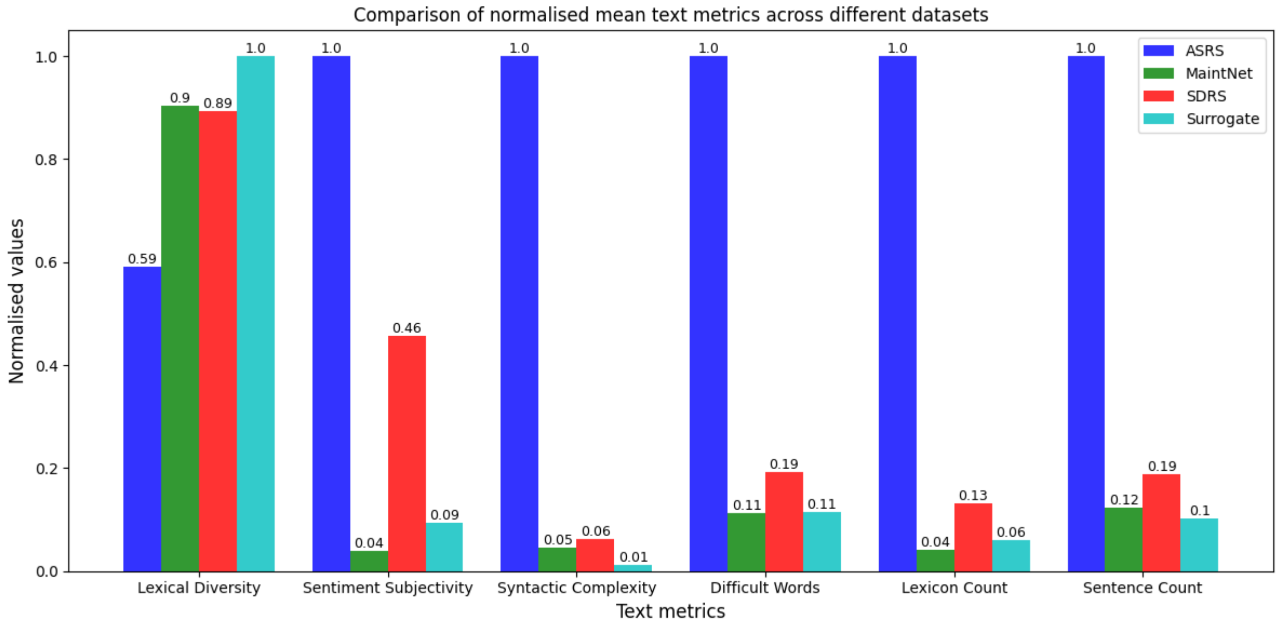

Figure 3 – Comparison of the normalised mean of text metrics across the sample datasets and the surrogate dataset.

To illustrate the output of the Word2Vec model, which is trained on the surrogate dataset, as described in Section 2.2, Figure 4 visualises the cosine similarity of various terms using t-Distributed Stochastic Neighbor Embedding (t-SNE) [20], which is a machine learning algorithm for dimensionality reduction. Here, *relative* positions of terms are of greater relevance than their *absolute* positions, hence the axis values are neglected. Terms are grouped by colour-coded clusters indicating related terms. Each point represents a term, such as "lamp," "bulb," and "light" in one cluster, indicating these terms are closely related in the context. While clusters including terms "rudder," "aileron," and "elevator", which relate to aircraft control surfaces and the term "actuator". There are more structural terms like "corrosion", which relate to "rivet" and "frame".

3.2 Results and validation

In this section, the introduced NLP framework is applied to the surrogate dataset. The first two phases of the NLP framework workflow, take less than 30 minutes for processing the reports contained in the surrogate dataset. It should be noted this was computed using a laptop with 8 cores, capable of handling 16 total processor threads (3.30 GHz base clock speed) and 32 GB of RAM.

Moving to the third phase of the NLP framework, Table 2 is an representative case where the NLP Framework identified a trend of recurrent defects relating to actuators and hydraulics for a specific

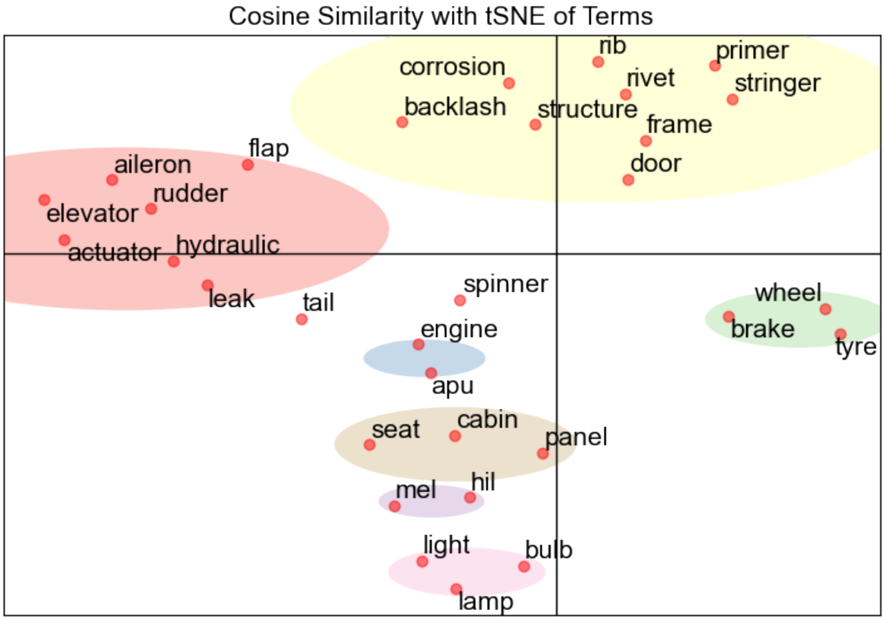

Figure 4 – Illustrative example of Cosine Similarity between terms - showing the clustering of terms.

aircraft. The sequence of reports is characteristic of an initial report where no fault is found and acquitted, but then some days later the defect returns. Such reports are collated and flagged in a dashboard view for maintainers to view and supplement tactical decision-making, Table 2 is an extract from this dashboard.

Important aspects to note in this clustering of six reports in Table 2 is the mix of sources, maintenance and pilot reports, as well as the ATA (Air transportation Association, a Aircraft numbering system) chapters. Pilot and Maintenance reports tend to have different perspectives on the condition of the aircraft or descriptions of a fault. In the two Pilot reports they describe the system response outputs experienced in the flight deck, however the four Maintenance reports describe component condition. Furthermore, in conventional reliability trend analysis ATA chapters are the primary method for identifying aircraft or fleet wide emerging problems, however in this case the ATA chapters are from 27 (flight controls) and 29 (hydraulic power), which would reduce occurrence counts if only analysing ATA chapters.

Overall, the NLP framework is able to effectively extract and cluster relevant reports to hydraulic defects, which is recurring approximately over a forty day period, beyond maintainer shift times and aircraft positioning, with some potentially interrelated defects with actuators. If the similarity threshold value were to increase to 0.7, then reports #2, #3, #5 and #6 would be filtered out, leaving #1 and #4, which clearly shows a hydraulic recurring defect. Given the risk rating based on a 'replace' corrective maintenance action word and occurrence rate, this would be indicative of a lead indicator for a problematic component. This is demonstrated with subsequent pilot and maintenance reports #5 and #6 respectively, where replacement is planned and carried out days later for hydraulic system components. Such recurring defects could have potentially been addressed and planned for advance.

4. Discussion

The application of NLP in aircraft fleet management presents a step-change in streamlining and enhancing the efficiency of maintenance operations. As the case study demonstrated in the previous section, NLP was able to rapidly identify reports leading to a significant recurring defect. In such a case, normally this would be a labour-intensive process, with maintainers relying on manual analysis gathering relevant reports and using their experience and intuition to diagnose issues. NLP, with its ability to process, interpret, and analyse large volumes of unstructured data, can help reduce the time required for data analysis.

Table 2 – Example of reporting with recurrent actuator defect and lead indicators.

#	Date	Type	ATA	Report Text	Word	Action	Risk	Similarity
1	30/Jul	Maint	29	HYDRAULIC TUNNEL DIRTY AND CONTAMINATED WITH OLD HYDRAULIC FLUID. ACTION: HYDRAULIC TUNNEL EXAMINED. NO LEAK VISIBLE. TUNNEL CLEANED WITH HOT WATER.	hydraulic	examine	Low	0.80
2	4/Sep	Maint	27	PERFORM BACKLASH CHECK OF L/H WING MID ACTUATOR AND INBD ACTUATOR ACTION: 1) LH WING FLAP MECHANICAL ACTUATOR INBOARD HAS ALREADY BEEN REPLACED. 2) LH WING FLAP MECHANICAL ACTUATOR MID ACTUAL BACKLASH. WO ISSUED FOR REP. INSPECTION - ACTUATOR NUT CAN STAY ON ACFT 1000 FC MORE.	actuator	replace	medium	0.63
3	4/Sep	Maint	27	PERFORM BACKLASH CHECK OF RH WING INNER FLAP ACTUATOR ACTUAL VALUE. ACTION: FUNCTIONAL CHECK OF FLAP ACTUATOR NUT BACKLASH ON RH INNER FLAP ACC PERF.	flap	check	Low	0.55
4	8/Sep	Pilot	29	HYDR QTY NO.2 SYS INDICATION ONLY PARTLY READABLE AFTER PWR CHANGE. ACTION: SYSTEM CHECKED, FOUND NORMAL OPS.	change	change	high	0.71
5	9/Sep	Pilot	29	HYDRAULIC QUANTITY INDICATION 2 UNRELIABLE. ACTION: A/C DISPATCHED ACC MEL *HYD FLUID INDICATION 2 INOP* ITEM SET ON B/C SECTION HIL MAINT PROCEDURE PERFORMED. HYD CTRL PANEL REPLACED. TEST OK ITEM CLEARED FROM B/C SECTION HIL INSTALLED IN S/A CONDITION FROM LVD.	hydraulic	replace	high	0.42
6	11/Sep	Maint	29	HYD QTY INDICATION 2 UNRELIABLE. ACTION: HYD PX QTY INDICATOR ASSY REPLACED.	hydraulic	replace	high	0.62

Moreover, NLP can assist maintainers by providing evidence-based insights that support their intuitive decision-making and strengthen submissions to review the performance of maintenance programs. By analysing historical data, NLP systems can identify patterns and lead indicators that might not be immediately obvious, enabling technicians to predict potential issues before they become critical.

The NLP framework presented is a proof-of-concept and requires further enhancements to overcome some limitations. In this case study, a pilot report was mistakenly marked as ‘high’ risk for a check. This was erroneous due to the system prioritising the term "change" over "check" in the risk rating logic for maintenance actions However, the context of ‘change’ in this report was different and should not have led to this ‘high’ risk classification. For future work, there is opportunity to build a training dataset of similar recurrent defects and recommend early to maintainers that an aircraft has a probability, with a level of confidence, of following a similar trend.

The word cloud shown in Figure 5 summarises the analysis of qualitative feedback from eight respondents who were asked to evaluate the NLP framework and dashboard. This highlights the key priority areas that are of particular interest to maintainers in aircraft fleet management. The core function of most respondents was engineering analysis. The word cloud represents the top 50 words from the feedback forms, where words with larger font sizes appear more frequently than smaller ones. Dominant terms like "maintenance", "program", "reliability" and "parts" emphasise the focus on maintenance program optimisation. The frequent appearance of "data" and "trends" indicates a significant interest in data-driven approaches and trend analysis for proactive maintenance. Additionally, terms such as "failure", "risk" and "replacement" indicate the critical need for effective risk management in part replacement strategies.

Overall, the word cloud reflects the needs and integration of data analytics for proactive manage-

Figure 5 – Feedback on Alerts Dashboard from maintainers and engineers.

ment in maintenance and reliability programs. Examples of the verbatim feedback supports this analysis, for example respondents wrote "...maintain a robust reliability program capable of optimising spares parts inventory for both schedule and unscheduled events.", and "...parts forecast prediction for on time logistics replacement.", also adding it would be useful "highlighting Line Replaceable Units (LRUs) with repeat defects and improving identification of trends (reduce manual processing/identification)." Confirming that unstructured text fields contain information that supports engineering decision-making, but in present form remains a manual process without NLP tools.

5. Conclusion

This paper demonstrates the capability for NLP to process large amounts of aircraft textual reports in minutes for the purposes of fleet management analysis and generate new insights, not readily available due to the laborious manual barrier of analysing text. The novelty of this research is in the application of the NLP and ML framework, including the development of risk ratings and identification of similar lead indicator defects. As well as demonstrating the capability for new correlations to be quantified across an aircraft fleet, supporting qualitative human intuition and maintainer heuristic decision-making.

Future work will investigate Large Language Models (LLMs) (*e.g.*, Bidirectional Encoder Representations from Transformers (BERT)) for use with ML classification methods to automatically identify reports that are ‘high’ risk and worthy of further inspection and greater characterise the sentiment of maintenance narratives. Such a ML approach goes beyond word-level features and enables more nuanced contextual analysis when the model is fine-tuned to predictive maintenance applications.

6. Acknowledgements

Thank you to the Australian Defence Science and Technology Group (DSTG) and the Capability Acquisition and Sustainment Group (CASG) for their support in this collaborative research with RMIT University. The authors would like to thank and acknowledge the contributions of the NLR (Royal Netherlands Aerospace Centre - Koninklijk Nederlands Lucht- en Ruimtevaartcentrum) Aerospace Vehicles Integrity & Lifecycle Support (AVIL) Team. This research was also supported by the Defence Science Institute, an initiative of the State Government of Victoria, under Research Higher Degree (RHD) Student Grant number 35761.

7. Contact Author Email Address

mailto: michael.scott@rmit.edu.au

Michael J. Scott received a BEng (Hons I) from RMIT University, Melbourne, Australia. He is undertaking PhD research into Aircraft Structural Prognostics and Health Management, with a focus on structural diagnostics, mixed-reality and predictive maintenance.

8. Copyright Statement

References

[1] Michael J. Scott, Wim J. C. Verhagen, Marie T. Bieber, and Pier Marzocca. A systematic literature review of predictive maintenance for defence fixed-wing aircraft sustainment and operations. *Sensors*, 22(18):7070, 9 2022.

[2] Nathaniel C Rigoni. Natural Language Processing for Identification of Ground Truth Events in Data Curation. *AIAC 2023: 20th Australian International Aerospace Congress*, 2023.

[3] Rodrigo L. Rose, Tejas G. Puranik, and Dimitri N. Mavris. Natural language processing based method for clustering and analysis of aviation safety narratives. *Aerospace*, 7(10), 2020.

[4] Farhad Akhbardeh, Travis Desell, and Marcos Zampieri. Maintnet: A collaborative open-source library for predictive maintenance language resources, 2020.

[5] Aircraft Commerce. Eliminating the problem of repetitive aircraft technical defects faults. *Aircraft Commerce*, (125), August/September 2019. Issue No. 125.

[6] Ayaka Miyamoto, Mayank V. Bendarkar, and Dimitri N. Mavris. Natural language processing of aviation safety reports to identify inefficient operational patterns. *Aerospace*, 9(8), 2022.

[7] Chuyang Yang and Chenyu Huang. Natural Language Processing (NLP) in Aviation Safety: Systematic Review of Research and Outlook into the Future. *Aerospace*, 10(7), 2023.

[8] Jon Ricketts, David Barry, Weisi Guo, and Jonathan Pelham. A scoping literature review of natural language processing application to safety occurrence reports. *Safety*, 9(2), 2023.

[9] Patrick Jonk, Vincent de Vries, Rombout Wever, Georgios Sidiropoulos, and Evangelos Kanoulas. Natural language processing of aviation occurrence reports for safety management. In *Book of Extended Abstracts for the 32nd European Safety and Reliability Conference*. Research Publishing Services, 2022.

[10] Irene Buselli, Luca Oneto, Carlo Dambra, Christian Verdonk Gallego, Miguel García Martínez, Anthony Smoker, Nnenna Ike, Tamara Pejovic, and Patricia Ruiz Martino. Natural language processing for aviation safety: extracting knowledge from publicly-available loss of separation reports. *Open Res Eur*, 1:110, February 2022.

[11] Christer Stenström, Mustafa Aljumaili, and Parida Aditya. Natural language processing of maintenance records data. *International Journal of COMADEM*, 18:33–37, 04 2015.

[12] Keyi Zhong, Tom Jackson, Andrew West, and Georgina Cosma. Natural Language Processing Approaches in Industrial Maintenance: A Systematic Literature Review. *Procedia Computer Science*, 232:2082–2097, January 2024.

[13] Martin Kála, Andrej Lališ, and Tomáš Vojtěch. Analyzing Aircraft Maintenance Findings with Natural Language Processing. *Transportation Research Procedia*, 65:238–245, January 2022.

[14] Emily Joback, Cynthia Engholm, Q. Chelsea Curran, Yan Glina, Erin Mitchell, Allison Chang, and C. Brian Bassham. Application of natural language processing techniques to kc-135 aircraft maintenance data. *Military Operations Research*, 28(3):pp. 67–86, 2023.

[15] Syed Meesam Raza Naqvi, Mohammad Ghufran, Safa Meraghni, Christophe Varnier, Jean-Marc Nicod, and Noureddine Zerhouni. CBR-Based Decision Support System for Maintenance Text Using NLP for an Aviation Case Study. In *2022 Prognostics and Health Management Conference (PHM-2022 London)*, pages 344–349, May 2022. ISSN: 2166-5656.

[16] Farhad Akhbardeh, Travis Desell, and Marcos Zampieri. NLP Tools for Predictive Maintenance Records in MaintNet. In Derek Wong and Douwe Kiela, editors, *Proceedings of the 1st Conference of the Asia-Pacific*

Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, pages 26–32, Suzhou, China, December 2020. Association for Computational Linguistics.

[17] Sarvesh Sundaram and Abe Zeid. Technical language processing for Prognostics and Health Management: applying text similarity and topic modeling to maintenance work orders. *Journal of Intelligent Manufacturing*, February 2024.

[18] Dave Nakata. An Introduction to MSG-3. SAE Technical Paper 841481, SAE International, Warrendale, PA, October 1984. ISSN: 0148-7191, 2688-3627.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.