Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 277 (2026) iii-xxvi

Procedia Computer Science

www.elsevier.com/locate/procedia

Table of Contents

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 277 (2026) 2011–2020

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Industry of the Future and Smart Manufacturing
(former International Conference on Industry 4.0 and Smart Manufacturing)

On the use of Natural Language Processing for improving automatic maintenance report labelling in the aviation field

Mattia Mauri[a], Roberto Sala[a]*, Duarte Dinis[b,c], Fabiana Pirola[a]

[a] *University of Bergamo, Department of Management, Information and Production Engineering, Viale Marconi, 5, Dalmine (BG), 24044, Italy*
[b] *UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal*
[c] *CEGIST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

Abstract

The use of Natural Language Processing (NLP) is spreading more and more everyday and in almost every sector; it is applied for example to automatize activities, improve efficiency, and support data-driven decisions. As with other sectors, in view of the digital revolution brought by Industry 4.0, maintenance in aviation should be supported by NLP in many different activities; precisely, this paper aims to demonstrate the importance of it in the improvement of data collection and analysis. In fact, the host Maintenance, Repair, and Overhaul (MRO) company counts with an enormous quantity of maintenance data available that, at the moment, are not suitable for data analysis due to the presence of many errors in classification; the analyzed models have the objective to support MRO companies in classifying data into the correct physical zone of the aircraft.

Keywords: Aircraft Maintenance; Maintenance Management; Natural Language Processing; Decision Support Systems

1. Introduction

The aviation sector is among the most technology intensive ones [1] and it is a global oligopoly, since just few companies produce civilian and / or military aircraft. From an economic standpoint, data show a continuous increase

* Corresponding author. Tel.: +39 0352052005
E-mail address: roberto.sala@unibg.it

in terms of generated value and, in 2023, it has been estimated that it represents 4.1% of global GDP [2]. Costs, monetary losses and safety related to aircraft are extremely impactful, therefore maintenance plays a crucial role in this sector even more than in others. In general, maintenance is the process of keeping something in operation at the same level of reliability, such as industrial machinery or other productive equipment, and it is divided into two main types: preventive, also known in aviation as scheduled maintenance, and corrective, known in aviation as unscheduled maintenance [3].

Over the years, the role of maintenance became more and more strategic for companies, as a well performed maintenance contributes to their competitiveness, especially in view of the digital revolution brought by Industry 4.0. The greater availability of data and the development of new, advanced, hardware and software means allowed companies to develop new, advanced and data-based, maintenance approaches. The digital transformation allowed to collect increasing quantities of numerical data, to be analyzed with machine learning (ML) approaches, but also allowed to digitalize and collect textual data, like the ones generated before, during, and after the maintenance intervention by the technicians. Such data, frequently unstructured (e.g., technician reports, maintenance logs, etc.), require specific approaches to be analyzed.

Natural Language Processing (NLP), a branch of Artificial Intelligence (AI) capable of interacting of human-language [4], can be used to analyze unstructured textual data , optimize, and automatize lengthy processes that may suffer from human errors. Such approaches can be helpful in the context of maintenance management as they increase and standardize document processing, allowing companies to create a standardized knowledge base to be used for descriptive, predictive and prescriptive purposes within decision support systems. This work aims to analyze the implementation of predicting models based on NLP algorithms to support the collection and classification of data related to maintenance performed on aircraft. The necessity derives from the enormous amount of data collected by the host Maintenance, Repair and Overhaul (MRO) company that cannot be exploited given the existence of classification errors in the dataset due to the human technicians. After terminating the maintenance action, workers must fill out the dataset with a label representing the physical area; in doings so, operators often insert the wrong zone for different reasons (i.e. typing mistakes, low concentration, or difficulties in remembering the correct code). Thus, despite having a useful and available set of data, the company is not able to exploit it due to errors in the dataset. The aim, with this work, is to test whether the classification process of intervention can be automatized by means of NLP approaches and, if so, use the processed data to run analysis and extract useful statistics.

The paper is structured as follows: Section 2 briefly explains the literature related to this topic; Section 3 gives an overview of the methodology applied to reach the final objective; Section 4 clearly details the case study from the first phase to the achievement of final result; Section 5 discusses the application of the models in the host MRO; and Section 6 is dedicated to conclude the paper.

2. Literature review

NLP is a subfield of AI and Computer Science that makes use of ML to allow computers to comprehend and communicate with textual data [4]. NLP has been introduced in several sectors. In aviation it has been used to analyze data, optimize maintenance and build safety reports [5].

The main objective of NLP is to ease the interaction between human-users and computer based tools; more precisely, the goals include interpretation, analysis and manipulation of natural language data for the intended purpose with the use of various algorithms, tools and methods [4]. Consequently, as with other sectors, maintenance activities could be performed better with the support of NLP: efficiency can be improved for the partial elimination of human work and the consequent decrease in needed time; accuracy can be higher due to lower likelihood of errors; and, finally, more data-driven decisions could be taken thanks to its functionalities [6][7][8].

Major American airlines have tried to adopt Sentiment Analysis in order to collect data from an enormous number of Tweets written by final-users [9]; in fact, customers very often post regarding their experience, but this is not directly usable for data analysis due to their unstructured format. Researchers have used the library ‘Tweepy’ and performed a deep Pre-Processing phase; the desired outcome was to classify whether a tweet was positive or negative, and this has been predicted correctly for 91.86% [9].

During a conference at Aviation Week’s Aerospace IT (Chicago, October 6th, 2022) Tassio Carvalho, American Airlines’ senior manager of operations research, has explained that NLP could be introduced to support everyday

activities for many different reasons. This ranges from speech recognition to instantly transcript machinal descriptions, to the support in the creation of well-structured dataset where operators could record how they operate; here there is a strong relationship with this research work. NLP is also exploited for the construction of knowledge graphs from maintenance data [10]; more precisely, the objective is to use Echidna and MWO2KG to transform chaotic and unstructured data into intuitive graphs. In this case, words are transformed into numerical representations by Word Embeddings technique [11].

Safety is for sure the most delicate issue to face in aviation and it takes a lot of effort from working personnel; in the last years, with a small bracket related to Covid-19, demand for flying has kept growing while available work force has not followed the same trend. For this reason, there have been attempts to automatize actions such as analysis or categorizations where possible. It is the case, for instance, of the attempt to categorize descriptive incident reports into three different categories [12] that are the followings: Unsafe Supervision, Precondition for Unsafe Acts, and Unsafe Acts according to Human Factor Analysis and Classification System (HFACS) [13]. The main algorithm applied in this case is Labeling Spreading, after a precise and iterative phase of Pre-Processing. Similarly to this last work of research, there is an example of extraction of Cause and Effect entities as well; more precisely, the final aim is to analyze long maintenance documents, Work Request Notifications (WRN), in order to extract, collect and analyze entities [14].

It is fundamental to mention the current scenario regarding the implementation of Technical Language Processing (TLP)[15][16][17]. As it is understandable from the acronym, it is a sub-field of NLP specialized in the comprehension of technical language [18]. This comes from the lack of this branch of language in NLP; in fact, in various research works, this one included, one of the main issues is to train models on the base of data that is not composed by everyday common vocabulary. TLP should be implemented with the support of technical operators too; as a matter of fact, they should classify elements in the right corresponding category, and then experts should collaborate in order to create an efficient process of Named Entity Recognition (NER) [19] because it does not exist in the field of maintenance.

From this brief literature review it is clear that NLP is spreading in the aviation industry, as well as in other industries; despite this, there are many issues and challenges still to be faced such as the lack of technical language to train models in the automatic recognition of physical parts of aircrafts. Therefore, this research work aims to propose the application of NLP to improve the overall performance in aviation maintenance by increasing the effectiveness and efficiency of data-driven decisions.

3. Research approach

This section explains how the research has been conducted. This work of research adopts a computational and data-driven approach, that has been structured in different phases: data retrieval, analysis of dataset structure, pre-processing, identification of different approaches, models' building, and results' analysis. The methodology is based on Natural Language Processing (NLP) techniques that are exploited to extract, transform and analyze textual data aiming to gain knowledge from it. The host MRO is a Portuguese company operating in Civil Aviation; maintenance works comprehend both internal and external structure of aircrafts and, once action is terminated, details related to the work are manually filled out by technicians and stored in a database.

Each maintenance activity is associated with the aircraft zone where it has taken place. Such zones are identified by unique numerical codes composed of three digits (e.g., zone 100 is Lower Fuselage while 110 and 120 are subzones of zone 100. In turn zones 111 and 112 are subzones of zone 110[3]). One of the main problems that the host company is facing is the mislabeling of the maintenance activities, mainly due to the association with the wrong zone. The supervisor is responsible for taking care of the correctness of the maintenance activity, a task that, depending on the amount of work executed might take a considerable amount of time. The objective of this research is to verify whether Natural Language Processing (NLP) can be used to automatically assign the correct zone code to the maintenance task, based on the description that the technicians manually enter in the system.

The research has been carried out by going through the following steps (shown also in Fig. 1):

1. *Data retrieval.*
2. *Analysis of the dataset structure.* Where the authors analyzed the dataset structure and removed all non-useful columns and information.

3. *Selection of the approaches to test.* Where the NLP approaches to test were identified. In particular, an approach based on Logistic Regression [20] and another one based on Word Embeddings [11].
4. *Analysis of Labels distribution and Data Pre–Processing.* Where data were prepared for the analysis through common NLP techniques such as stop-word removal and jargon standardization. Additionally, labels were adjusted and, following the distribution analysis, the dataset was balanced.
5. *Application of the approaches.* Where the authors have built three different models, this has been executed in order to analyze them and see which one would bring better results in the actual production. They have started from the most "conventional" one and concluded with the one considered the most advanced. Logistic Regression for predicting single label descriptions has been the first model; then Logistic Regression for predicting multi label descriptions; and, finally, authors have concluded with a Word Embeddings approach. It is important to mention that, despite higher expectations, Word Embeddings approach has brought lower performance rates because of the lack of topic-related documentation needed for the training.
6. *Results comparison.* Where performance rates have been extensively analyzed and compared.
7. *Identification of improvement actions.* Where additional changes to the structure of data collected have been identified and proposed to reach improved results.

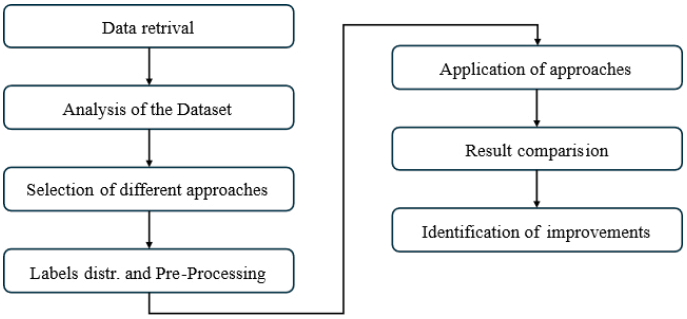

Fig. 1 - Research approach.

4. Case study

This section aims to present the application of the methodology described in Section 3. As already stated, the host company operates in the aviation MRO sector; once the operators finish a maintenance task, they fill out a form where they enter a description of the task, a description of the activity, assign the zone code to the work, and enter other information not relevant for the current analysis. The data entered are then stored into a database from which the dataset under analysis has been extracted. From this point onwards, the authors will refer only to the four columns used in the performed analysis: "description of anomaly", "description of corrective action", "zone-original" (filled out by the technicians), and "zone-corrected" (in which wrongly assigned zones have been corrected by experts). The aim of this research is to test the use of NLP to automate the classification task, preventing classification errors and providing a well-structured database for the company. Once analyzed, data can be extremely useful for various reasons, ranging from the recognition of most problematic parts of aircrafts, to the collection of most helpful solutions that technicians can take advantage of before starting a new work[3]. The aim is, therefore, to train an NLP model able to assign the correct aircraft zone to the maintenance activity, based on the content of the "description of anomaly" and "description of corrective action" columns.

Anomaly	Corrective Action	Zone - Original	Zone - Corrected
WING LH Upper shroud with c	Remove the corrosion and aplly alodine 120	541/641	530/630
ENGINE PYLONS AFTER REPAIR FAIRING P	DRILL, CUNTERSINK, AND INSTALL NEW RIV	414/424	410/420

Fig. 2 - Host company dataset example.

4.1. "As-is" situation

The dataset is an Excel file composed of 13 columns and 62508 lines, with each line representing a different maintenance work order, or task. Out of the 13 columns, only the four mentioned before were used for the analysis (Fig. 2). If well classified, "zone-original" and "zone-corrected" should contain the same Label (or group of Labels). A brief explorative analysis allowed to identify thousands of misalignments between these columns, which forced the supervisor to spend an enormous amount of time correcting labels. The main problem is that despite the large data availability, if not corrected, such data cannot be used for analysis, preventing the host company from gaining useful knowledge from the data.

The authors used the Python coding language and the related libraries, such as 'Pandas' [21], to run the NLP analysis; Python was selected due to the vast availability of libraries developed for NLP tasks. First, the label distribution was analyzed using the'Matplotlib' [22] and 'Pandas' [21] libraries. Results showed that the distribution of labels was very unbalanced and chaotic (Fig. 3): zone 240 (the most common) was present 5834 times, while zone 220/510 (the least common), just once. This could have prevented the model from being trained correctly, limiting thus its usefulness. To solve this, the Oversampling technique was used; it will be better explained in the next section. Another aspect of interest was the presence of multiple labels associated with a single maintenance task. This forced the authors to consider also models able to handle multi-labelled data. In addition, a small amount of alphabetical labels was identified. These have been removed from the dataset because they were not meaningful for the analysis.

4.2. Pre-Processing and Data Cleaning

In this section, the 'Pre-Processing' phase will be illustrated; the primary goal of this phase was to transform raw data, filled out by human-beings, into a more structured format that can be more easily used by machines [23]. The authors have initially started from the most common cleaning actions and Pre-Processing phases for a dataset and then, through different attempts, found the most suitable and helpful for this specific case. In Fig. 4, it is shown an example of the steps implemented for Pre-Processing and Data Cleaning on a generic sentence.

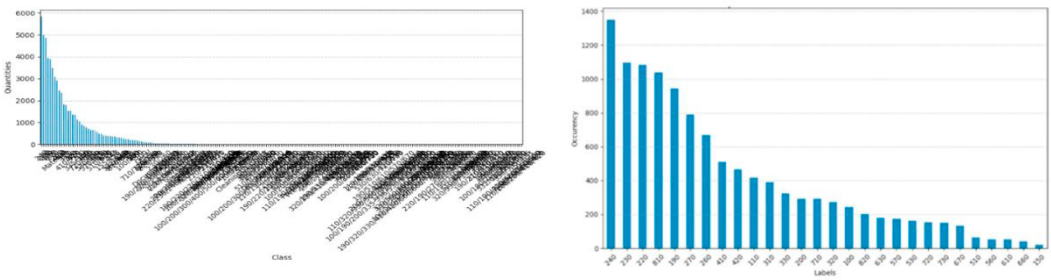

Fig. 3 - Distribution of classes in the 'as-is' scenario vs Distribution of classes after Pre-Processing

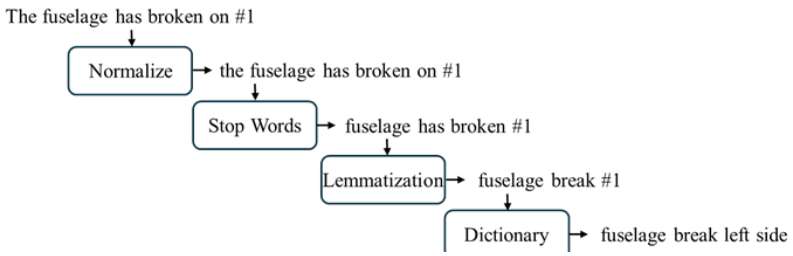

Fig. 4 - Example of Pre-Processing and Data Cleaning.

The research has started with the implementation of simple models where no specific Pre-Processing or other actions have been implemented to understand the necessities of the models; of course, better results have been reached with the implementation of Pre-Processing:

1. *Stop Words.* These are the words that are not considered relevant in giving meaning because they are mainly used for grammatical reasons. Their removal is fundamental in order to increase the signal-to-noise ratio in unstructured data [24]; since models aim to recognize specific parts of aircraft, some descriptive words that are present in the list might be useful. An example of the stop words removed from the list used in the analysis is the following: "under, empty, over, down, after, no, not, one, two, three, four, five, six, eight". The authors used Scikit – Learn [25] for this task.
2. *Lemmatization.* It is defined as an algorithm technique of finding the lemma of a word which is a root word rather than a root stem [26] and it is considered a standard phase of Pre-Processing; in this research, its application has actually brought lower results so it has been excluded from the process. This has been caused by the presence of technical words, whose meanings are better understood by models in their original form.
3. *Small dictionary introduction.* By exploring the dataset, it was noticed that operators used to identify specific parts of the aircraft like "left side" and "right sides" with numerical codes such as "#1" and "#2". In addition, they were used to use acronyms such as "AMM" or "IAW" instead of the full wording to refer to documentation or other zones. In response, the authors, with the support of an expert, created a dictionary to be used to substitute all the acronym with the "full" description, as this would have helped to standardize the vocabulary and improve the analysis.
4. *"0" replacement.* As mentioned before, labels are three digits number and, moving to the right, the level of detail increases; for instance, zone 412 where '4' stands for engine, '1' for left side and '2' in a specific area of it. In this research, it has been agreed that second digit precision would be enough, therefore all the labels have been transformed into label terminating with digit 0 with a specific script. This implementation has also brought a decrease in the number of labels to 28.
5. *Oversampling.* Even if the total number of classes is 28, dataset is not balanced and Oversampling is necessary because least common classes are the most problematic to predict. As previously said, Oversampling is the process of increasing the amount of smaller classes instances by producing some new instances or by repeating existing ones [27] For this purpose, it has been used the function *RandomOverSampler* included in the library Imbalanced-learn [28]. By doing so, all the classes have the same number of occurrences and models can be better trained. An example is reported in Fig. 5.

4.3. Single–Label Approach

In this section, the Single-Label Classification approach is described; this refers to the classification in which each datum belongs to only one class for the requested model [29]. It has been decided to first start with a simpler model in order to better comprehend issues and opportunities of the dataset. After multiple Pre-Processing steps, it has been decided to use Logistic Regression to determine to which class a datum belongs to [20]. This has been achieved with the support of Tf-idf Vectorizer and performance metrics present in the Scikit Learn [25] library. Tf-idf is one of the

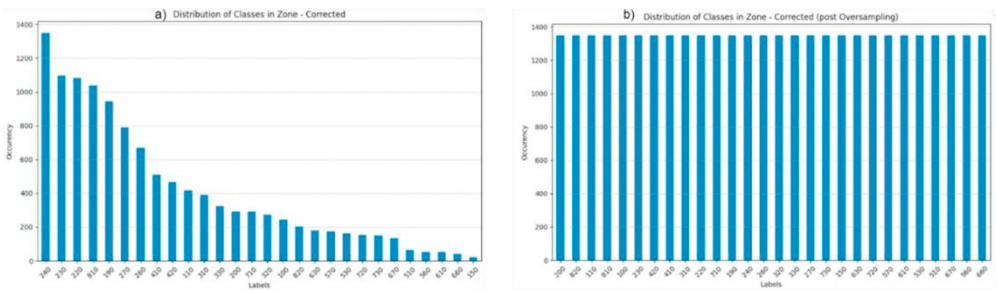

Fig. 5 - a) Labels distribution before Oversampling, and b) Labels distribution after Oversampling.

many techniques that can be used to transform words in a format that is usable by models; more precisely, it is a weighting function for the importance of words composed by two measures, and it translates words into numbers [30]. Once the algorithm is applied to descriptive columns of the dataset, words are transformed into numbers, and the construction of the model can start; 80% of the data is now used to train the model and the remaining 20% is used to test it. It is important to underline, again, that the model aims to predict the correct numerical label, referring to a specific physical part of the aircraft, by recognizing words, that have been transformed with Tf-idf Vectorizer, in the descriptive parts. Results will be better discussed in the specific section, but it is worth mentioning that the final model for single-label predictions has reached an accuracy equal to 95.84%.

4.4. Multi-Label Model

In this section, the Multi-Label classification model is described; this is the scenario where each maintenance task, can be associated with one or more classes describing different parts of aircraft [29]. Since labels are manually inserted by technicians, they are composed of many division characters between them. Therefore, together with previous actions, it has been necessary to transform all the labels in a standard form.

For this task, the Label Powerset algorithm [25] was chosen. Also, other approaches were considered, but Label Powerset was chosen due to its higher performance. Label Powerset transforms the Multi Label Classification Problem into different multi-class Single Label Problems; in fact, it considers each unique combination of Labels as a class. For this reason, it can get very complex as the number of Labels increases [31]. At the same time, it is very useful since it is capable of comprehending dependencies between different labels. In addition to this, the Oversampling technique has been applied to improve overall performances. It is worth mentioning that, differently from Oversampling in Single-Label Models, the number of each label is not balanced, since the balance occurs in terms of lines, i.e., combinations of labels. Label Powerset algorithm, together with Oversampling techniques and previous Pre-Processing actions, has resulted in being the best model for predicting labels in this specific research with a final accuracy of 93.48%; even if the accuracy in Single-Label Model was higher, the model itself was not as complete as this.

4.5. Word Embeddings Approach

Eventually, the Word Embeddings approach was tested. Word Embeddings represents words as vectors in a multi–dimensional space. More precisely, distances from vectors are used to identify semantic and syntactic similarities of words in a given context [11]; models are trained on pre-existing documents such as Google News or similar. For this research work, the model “Word2Vec” [32], with the library “Gensim” [33], was used. Despite its potential strength, this model has resulted in weaker performance due to the lack of topic-related documentation to train such a specific model as this one applied in the aviation sector.

4.6. Results

This section is dedicated to an analysis and comparison of results. Performance rates have been calculated using: Accuracy, which is the closeness between the found value and the true value, and it is fundamental to decide the degree of confidence that can be given to results [34]; Precision, which is the representation of how close the measurements of the same item are between each other; F-1 Score, which is the harmonic mean between Accuracy and Precision; Hamming Loss, used only in Multi-Label Models because it considers each single label inside of an instance with more than one label [35]; and absolute numbers of wrongly predicted labels. In Table 1 results are summarized and, for each model, only the metrics of the best one are presented; therefore Oversampling dataset for Single-Label Model, Label Powerset algorithm on oversampled dataset for Multi-Label Model, and training on Google News for Word Embeddings.

The MultiLabel model was selected as the best one, given the requirements of the host company, despite the slightly lower performance compared to the Single-Label, as multiple labels could be associated to the same maintenance task.

Table 1: Best Results for each Model

Metrics	Single-Label Model	Multi-Label Model	Word Embeddings
Accuracy	95.84%	93.48%	67%
Precision	95.78%	N.A.	N.A.
F-1 Score	0.95	0.97	N.A.
Hamming Loss	N.A.	< 1%	N.A.
Nr of wrong terms	7246	853	N.A.
Nr of right terms	314	12242	N.A.

5. Discussion

The objective of this research is to determine if and how the application of NLP could be helpful and supporting in the data management functions in the aviation MRO sector. In the current scenario, not only data are difficult to use due to errors in classification, but they also require a substantial amount of supervisor's time to check and correct them. With the found performance rates, predicting Model, especially Multi-Label one, could be implemented for support technicians inserting the right label; the developed models are capable of recognizing descriptions and predicting right labels more than 90% of cases. As it is visible in Table 1, accuracy has reached a value greater than 95%; in this situation, the utility of the dataset has improved significantly since the model can almost always predict the right label for the work of maintenance. With the full practical implementation, this model could become an actual supporting tool for operators in everyday activities.

This supporting tool should not be considered only in giving support through the fulfillment of the correct labels but, having a well-structured database, it could be helpful before performing a new maintenance intervention. For instance, moving to a practical example, consider a group of workers that must intervene a problematic situation localized on the lower part of the right engine; before performing the maintenance action, they could filter the dataset for the label referring to that specific part of the aircraft and take advantage of all previous recorded actions. The final result would be a knowledge base that could be used by the host company for several scopes as a supporting tool for technicians, material for training new employees and base of data for analyzing most problematic aircraft zones.

As just mentioned, the new structured dataset could be used to analyze most problematic zones in the aircraft and, in the long period, try to study them to anticipate issues instead of solving them. In all sectors, data collection and analysis is extremely important to gain competitive advantage by understanding where and how to improve; but in the aviation sector it is even more important due to the costs and losses related to grounded aircraft. Of course, before costs and losses, there is safety, and it could be improved as well.

Performance metrics could increase even more with a few changes from the operative side that derives from an analysis of Models' performances. First of all, it has been studied which are the most relevant words in each zone by using Chi2 Technique [36], which allows to measure relationships between categorical variables such as words; this has highlighted words that are more recognizable by Models therefore operators should be encouraged to comprehend them in their descriptions for further improvements. Then, abbreviations need to be mentioned; even if there has been the introduction of a small dictionary, there are more and more that are deeply used for descriptions and, without a complete dictionary, they create issues. The host company should build a complete dictionary to be applied in the predicting model for recognition of those. Similarly, workers should be inserting descriptive texts in a standardized way; at the moment, descriptive texts are chaotic since they are filled quickly. There should be a standardization in labels entry as well; now in multi-label lines, they are inserted with too many separators characters between them, which requires a cleaning before using machines.

Moving to labels, it has been analyzed which are the most problematic to predict and they have resulted to be the ones related to the fuselage part; after a deeper search, it has been highlighted that both external and internal parts of fuselage are associated with same group of labels and so it is really hard to recognize them for both technicians and machines. New labels should be created to better identify these critical areas.

For such an important company, this list of suggestions should not be hard to study and implement. Together with those, NLP Models could bring enormous advantages and an important gain in overall efficiency in such a complicated industry where it is fundamental to keep technology and innovation at its maximum level possible. As conclusion of this section, it is true to say that potential advantages in implementing the developed NLP models outweigh the fatigue necessary to implement an accurate and precise database by hand.

6. Conclusion

Considering the previously listed results, it is possible to affirm that the implementation of NLP models could actually bring competitive advantages to a sector where technology and innovation are between every day's necessities. Methodologies and approaches have perfectly fit the scenario since the highest score is close to 95%; this result can be considered sufficiently high due to the fact that the research aims to propose a supporting tool and not the total automatization of the process. This complete automatization could be achieved including a few different changes, with TLP at the top, since most of the problems occurred with technical language. Better results could be reached also with a strong collaboration with technicians; in fact, many actions of Pre-Processing have been necessary due to the presence of many errors caused by human-beings. Sorts of regular expressions should be created to have a better organized dataset ready to be used by models.

Despite this, from a theoretical point of view, it is important to highlight results; in fact, better results have been reached with algorithms that are considered less advanced than others: Logistic Regression has brought higher rates than Word Embeddings. Therefore, the general approach for building predicting models has resulted in great performances in both single-label and multi-labels situations.

From the practical point of view, this research has confirmed the utility of NLP in real companies; this Paper is concentrated with maintenance in the aviation sector. Nonetheless, it should be applicable to other maintenance sectors as well. In fact, a model capable of predicting physical zones of an object could be used with other means of transport such as cars, trains or ships; specific industrial machinery; and with large infrastructures. Each scenario would need to be trained with a strictly topic-related dataset; as seen in the research, this is the main issue when there is the presence of technical language. Again implementation of TLP is fundamental for future progresses.

In conclusion, considering both points of view, the Paper has reached its target of demonstrating the applicability and utility of NLP for improving classification process of maintenance actions performed by an aircraft MRO. The host company should implement this model in order to gain competitive advantages through the precise collection and categorization of enormous amounts of data that currently cannot be used in its completeness. Through the use of the developed models, the entire dataset could be used for analysis intended to improve the overall efficiency of the company.

Acknowledgements

Duarte Dinis acknowledges the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for its financial support via the project UIDB/00667/2025 and UIDP/00667/2025 (UNIDEMI). Duarte Dinis also acknowledges FCT for the financial support under the project UIDB/00097/2025 and UIDP/00097/2025 (CEGIST).

References

[1] Knoll-Csete, Edit, e István Kárász. (2021 [citato 1 ottobre 2024]) «The Characteristics of the Aviation Industry». [Internet]Disponibile su: https://www.cambridge.org/engage/coe/article-details/613f813242198e4dd6754490

[2] Raut, Prasanna. «The Economic Impact of Aviation».

[3] Dinis, Duarte, Ana Barbosa-Póvoa, e Ângelo Palos Teixeira. (2019) «A supporting framework for maintenance capacity planning and scheduling: Development and application in the aircraft MRO industry». *International Journal of Production Economics* **218** : 1–15.

[4] Chowdhary, K.R. (2020 [citato 19 settembre 2024]) «Natural Language Processing». in *Fundamentals of Artificial Intelligence* New Delhi: Springer India; p. 603–49. Disponibile su: http://link.springer.com/10.1007/978-81-322-3972-7_19

[5] Yang, Chuyang, e Chenyu Huang. (2023) «Natural Language Processing (NLP) in Aviation Safety: Systematic Review of Research and Outlook into the Future». *Aerospace* **10** (**7**): 600.

[6] Stenström, Christer, Mustafa Aljumaili, e Aditya Parida. «Natural Language Processing of Maintenance Records Data». *International journal of COMADEM*.

[7] Sala, Roberto, Fabiana Pirola, Giuditta Pezzotta, e Sergio Cavalieri. (2023) «Improvement of maintenance-based Product-Service System offering through field data: a case study». *Production & Manufacturing Research* **11** (**1**): 2278313.

[8] Lukens, Sarah, Lucas H. McCabe, Joshua Gen, e Asma Ali. (2024) «Large Language Model Agents as Prognostics and Health Management Copilots». *Annual Conference of the PHM Society* **16** (**1**). Disponibile su: http://www.papers.phmsociety.org/index.php/phmconf/article/view/3906

[9] Wu, Shengyang, e Yi Gao. (2024) «Machine Learning Approach to Analyze the Sentiment of Airline Passengers' Tweets». *Transportation Research Record* **2678** (**2**): 48–56.

[10] Stewart, Michael, Melinda Hodkiewicz, Wei Liu, e Tim French. (2024) «MWO2KG and Echidna: Constructing and exploring knowledge graphs from maintenance data». *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* **238** (**5**): 920–32.

[11] Li, Yang, e Tao Yang. (2018 [citato 30 settembre 2024]) «Word Embedding for Understanding Natural Language: A Survey». in S Srinivasan (a c. di) *Guide to Big Data Applications* Cham: Springer International Publishing; p. 83–104. (Studies in Big Data; vol. 26). Disponibile su: http://link.springer.com/10.1007/978-3-319-53817-4_4

[12] Madeira, Tomás, Rui Melício, Duarte Valério, e Luis Santos. (2021) «Machine Learning and Natural Language Processing for Prediction of Human Factors in Aviation Incident Reports». *Aerospace* **8** (**2**): 47.

[13] Wiegmann, Douglas A., e Scott A. Shappell. (2017 [citato 14 ottobre 2024]) «The Human Factors Analysis and Classification System (HFACS)». in *A Human Error Approach to Aviation Accident Analysis* 1ª ed. Routledge; p. 45–71. Disponibile su: https://www.taylorfrancis.com/books/9781351962360/chapters/10.4324/9781315263878-3

[14] Hershowitz, Brad, Melinda Hodkiewicz, Tyler Bikaun, Michael Stewart, e Wei Liu. (2024) «Causal knowledge extraction from long text maintenance documents». *Computers in Industry* **161** : 104110.

[15] Kála, Martin, Andrej Lališ, e Tomáš Vojtěch. (2022) «Analyzing Aircraft Maintenance Findings with Natural Language Processing». *Transportation Research Procedia* **65** : 238–45.

[16] Amin, Nadine, Tracy L. Yother, Mary E. Johnson, e Julia Rayz. (2022) «Exploration of Natural Language Processing (NLP) Applications in Aviation». *Collegiate Aviation Review International* **40** (**1**). Disponibile su: https://ojs.library.okstate.edu/osu/index.php/CARI/article/view/8566/7733

[17] Miyamoto, Ayaka, Mayank V. Bendarkar, e Dimitri N. Mavris. (2022) «Natural Language Processing of Aviation Safety Reports to Identify Inefficient Operational Patterns». *Aerospace* **9** (**8**): 450.

[18] Brundage, Michael P., Thurston Sexton, Melinda Hodkiewicz, Alden Dima, e Sarah Lukens. (2021) «Technical language processing: Unlocking maintenance knowledge». *Manufacturing Letters* **27** : 42–6.

[19] Zhang, Ying, e Gang Xiao. (2024) «Named Entity Recognition Datasets: A Classification Framework». *International Journal of Computational Intelligence Systems* **17** (**1**): 71.

[20] Sperandei, Sandro. (2014) «Understanding logistic regression analysis». *Biochemia Medica* : 12–8.

[21] McKinney, Wes. «pandas: a Foundational Python Library for Data Analysis and Statistics».

[22] Hunter, John. (2007) «Matplotlib: A 2D Graphics Environment». *Computing in Science & Engineering* **9** : 90–5.

[23] Roy, Swarup, Pooja Sharma, Keshab Nath, Dhruba K. Bhattacharyya, e Jugal K. Kalita. (2019 [citato 2 ottobre 2024]) «Pre-Processing: A Data Preparation Step». in *Encyclopedia of Bioinformatics and Computational Biology* Elsevier; p. 463–71. Disponibile su: https://linkinghub.elsevier.com/retrieve/pii/B9780128096338204573

[24] Sarica, Serhad, e Jianxi Luo. (2021) «Stopwords in technical language processing».Diego Raphael Amancio (a c. di) *PLOS ONE* **16** (**8**): e0254937.

[25] Pedregosa, Fabian, Fabian Pedregosa, Gael Varoquaux, Gael Varoquaux, Normalesup Org, Alexandre Gramfort, et al. «Scikit-learn: Machine Learning in Python». *MACHINE LEARNING IN PYTHON*.

[26] Khyani, Divya. «An Interpretation of Lemmatization and Stemming in Natural Language Processing».

[27] Mohammed, Roweida, Jumanah Rawashdeh, e Malak Abdullah. (2020 [citato 24 settembre 2024]) «Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results». in *2020 11th International Conference on Information and Communication Systems (ICICS)* Irbid, Jordan: IEEE; p. 243–8. Disponibile su: https://ieeexplore.ieee.org/document/9078901/

[28] Lemaıtre, Guillaume, e Fernando Nogueira. «Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning».

[29] Ji, Xiaowan, Anhui Tan, Wei-Zhi Wu, e Shenming Gu. (2023) «Multi-label classification with weak labels by learning label correlation and label regularization». *Applied Intelligence* **53** (**17**): 20110–33.

[30] Robertson, Stephen. (2004) «Understanding inverse document frequency: on theoretical arguments for IDF». *Journal of Documentation* **60** (**5**): 503–20.

[31] Junior, J.D. Costa, E.R. Faria, J.A. Silva, e R. Cerri. (2017) «Label Powerset for Multi-label Data Streams Classification with Concept Drift».

[32] Jatnika, Derry, Moch Arif Bijaksana, e Arie Ardiyanti Suryani. (2019) «Word2Vec Model Analysis for Semantic Similarities in English Words». *Procedia Computer Science* **157** : 160–7.

[33] Řehůřek, Radim, e Petr Sojka. (2010) «Software Framework for Topic Modelling with Large Corpora». in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* Valletta, Malta: ELRA; p. 45–50.

[34] Menditto, Antonio, Marina Patriarca, e Bertil Magnusson. (2007) «Understanding the meaning of accuracy, trueness and precision». *Accreditation and Quality Assurance* **12** (**1**): 45–7.

[35] Wu, Guoqiang, e Jun Zhu. (2020 [citato 8 ottobre 2024]) «Multi-label classification: do Hamming loss and subset accuracy really conflict with each other?». [Internet]arXiv; Disponibile su: http://arxiv.org/abs/2011.07805

[36] Huan Liu e R. Setiono. (1995) «Chi2: feature selection and discretization of numeric attributes». in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence* p. 388–91.