Literature Survey on Aircraft Maintenance Issues with Human Errors and Skill Set Mismatch Using Document Mining Technique

T. Nanthakumaran Thulasy[1(✉)] , Puteri N. E. Nohuddin[2] ,
Norlizawati Abd Rahim[3] , and Astuty Amrin[4]

[1] Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia,
54100 Kuala Lumpur, Malaysia
tnanthakumaran@graduate.utm.my
[2] Institute of IR4.0, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia
[3] Department of Science, Management & Design, Razak Faculty of Technology and
Informatics, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia
[4] Department of Engineering, Razak Faculty of Technology and Informatics,
Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia

Abstract. Aviation encompasses a wide range of mechanical flights as well as the aircraft business. The industry is quickly expanding as travelers choose to travel by air due to the concept of a world without borders, lower rates, and global business activities. Thus, aircraft maintenance and repair are critical jobs in the aviation sector for improved aircraft service and safety. Despite this, numerous aviation disasters occur due to a variety of factors such as malfunctioning equipment, hazardous weather conditions such as turbulence, and human error. In this research, we used text mining to explore a set of internet articles and reports about aviation occurrences and difficulties. The method retrieves frequent and significant terms from the document set. The experiment is divided into four (4) modules: (i) collecting of web articles and reports, (ii) document preprocessing, (iii) text analytics, and (iv) visualization. The experiment's results show that the bulk of records describe aircraft maintenance, skill mismatch, and training shortages.

Keywords: Aviation · Industry revolution 4.0 · Aircraft maintenance issues · Skill mismatch · Document mining

1 Introduction

The combination of Artificial Intelligence and engineering automation is driving the growth of Industry Revolution 4.0. (IR4.0). As a result, it sort of creates surplus and shortage in specific groups of employees and specialists especially in which the workforce requires the replacement of appropriately skilled personnel. Some work tasks can be replaced by machines to reduce production costs. In general, many work skills are acquired through formal schooling (TVET, colleges, and universities), but they can also be acquired through alternative means [1]. The changes in the IR4.0 workforce setting

have necessitated the acquisition of new skills and competences in order to keep up with the advancement of current innovations [2]. The recent situation revealed that many job domains have become extremely demanding, and as a result, current employees' skill sets are insufficient to sustain a significant transformation in the entire area of work. As a result, additional efforts should be made to implant critical job skills in recent graduates and even present employees, so that they are in sync with the current needs of the organization.

Graduates should be exposed to appropriate skill sets in order to guarantee their work and remain employed in a constantly changing industry and innovation climate. As such, the purpose of this article is to support claims that one of the aircraft maintenance skill sets is challenging. The aviation sector is being shaped by IR4.0 to adopt current technologies in its aircrafts and human resources [3]. In the production of aircrafts and related equipment, the aviation sector is moving forward with the use of technology such as artificial intelligence, machine learning, automation, sensors, and remote monitoring. Therefore, maintaining modern aircrafts has become a difficult undertaking and it is critical to guarantee that engineers and technicians have a broad skill set and the best competences. According to the Malaysian Aerospace Industry Blueprint 2030 [4], some education and training programs are not being developed sufficiently, and there is a skills mismatch between academics and industry. Graduates from academies do not meet workforce requirements for the proper talent. Recently, skill mismatch among aircraft maintenance professionals has emerged as a widespread personnel management issue in the aviation industry [5].

The aim of this paper is to outline a document mining experiment conducted on a set of documents pertaining to human factors, issues and aircraft maintenance using a text mining algorithm. The following is how the paper is structured: Sect. 2 goes over the backdrop and associated work on the influence of IR4.0 on aircraft maintenance, skill sets issues and incidents, and document mining or document analytics procedures. Sections 3 and 4 elaborate on document mining framework and the results of document analytics utilizing the proposed methodology. Finally, Sect. 5 concludes with a summary and recommendations for further research.

2 Literature Review

This section covers a variety of topics relating to IR4.0, such as aircraft maintenance and skill set concerns and mishaps, as well as document mining.

2.1 The Effects of IR4.0 on Aviation

IR4.0 ushers in Industry 4.0 by transforming the manufacturing sector to embrace digital transformation through the incorporation of sensing devices in virtually all manufacturing parts, products, and equipment. The implementation of a ubiquitous system enhances the ability to examine digital data and physical equipment, allowing every worldwide industrial sector to progress much more quickly [6]. [7] defines Artificial Intelligence (AI), Big Data, the Internet of things (IoT), the Internet of services (IoS), cyber-physical systems (CPS), and smart production are all listed as essential components of IR4.0.

As a result, this gyration is the culmination of all preceding revolutions' inventions. According to the literature, this revolution will convert and have an irreversible impact on the built environment business as profoundly and irreversibly as each of its three predecessors, and faster.

In the aviation industry, one example of IR4.0 adaptation is the development of predictive maintenance systems, which include the function of sensors and IoT in maintaining and monitoring aircraft components to be replaced before obvious flaws arise. This approach teaches a valuable lesson in how regular aircraft maintenance processes can be translated by the smart factory [8]. Similarly, in the manufacturing industry, components of smart factories are introduced as part of engineering science integrations. The distributed manufacturing branches or alliances are linked by leveraging technological improvements such as big data analytics and remote monitoring to assist factories in increasing productivity, creating a safe environment, and maintaining safe operations [8].

2.2 Aircraft Maintenance and Skill Sets Issues and Incidents

Aircraft maintenance refers to the repair and servicing actions conducted to maintain the aircraft operational and reliable. Aircraft maintenance is a critical responsibility for ensuring flight safety throughout the aircraft's life cycle [9]. Despite the fact that aircraft maintenance is considered a high-risk job assignment in aviation due to its critical impact on aviation safety, it nevertheless plays a substantial role in aircraft accidents and incidents [10].

Saleh et al. [11] did a study to explore maintenance safety issues. Human factors analysis in both studies reveals that the majority of the factors impacting the maintenance event are related to incorrect component inspection and installation methods, as well as transitory factors that come from the organizational and management levels.

According to Khan et al. [12] based on International Civil Aviation Organization (ICAO) official statistics, around 0.9 percent of aircraft incidents involved maintenance. It also said that maintenance accidents were 20% more likely to result in one or more deaths than all official ICAO mishaps (14.7%). Nonetheless, it was discovered that the number of accidents caused by maintenance errors each year dropped over the study period, and the statistic rate decreased from 5% per year to 2% per year. The findings revealed that aircraft with maintenance contributions were typically between the ages of 10 and 20 years, with aircraft over 18 years old being more likely to result in fleet loss and aircraft over 34 years old being more likely to result in death. [13] discovered that the primary causes of technical failures were inadequate repair processes, irresponsibility, and erroneous installations, which might be caused by worker skill sets.

Many of the previous publications explored key concerns concerning human factor errors in aircraft maintenance. This paper offers a study on document analytics on additional published publications about these topics. As a result, it is projected that this analysis will reveal human factors mistake, particularly skill mismatch among aircraft maintenance personnel, as a common talent management issue in the aviation business.

2.3 Document Mining

Document mining is one of the Data Mining (DM) approaches [14]. It seeks knowledge in a stack of documents that may contain comparable conversations and arguments by determining phrase frequencies and word connections in the texts. DM is well-known as a method of obtaining nontrivial, valuable, and meaningful insights from databases, also known as Knowledge Discovery in Databases (KDD) [15]. The KDD stages are as follows:

(i)   Selection – For the mining process, data set is carefully selected.
(ii) Preprocessing – The selected data is preprocessed into an acceptable data format and data type.
(iii) Transformation – The cleaned data is turned into appropriate data ranges and aggregations.
(iv) Data Mining – The primary process of discovering knowledge in terms of patterns and rules.
(v)   Interpretation – Patterns and rules are translated into knowledge and information for users, stakeholders, and decision makers.

A document is commonly regarded as a single piece of information, and a collection of them is referred to as a document dataset. A proper document collection is essential for both formal and informal readers to access material as part of knowledge and information sharing [16]. Many documents, such as articles, contracts, reports, and news, can be collected as data files produced or converted into digital papers thanks to ICT. Thus, document mining can be used to process digital documents for speedier information transmission. Document mining is mostly used to directly find term relationship and similarity in a document dataset [17].

In this study, document mining utilizes Text Mining (TM) which is one of the data mining approaches used to derive insights from text data. TM is a text analytics methodology that uses NLP and DM to convert free (unstructured) text in documents and databases into standardized and structured data patterns. During the preprocessing stage, NLP is used to clean, convert, and standardize text data (text terms) formatting.

In TM, cleaned terms are ranked using Term Frequency-Inverse Document Frequency (TF-IDF) to apply term count and weighting to each term that appears in a set of documents or corpus. As a result, it generates a list of keywords ranked by TF-IDF. Then, in a corpus, a Term Document Matrix (TDM) is built, which is a two–dimensional matrix or table with rows representing words and columns containing document ids. The goal of TDM generation is to record calculated word frequency counts across documents in a corpus [18].

TF-IDF is a popular method for computing word frequency counts and weights across documents in a corpus, and it is widely used in information retrieval and TM. It is intended to assess how important a term is to a document in a corpus. As a result, TF is the term frequency for terms that appear in a document, whereas IDF is used to determine how relevant a term is for a collection of documents or corpus [19]. Each term

has its own TF and IDF score, as well as its own TF*IDF weight, as indicated in Eq. 1:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

- $tf_{i,j}$ is total of frequencies of i in j.
- df_i is the number of documents have i.
- N is total number of documents in a corpus.

The $w_{i,j}$ value of a term grows in proportion to the number of terms that appear in the documents. It is used to determine the significance of a phrase across all papers in a corpus. Non-data science experts may find document mining to be technical, yet there are several useful text or document analytics products on the market, such as SAS, WordStat, Voyant Tools, RapidMiner, and many more. Some of these are beautifully designed, powerful, and ready to be used by non-technical users in constructing their own text or document analytics tasks. In this work, we employ a document analytics technique to evaluate a collection of internet articles about aircraft incidents, maintenance personnel issues and skill mismatch difficulties. As an absolute combination method, we combine NLP and TF-IDF.

3 Document Analytics on Aircraft Maintenance Issues
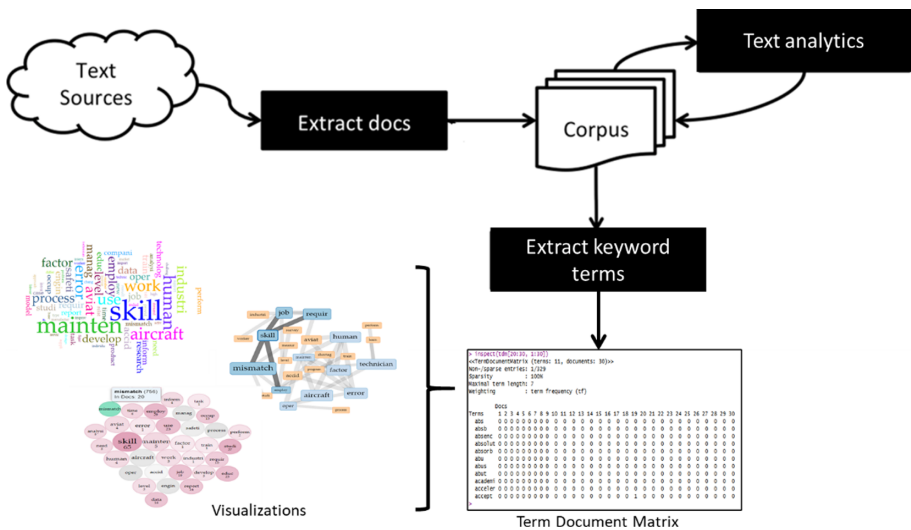
3.1 Document Mining Framework

Fig. 1. Document mining framework for aircraft maintenance documents.

This proposed document analytics process includes multiple integrated procedures in this study. Figure 1 depicts the framework, Document Mining Framework (DMF),

which consists of four (4) stages: (i) online articles and report collections on aircraft maintenance and incidents, and. (ii) preprocessing of document data (iii) text analytics (iv) visualizations. The goal of this experiment is to identify the key keywords that appear in the document set, also known as the corpus.

Initially, DMF gathers online papers that particularly explain and discuss aircraft maintenance faults and incidents, engineers' and technicians' competence and competency, and aircraft maintenance personnel training. These web articles and reports are gathered at random, but the extent of document content descriptions is carefully chosen. The documents are in text (.txt) file format. The files are then subjected to a document preprocessing mechanism. The R programming language was used to create the module. The module is built on the (NLP) principle, with two (2) key processes: (i) removing stop words and (ii) stemming terms to become root words. Document preprocessing is an important step in document analytics since it is a standard approach used to clean original documents to reduce noise, unstructured data, and inconsistent data.

The following module extracts relevant terms for use in creating the TDM. The ranking and extraction of these keywords is based on the TF-IDF methodology. Text analysis is carried out in accordance with their important values and word occurrences. The framework uses the TF-IDF algorithm to rank and count keywords at this level. A TDM is created throughout the mining process. The TDM is then used as the data input for term pattern visualization using the Voyant application [20], which is an open-source web-based program that reads and analyses texts in a variety of formats. It facilitates scholarly text analytics of texts or corpora, primarily by analysts, students, and the general public. Finally, the identified term patterns are represented in three (3) forms of visualization graphics, such as word clouds, term links and term berry. From here, we can identify the gist of document contents amongst various studies undertaken under the areas of aircraft maintenance and repair, technician skills, human errors, and aviation mishaps.

4   Results and Discussion

The experiment's results are reported and discussed in this subsection. The document analytics data are displayed in a variety of visualizations, including a word cloud, term links and termberry. Each visualization shows the key phrases that describe the gist of the document's contents.

The corpus includes 43 documents, including (i) 20 Web of Science indexed journals, (ii) 3 Scopus indexed journals, and (iii) 20 technical reports and white papers. Maintenance and Reliability (Web of Science), Human Factors: The Journal of the Human Factors and Ergonomics Society (Sage Journals), International Journal of Industrial Ergonomics (Scopus), Aerospace, and other indexed journals are among those chosen. These articles discussed and reviewed aircraft maintenance, repair, and overhaul, technician skills, human errors, and aviation incidents. This document dataset was published between 2010 and 2020.

All 43 documents were subjected to the document preprocessing module in the second stage of DMF to remove stop words and stemming. All papers are integrated as a corpus during document preprocessing. The text in the papers is then transformed to

lowercase, and all symbols and numerals are eliminated. The next step is to remove all numbers, symbols, stop words such as 'a', 'an', 'the', and white space from the corpus. The goal of this method is to improve the quality of the corpus by removing nonsensical phrases. Finally, stemming is a critical procedure. Stemming is a process in which terms are trimmed down to their underlying term. For example, the terms 'running' and 'runner' are trimmed down to their base word 'run'. The stemming procedure is essential to ensure that terms with the same root words are referred to as the same separate term.

Table 1. Part of the term document matrix.

	file1	file2	file3	file4	file5	file6
Accept	3	4	4	5	3	4
Access	9	1	25	1	23	0
Accord	1	34	8	1	2	13
Activ	1	6	10	4	6	12

The third stage of DAF then employs the TF-IDF equation on the cleaned corpus to generate TDM. Table 1 depicts a portion of TDM's rows and columns used to store documents and phrase frequencies. These frequency counts are used to determine the links between terms and documents.

The summary sheet, as shown in Fig. 2, contains information corpus and term details such as document length, vocabulary density, average words, most frequent words, and distinctive words. This corpus contains 43 documents with a total of 278,807 words and 13,045 unique terms. The longest and shortest documents in the corpus are described by document length. The top five longest documents are file7 (32757 terms), file27 (30792 words), file29 (21910 words), file14 (21425 terms), and file3 (12706 terms). The shortest documents, on the other side, are file24 with 353 terms, followed by file26 with 581 terms, file21 with 589 phrases, file16 with 658 terms, and file20 with 794 terms.

It also reveals that the most often occurring terms in the corpus are 'skill', 'mainten', 'aircraft', 'error', and 'human'. Other visualization statistics use these keywords to demonstrate word relevance and relationship in the document contents.

Figure 3 depicts a word cloud, which is a graphical representation of the frequency of words in the corpus. The larger the words appear in the word cloud, the more frequently the terms are used. The word cloud indicates larger words for 'skill' (4350 frequencies), which indicates it appears 4350 times in relation to the 43 long datasets, followed by 'mainten' (2719 frequencies), 'aircraft' (1968 frequencies), 'error' (1772 frequencies), and 'human' (1772 frequencies) (1673 frequencies). The terms in the word cloud have a high phrase count among the 13,045 unique terms. Terms such as 'employ' (1191 frequencies), 'technolog' (939 frequencies),'mismatch' (756 frequencies), 'industry' (1494 frequencies), 'educ' (942 frequencies),'safeti' (1167 frequencies), 'aviat' (1436 frequencies), 'factor' (1359), 'train' (1058 frequencies), and 'accid' (1046 frequencies). We may determine from the word cloud visual that the corpus contains many document

This corpus has 43 documents with 278,807 total words and 13,045 unique word forms. Created 2 seconds ago.

Document Length:
- Longest: file7 (32757); file27 (30792); file29 (21910); file14 (21425); file3 (12706)
- Shortest: file24 (353); file26 (581); file21 (589); file16 (658); file20 (794)

Vocabulary Density:
- Highest: file24 (0.688); file25 (0.529); file20 (0.526); file26 (0.504); file21 (0.499)
- Lowest: file27 (0.073); file29 (0.093); file14 (0.104); file7 (0.113); file3 (0.151)

Average Words Per Sentence:
- Highest: file7 (32757.0); file27 (30792.0); file29 (21910.0); file14 (21425.0); file3 (12706.0)
- Lowest: file24 (353.0); file26 (581.0); file21 (589.0); file16 (658.0); file20 (794.0)

Most **frequent words** in the corpus: skill (4350); mainten (2719); aircraft (1988); error (1772); human (1673)

Fig. 2. Summary of information corpus and term.

contents about aircraft maintenance, human factors, accidents and errors, skill mismatch, the aviation industry, and education and training.

Fig. 3. A word cloud.

Figure 4 depicts the link analysis of the frequent phrases found in the 43 documents in this text analytics. Term links associate terms with how they are written or described in texts. As a result, the created term linkages provide support for details pertaining to term representation in a word cloud. The term links display the descriptive terms associated with the primary keywords. For example, the term 'skill' associated to 'mismatch', 'technician', and 'requir' described that many of the documents described the technician having mismatch skills required for their aircraft maintenance job responsibilities. The thicker the ties between two (2) keywords, the greater their association.
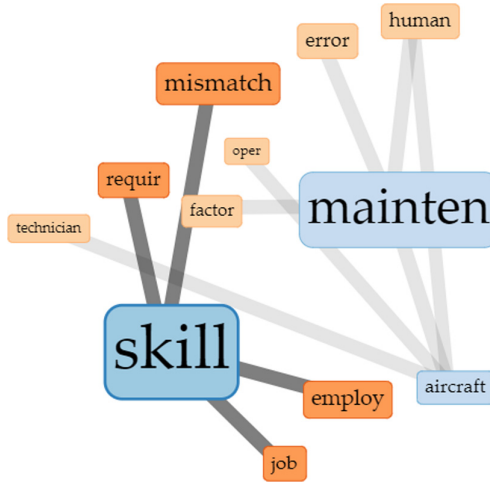
Fig. 4. Term links.

We may use the term link to zoom in and see what other terms are related to the core keywords. Figure 5 depicts the sub term connection, which reveals a connection analysis of more related keywords associated with the 43 documents' common keywords. In this graphic, we will get a larger picture of how the issues in these publications are comparable. Many publications addressed skill mismatch and inadequacy, as well as aircraft maintenance. The words 'technician' are synonymous with 'learn', 'program', and 'aircraft'. The connections between 'human', 'error', 'mainten' and 'accid' emphasize one of the critical challenges that require more attention from the allied parties. Overall, the link analysis correlates the primary keywords and terms such as 'skill', 'aircraft', 'technician', 'error', 'human', 'mismatch', 'measur' and 'train'.

A TermsBerry representation is depicted in Fig. 6. The TermsBerry aims to integrate high frequency terms with co-occurring phrases. In this way, it is an extension of displaying how close the terms appear to one another. It is also used in the same way as word cloud visualization, but it is more effective because it includes term statistics and corpus coverage information. In this figure, we highlight the term 'mismatch' which has 756 frequencies and appears in 20 publications and is strongly connected to keywords such as 'skill', 'educ', 'employ', 'use' and 'studi' (high statistics).

Overall, we can deduce from the visualizations that the document dataset focuses and discusses on aircraft maintenance incidents and skill set issues. Because aircraft maintenance is seen as a complicated problem to solve [21], it is critical to address issues of talent inadequacy and mismatch among aircraft maintenance specialists.
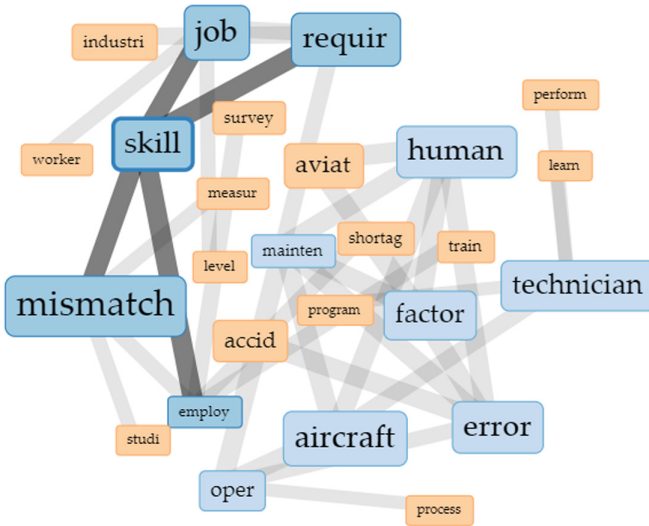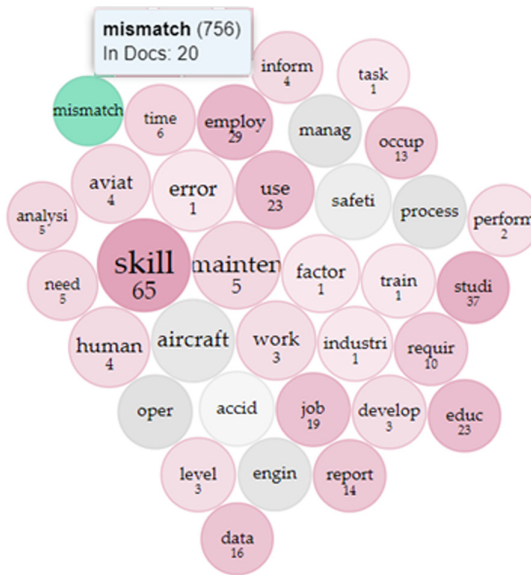
Fig. 5. Sub-term links.

Fig. 6. Visualization of TermBerry

5   Conclusion

We use a document mining technique to analyze a collection of web articles and reports on aircraft incidents and aviation competence problems in this study. The impact of IR4.0 on aviation, aircraft maintenance and competence challenges and incidents, document

mining, TDM, and the TF-IDF technique are also discussed in this paper. It introduces the DMF and shows the results of document analytics performed with the framework in question. We can deduce that the document collection under consideration focuses mostly on aircraft maintenance and skill set problems. According to the publications reviewed, aircraft maintenance is considered a challenging topic to solve, and it is critical to address skill gaps and mismatch among aircraft maintenance professionals. The importance of aircraft maintenance, repair, and safety were emphasized in these contents. According to reports and news, many aviation disasters occur due to a variety of factors, including malfunctioning equipment, hazardous weather conditions, such as turbulence, human mistake, and worker technical skill. To address IR 4.0 and skill mismatch concerns in aircraft maintenance, the research will be developed in the future to address the need for a competency-based skill evaluation and training mapping model for Royal Malaysian Air Force technicians.

References

1. Verhaest, D., Omey, E.: The relationship between formal education and skill acquisition in young workers' first jobs. Manch. Sch. **81**, 638–659 (2014). https://doi.org/10.1111/j.1467-9957.2012.02305.x
2. Mohd Kamaruzaman, F., Hamid, R., Mutalib, A.A., Rasul, M.: Comparison of engineering skills with IR 4.0 skills. Int. J. Online Biomed. Eng. (iJOE). **15,** 15 (2019). https://doi.org/10.3991/ijoe.v15i10.10879
3. Xu, M., David, J., Kim, S.: The fourth industrial revolution: opportunities and challenges. Int. J. Finan. Res. **9**, 90 (2018). https://doi.org/10.5430/ijfr.v9n2p90
4. Samah, S.K.: Advancing Malaysian Aerospace Industry Through Research. Malaysian Industry-Government, Kuala Lumpur (2015)
5. Kumar, P., Kamalakannan, P.: Skill gap analysis and training needs in Indian aerospace industry. J. Airl. Airpor. Manag. **6**, 115–132 (2016). https://doi.org/10.3926/jairm.56
6. Mrugalska, B., Wyrwicka, M.: Towards lean production in industry 4.0. Procedia Eng. **182**, 466–473 (2017) https://doi.org/10.1016/j.proeng.2017.03.135
7. Maskuriy, R., Selamat, A., Maresova, P., Krejcar, O., David, O.: Industry 4.0 for the construction industry: review of management perspective. Economies **7**, 14 (2019) https://doi.org/10.3390/economies7030068
8. Tat, M., Kushan, M.: Impact of Industry 4.0 to aircraft maintenance, repair and overhaul. In: Conference: VII International Symposium Engineering Management and Competitiveness 2017 (2017)
9. PCAA: Approved Maintenance Organisations-Air Navigation Order, 3rd ed Pakistan Civil Aviation Authority, Karachi, Pakistan (2019)
10. Rajee, O.D., Miller, M., Mrusek, B.M.: Managing safety risks in airline maintenance outsourcing. Int. J. Aviat. Aeronaut. Aerosp. **7**, 7 (2020)
11. Saleh, J.H., Tikayat, R.A., Zhang, K.S., Churchwell, J.S.: Maintenance and inspection as risk factors in helicopter accidents: Analysis and recommendations. PLoS ONE **14**, e0211424 (2019)
12. Khan, F.N., et al.: A preliminary investigation of maintenance contributions to commercial air transport accidents. Aerospace **7**(9), 129 (2020). https://doi.org/10.3390/aerospace7090129
13. Habib, K., Turkoglu, C.: Analysis of aircraft maintenance related accidents and serious incidents in Nigeria. Aerospace **7**(12), 178 (2020). https://doi.org/10.3390/aerospace7120178

14. ECM Connection: Data mining and document mining. https://www.ecmconnection.com/doc/data-mining-and-document-mining-0001. Access Online on 21 Jan 2021
15. Fayyad, P., Shapiro, S.: From data mining to knowledge discovery: an overview. In: Fayyad, P.-S., Smyth, U. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 1–34. AAAI Press/The MIT Press, Menlo Park, CA (1996)
16. O'Neil, B.: Framework for managing knowledge, content and documents. https://tdan.com/framework-for-managing-knowledge-content-and-documents/21065 (2017). Last Accessed 21 Jan 2021
17. Nohuddin, P.N.E., Noormanshah, W.M.U., Zainol, Z.: Content analytics based on random forest classification technique: an empirical evaluation using online news dataset. Int. J. Adv. Appl. Sci. **8**(2), 77–84 (2021). https://doi.org/10.21833/ijaas.2021.02.011
18. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF. J. Doc. **60**, 503–520 (2004). https://doi.org/10.1108/00220410410560582
19. Qaiser, S., Ali, R.: Text mining: use of TF-IDF to examine the relevance of words to documents. Int. J. Comput. App. **181**(1), 25–29 (2018). https://doi.org/10.5120/ijca2018917395
20. https://voyant-tools.org/ (2018)
21. Ward, M., McDonald, N., Morrison, R., Gaynor, D., Nugent, T.: A performance improvement case study in aircraft maintenance and its implications for hazard identification. Ergonomics **53**(2), 247–267 (2010). https://doi.org/10.1080/00140130903194138