Langit Biru: Jurnal Ilmiah Aviasi Vol. 18 No. 3 October 2025 Page. 118 : 129
ISSN (p) 1979-1534 ISSN (e) 2745-8695

IDENTIFICATION OF SUSTAINABILITY COURSE TOPICS RELEVANT TO THE NEEDS OF AIRPORT PROFESSIONALS USING NLP

Muhammad Farras Haidar[1], Toto Indriyanto[2]
[1,2] Institut Teknologi Bandung
e-mail: [1]farrashaidar09@gmail.com, [2]t.indriyanto@itb.ac.id
corresponding: [1]farrashaidar09@gmail.com

Received:	Revised:	Accepted:
28 May 2025	18 July 2025	3 October 2025

Abstract: This research is based on the condition of environmental degradation and public awareness of climate change caused by industry after the Paris Agreement held in 2015. One industry that causes environmental damage is the aviation industry. Increased awareness of the impact generated has made the aviation industry begin to improve to reduce emissions. This research focuses on identifying sustainability course topics relevant to airport professionals to improve their competency in dealing with environmental issues. This research uses Natural Language Processing (NLP) methods, where the NLP used is Unsupervised Learning, especially Latent Dirichlet Allocation (LDA). LDA was used to search for hidden topics in a set of airport sustainability reports accredited by Airport Carbon Accreditation (ACA). ACA is a globally recognized institutional carbon management certification program for airports. The analysis results show that the professional training topics related to green finance for sustainable airports match the current market needs and are not yet organized by course providers at leading international aviation organizations.

Keywords: Sustainability Course, Aviation Industry, NLP, LDA, Professional Training

Introduction

Global warming, driven by environmental degradation, has resulted in rising global temperatures that are increasingly experienced worldwide. The impacts of rising temperatures are becoming more evident, including an increase in the frequency and intensity of natural disasters. These effects make people begin to realize the need for actions that can protect the planet (Agustina and Pradesa, 2024). One of the regions significantly affected by global warming is Europe. The Copernicus Climate Change Service noted that Europe experienced the hottest summer ever recorded. Especially in the summer, from June to August, it was recorded that the temperature increased 0.7 degrees Celsius above the average temperature in 1991 - 2020 (Copernicus Climate Change Service, 2024).

The Paris Agreement, adopted in 2015, established a global framework to address climate change. The agreement aims to limit the increase in global temperature to below 2 degrees Celsius above pre-industrial levels. It seeks to limit the temperature increase to 1.5 degrees Celsius above pre-industrial levels (COP21, 2015). All industries must reduce their emissions, including the aviation industry. The roles of each stakeholder in reducing emissions are expected to have a significant impact on reducing emissions from aircraft manufacturing to airports are required to implement sustainability practices.

Sustainability at airports has progressed significantly from the initial initiatives that focused solely on emissions reduction and waste management to a more comprehensive approach that includes social and economic aspects of operations. Supported by environmental policies and the adoption of renewable energy technologies, airports around the world continue to become more environmentally friendly and sustainable. Airport operators are facing increasing pressure from

DOI: https://doi.org/10.54147/langitbiru.v18i03

stakeholders, such as governments, shareholders, customers, employees, and the general public, to assess their socio-economic impacts and manage their operations sustainably (Dimitriou and Karagkouni, 2022). In response to the increasing awareness of sustainability and environmental issues, there is an urgent need to develop initiatives that support sustainability practices in airports around the world.

A collaborative approach involving a wide range of stakeholders is needed to overcome the regulatory complexities of implementing green technology in the aviation industry. According to the World Economic Forum, success in implementing sustainability depends on supportive policies, innovative partnerships, financing, and appropriate education and information approaches. Requiring a wide range of initiatives to achieve ambitious CO2 emission reduction goals, industry players and the government are expected to collaborate to ensure implementation in the aviation industry. Such collaboration can include research and development of new technologies to reduce emissions (World Economic Forum, 2011).

In recent years, the aviation sector has increasingly adopted Natural Language Processing (NLP) methods, particularly in the field of safety management. As reviewed by Yang and Huang (2023), NLP has been widely used to process large collections of narrative safety reports and accident investigation documents, enabling the extraction of risk factors and identification of safety trends. Their review shows that techniques such as text classification, topic modeling, and information extraction have supported safety monitoring and management in aviation. NLP's effectiveness in identifying hidden patterns in aviation safety data provides a strong basis for applying this method in this study. Therefore, NLP was chosen in this study to help analyze hidden patterns in airport sustainability reports to find topics that could be raised in a course for professionals working at airports.

As such, this research aims to identify sustainability topics that meet the needs of aviation professionals using market analysis. This research will provide important insights into the development of training programs that can help the aviation industry to achieve its sustainability goals while supporting global efforts to address climate change.

Method

This research used a machine learning approach based on Latent Dirichlet Allocation (LDA), a topic modelling using unsupervised learning. The methodology used in this research is as follows:

1. Natural Language Processing (NLP)

Computing machines are only able to understand and execute instructions in the form of binary codes, i.e., 0 and 1, while humans use everyday language to communicate. Therefore, a mechanism is needed that can convert human language into machine-understandable language. Natural Language Processing (NLP) is a branch of computer science that focuses on developing techniques to analyze, formulate, and interpret human language. NLP is divided into two main categories: speech recognition and text processing (Bastian, 2023). Text processing can be applied to analyze narrative data on a large scale, such as narrative data in sustainability reports, so that analysis can be carried out more efficiently both in terms of time and cost. NLP is able to analyze many reports, one of which is analyzing company sustainability reports. It can be used to extract meaningful insights and patterns from large and complex text data (Kang & Kim, 2022).

Regular Expression is one of the main tools to describe patterns in text. Regular Expression can identify character sets that can be extracted from a document (Vestermark, 2024). Regular expressions have an important role in the text preprocessing stage. Text preprocessing serves to convert text into a form that is more easily interpreted by computers. Various methods of text

normalization will be further explained below:

1.1 Tokenization

Tokenization is a standard method for breaking text into separate words. This process is done to facilitate further analysis. However, not all words are separated by spaces, as in the examples of "He's" or "Don't". In cases like this, the tokenization method must include various text splitting techniques to separate the words appropriately. Breaking a text into individual words or tokens allows the system to understand the language better (Kang and Kim, 2022).

1.2 Lowercasing

Regular Expression is sensitive to uppercase and lowercase differences, so each capital letter will be considered different from its lowercase counterpart. Therefore, it is important to standardize all tokens into lowercase or uppercase letters, so that NLP algorithms can work more effectively. However, in text processing, uniformity to uppercase is rarely used. Instead, it is more common as it is supported by libraries such as NLTK, spaCy, and others (Rose et al., 2020).

1.3 Stop-words Removal

Before representing a document, it is important to remove words that appear frequently, because common words that have a high frequency tend to have low semantic value (Mayola et al., 2024). These words are known as stop-words, such as and, or, but, and so on.

1.4 Lemmatization

Lemmatization aims to identify that some words have the same root, even though their surface forms are different. For example, the words fall, fell, and fallen all come from the same root form, fall. Lemmatization is essential in processing languages that have complex word forms and structures (Lopez Perez et al., 2025).

2. Unsupervised Learning

Unsupervised Learning is a type of machine learning that uses algorithms to analyze and classify datasets without labels. These algorithms can detect hidden patterns or group data independently without human intervention. Unsupervised Learning models are used in three main tasks: clustering, association rules, and dimensionality reduction.

The main goal of Unsupervised Learning is to discover patterns, structures, and relationships in data without the use of explicit labels or predefined target variables. In contrast to Supervised Learning, which aims to enable machines to learn from labeled examples and make predictions, Unsupervised Learning focuses more on extracting insights and knowledge from unlabeled data.

3. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probability model for discrete data sets such as text collections. LDA models are a three-level Bayesian hierarchy where each item in the data set is considered as a mixture of several underlying topics. Each data element is assessed as a result of a descent process involving several related topics. The purpose of using LDA is to identify the optimal number of topics in a text corpus and to determine the distribution of words in each topic (Mustahidah, 2021). To make it easier to understand the algorithm of LDA, an illustration of topic modeling using LDA can be seen in Figure 1 (Rishu & Kukreja, 2024):

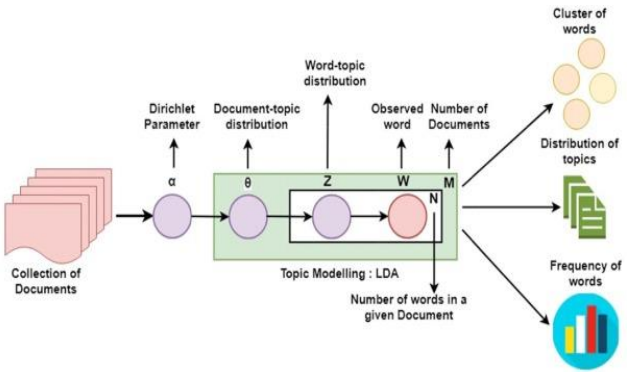

Figure 1: Model of LDA

In Figure 1, the generative model is illustrated through three different levels. The first level is corpus-level parameters, the second level is document-level variables, and the third level is word-level variables. In text analysis, topics can provide a clear picture of the content of a document. The parameter values α and β are Dirichlet distributions, the joint distribution of the mixture of topics θ, the set of topics z of N, and the set of words w of N can be formulated as follows (Lopez Perez et al., 2025.):

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (1)$$

where:

α = Dirichlet parameters on the distribution of topics over documents.
β = Dirichlet parameter of the distribution of words over topics.
θ = The distribution of topics in the document, selected from the distribution. Dirichlet Dir(α).
z_n = Topic for the n-th word, chosen from the multinomial distribution based on θ.
w_n = The n-th word, chosen from a multinomial distribution based on the topic z_n and parameter β.

The distribution θ is chosen as the Dirichlet distribution (α) and for each word w_n, the topic z_n is the multinomial probability of θ, and the word Wn of $p(w_n|z_n, \beta)$ is the multinomial probability that depends on the topic z_n.

4. Workflow of The NLP – LDA

The workflow of the NLP – LDA process is illustrated in Figure 2. The process begins with the collection of sustainability reports from airports accredited at ACA Level 4+ and Level 5, which represent the highest level of maturity in carbon management. These sustainability reports are then processed through a series of text preprocessing steps to prepare the data for topic modeling.

In the Data Preprocessing stage, the text data undergoes tokenization to separate the text into individual words or tokens. The tokens are then converted to lowercase to ensure consistency, followed by the removal of stop words and irrelevant characters that do not contribute to semantic meaning. Lemmatization is applied to reduce words to their base or dictionary form, to improve linguistic accuracy, while stemming is performed to normalize word variations. These preprocessing steps are essential for reducing noise and improving the quality of input data for the modeling stage.

After preprocessing, the data is analyzed using Latent Dirichlet Allocation (LDA). At this stage, several topic candidates are generated by running the LDA model with different numbers of

topics (K). Each candidate is evaluated using coherence and perplexity measures, which assess both the semantic interpretability and statistical quality of the topics. The model with the intersection score between coherence and perplexity is selected as the optimal representation. Finally, the resulting topics were subjected to visualization, which provided a clear depiction of the most representative keywords within each topic, enabling the labeling and interpretation of the results.

This workflow ensured that the research process was transparent, systematic, and reproducible, allowing the identification of sustainability-related themes that are most relevant for aviation professionals.

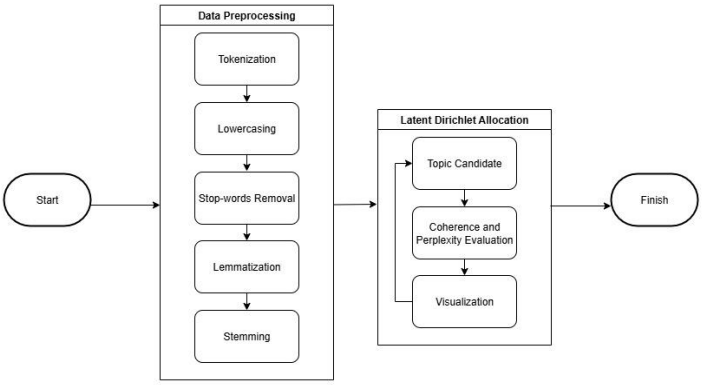

Figure 2: Workflow of NLP – LDA Process

5. Gap Analysis

Gap analysis is a method used to compare what is expected/desired with the actual existing. This analysis measures how optimal business processes implemented by the company. In conducting gap analysis, three main components need to be considered: actual performance, expected performance, and the gap between the actual and expected performance.

Results and Analysis

1. Market Size Determination

Market size is an important concept in business analysis that is used to measure the potential revenue or sales in a particular market. Market size gives an idea of how much opportunity there is in that market and helps companies make strategic decisions. There are three main components in market size analysis as depicted in Figure 3 below: Total Addressable Market (TAM), Serviceable Available Market (SAM), and Serviceable Obtainable Market (SOM) (Davalas, 2023).

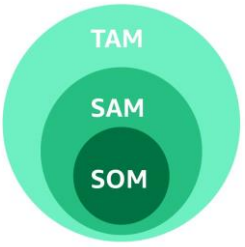

Figure 3. Market Size

1.1 Total Addressable Market (TAM)

The Total Addressable Market (TAM) gives an idea of the maximum addressable market potential that a course provider could achieve if they sold their product or service to all existing customers in the entire market. In the context of this research, TAM is calculated

based on the number of airports that are members of Airport Council International (ACI) World, which is 2,110 airports worldwide (Airport Council International, 2024). These airports are divided into 5 regions that adjust the branch of ACI, namely Europe, Asia Pacific – Middle East, North America, South America – Caribbean, and Africa.

1.2 Serviceable Available Market (SAM)

The Serviceable Available Market (SAM) is the part of TAM that makes it possible to reach course providers with the products offered. SAM considers factors such as market segmentation that are relevant to the product offered. In this study, SAM is identified based on the number of airports that have been accredited by ACA. This is because these airports have a strong commitment to sustainability practices and have become an example for other airports to be able to reduce carbon emissions. Thus, accredited airports need sustainability-related training for their employees.

ACA-accredited airport-focused course providers can develop more relevant and effective programs to cater to this market. According to the ACA's annual report, the number of members accredited by the ACA is increasing every year. By 2024, ACA members will amount to 575 airports, so the SAM for this study is 575 airports.

1.3 Serviceable Obtainable Market (SOM)

Serviceable Obtainable Market (SOM) represents the part of the Serviceable Available Market (SAM) that a company can realistically obtain within a certain period. In this study, SOM is the number of airports accredited by ACA that have level 4+. This is because they have shown a very strong commitment to sustainability and carbon emission reduction. Another point is that using the total number of ACA accredited airports with level 4+ as SOM helps course providers to identify the most relevant and potential market segmentation. Airports that have this level have specific needs related to carbon emission reduction and sustainability, which are expected to match the sustainability courses offered. The number of airports that have been accredited 4+ and above by ACA in 2024 is 68 airports, so the SOM in this study is 68 airports.

2. Market Opportunities

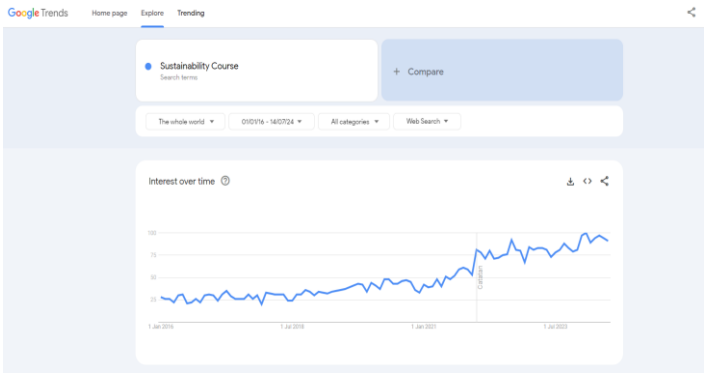

Figure 4: Results from Google Trends on Sustainability Course Search (Google Trends, 2024)

Figure 4 shows a significant increase in interest in sustainability courses globally. Data from Google Trends shows a consistent increase in searches related to "Sustainability Course" from 2018 to 2024. This graph reflects the increasing awareness and interest of society and industry in

sustainability and the importance of developing skills in this field. This surge in interest is driven by an increased global focus on environmental issues, sustainability policies, and green initiatives that are increasingly becoming priorities across various sectors.

This rise in interest provides a great opportunity for education and training providers to offer relevant and up-to-date courses related to sustainability. With more individuals and organizations looking for ways to reduce their environmental impact, sustainability courses are becoming increasingly important in helping to achieve global sustainability goals. Course providers can capitalize on this trend by developing programs that are comprehensive and tailored to market needs.

3. Airport Data Used

In this study, researchers focused on identifying sustainability course topics that are relevant to the needs of aviation professionals, especially airports. Data was obtained from existing sources, namely from airport sustainability reports that have been accredited by ACA with level 4+ and level 5 in 2022 – 2023. Sustainability reports at this level are used because they represent the highest commitment to carbon and sustainability management at airports. Currently, the highest level in ACA is level 5. Since not many airports have been accredited at level 5, level 4+ was added so that sustainability reports can illustrate how airports with a very high commitment to sustainability are working to reduce their carbon footprint to achieve net zero emissions as stipulated in the Paris Agreement. Reports from ACA level 1 until 4 were not included in this study because they represent lower stages of carbon management maturity, where commitments and actions are less comprehensive compared to level 4+ and 5 airports. In addition, only sustainability reports written in English were selected for this study to ensure consistency in text analysis and avoid language-related bias. The sustainability reports also describe the areas of concern that are the focus of their efforts to reduce their carbon footprint. The sustainability reports used in this study came from 38 airports. These reports provide information on sustainability practices, carbon emissions, and environmental management strategies implemented by airports to meet net zero emission targets. Table 1 lists the airports that have an accreditation level of 4+ or above in the last edition collected and used in this study:

Table 1: ACA Accredited Airports Level 4+ and 5 (*Airport Carbon Accreditation*, 2024)

North America	Europe	Asia Pacific Middle East
• San Diego (SAN) • Vancouver (YVR)	• Schipol (AMS) • Eindhoven (EIN) • Rotterdam The Hague (RTM) • London - Heathrow (LHR) • Rome - Fiumicino (FCO) • Rome - Ciampino (CIA) • Beja (BYJ) • Horta (HOR) • Faro (FAO) • Flores (FLW) • Santa Maria (SMA) • Lisbon (LIS) • Madeira (FNC) • Porto (OPO) • Porto Santo (PXO) • Kittilä (KTT) • Ivalo (IVL) • Kuusamo (KAO) • Rovaniemi (RVN)	• Kempegowda (BLR) • Christchurch (CHC) • Indira Gandhi (DEL)

North America	Europe	Asia Pacific Middle East
	• Helsinki (HEL) • London - City (LCY) • Milan - Linate (LIN) • Milan - Malpensa (MXP) • Stockholm - Arlanda (ARN) • Göteborg Landvetter (GOT) • Malmö (MMX) • Kiruna (KRN) • Åre Östersund (OSD) • Visby (VBY) • Budapest (BUD) • Luxembourg (LUX) • Pafos (PFO) • Larnaka (LCA)	

Discussion

1. Data Preprocessing

In this study, data preprocessing contained several steps. The original data were in '.pdf' format and were processed using the Python programming language. The narrative text was preprocessed with NLP libraries in Python, specifically NLTK and spaCy, which were used to perform tokenization, lowercasing, punctuation removal, stop-words removal, and lemmatization. Lemmatization was selected over stemming because it provides higher linguistic accuracy (Bastian, 2023). The preprocessing steps are summarized in Table 2. It should also be noted that potential confounding factors may affect the preprocessing stage. In particular, variations in report quality or level of detail may influence the accuracy of topic extraction. Although preprocessing was designed to minimize these effects, they were not systematically controlled and should be acknowledged as a limitation of this research.

Table 2: *Data Preprocessing* Results

Preproccesing	Results
Tokenization	['Schiphol', 'must', 'be', ',' , 'quieter', 'cleaner', 'and', 'better', '.', 'Firstly', 'because', 'we', 'simply', 'believe', 'that', 'this', 'is', 'necessary', '.', 'Schiphol', 'serves', 'a', 'societal', 'interest', 'which', 'includes', 'taking', 'good', 'care', 'of', 'the', 'neighbors', 'and', 'the', 'environment','.', 'We', 'are', 'very', 'transparent', 'on', 'our', 'position', 'due', 'to', 'the', 'broad', 'public', 'interest', '.', ...]
Lower-casing	['schiphol', 'must', 'be', ',' , 'quieter', 'cleaner', 'and', 'better', '.', ',', 'because', 'we', 'simply', 'believe', 'that', 'this', 'is', 'necessary', '.', 'schiphol', 'serves', 'societal', 'interest', 'which', 'includes', 'taking', 'good', 'care', 'of', 'the', 'neighbors', 'and', 'the', 'environment', '.', 'we', 'are', 'very', 'transparent', 'on', 'our', 'position', 'due', 'to', 'the', 'broad', 'public', 'interest', '.', ...]

Punctuation Removal	['schiphol', 'must', 'be', 'quieter', 'cleaner', 'and', 'better', 'because', 'we', 'simply', 'believe', 'that', 'this', 'is', 'necessary', 'schiphol', 'serves', 'societal', 'interest', 'which', 'includes', 'taking', 'good', 'care', 'of', 'the', 'neighbors', 'and', 'the', 'environment', 'we', 'are', 'very', 'transparent', 'on', 'our', 'position', 'due', 'to', 'the', 'broad', 'public', 'interest', ...]
Stop-words Removal	['quieter', 'cleaner', 'better', 'simply', 'believe', 'necessary', 'serves', 'societal', 'interest', 'includes', 'taking', 'good', 'care', 'neighbors', 'environment', 'transparent', 'position', 'broad', 'public', 'interest', ...]
Lemmatization	['quieter', 'cleaner', 'well', 'simply', 'believe', 'necessary', 'serve', 'society', 'interest', 'include', 'take', 'good', 'care', 'neighbor', 'environment', 'transparent', 'position', 'broad', 'public', 'interest', ...]

2. Topic Modeling

In topic modeling, the main programming language used is Python due to its versatility and wide functionality. After preprocessing, the corpus undergoes vectorization because computers cannot process words in their raw format. The text was first converted into a Bag-of-Words (BoW) matrix, and then re-weighted using Term Frequency–Inverse Document Frequency (TF-IDF) implemented in Python, which highlights words that are more important relative to others in the corpus (Mee, Homapour, Chiclana, and Engel, 2021).

The next step was to execute the Latent Dirichlet Allocation (LDA) model, which requires the specification of the number of topics (K) to be extracted. Since the value of K was not known previously, several candidate models with different K values were generated. K values in the range of 1–20 were tested, and each model was trained repeatedly to account for random initialization.

The quality of each candidate model was evaluated using two complementary metrics: Coherence measures the semantic similarity of high-probability words within a topic, while perplexity reflects statistical fit. The final choice of K was determined by identifying the intersection between coherence and perplexity, ensuring both semantic interpretability and statistical robustness (Tresnasari, Adji, and Permanasari, 2020).

The random state in the LDA model was fixed at 20 to ensure reproducibility. Coherence and perplexity were calculated for each value of K, and the results (Figure 5) indicated that the optimal number of topics was K = 2.

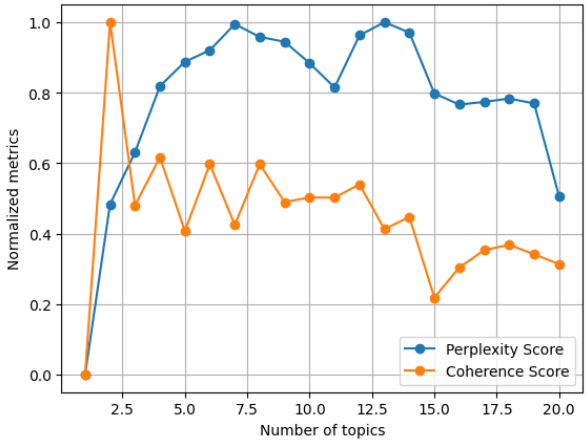

Figure 5: Coherence Score vs Perplexity Score

The LDA model generates topics located in the corpus, where each topic is represented by the most significant words in the topic. The top words for each topic and the topic proportion of the corpus are shown in Table 3.

Table 3: Top Keywords

Number of Topics	Top Keywords
1	*Emissions, Aviation, Operations, Risk, Management, Reduce, Energy, Project, Development, Aircraft.*
2	*Risk, Relate, Financial, Value, Asset, Change, Cost, Rate, Loss, Fair.*

From the 10 keywords in each topic, conclusions can be drawn to form the theme of each topic. The 1st topic contains the words *'emissions', 'aviation', 'operations', 'risk', 'management', 'reduce', 'energy', 'project', 'development', 'aircraft'*. From these *keywords*, it can be concluded that the 1st topic is about *Emission Reduction Strategies for Airports.* The topic of Emission Reduction Strategies for Airports is very relevant because it is supported by previous research showing that airports often face challenges due to inefficient and outdated energy system infrastructure. Midžić Kurtagić and Kadrić (2022) highlighted that improvements such as upgrading lighting systems, increasing the efficiency of pumps and HVAC (Heating, Ventilation, Air Conditioning) equipment, and integrating renewable energy sources can result in significant reductions in energy consumption and carbon emissions. These findings indicate that systematic energy efficiency initiatives are an important factor for airports to reduce their environmental impact, thereby reinforcing the relevance of the topic identified through the topic modeling results in this study.

In the 2nd topic, the top 10 *keywords* are *'risk', 'relate', 'financial', 'value', 'asset', 'change', 'cost', 'rate', 'loss', and 'fair'*, so the title of the 2nd topic is *Green Finance for Sustainable Airports.* The topic of Green Finance for Sustainable Airports is potentially very relevant to explore, as a study conducted by Liu and Zhu (2024) shows that green finance improves carbon emission efficiency by encouraging companies to achieve greater economic output with less energy use. Green finance also contributes to reducing emission intensity, promoting innovation, and improving efficiency. These findings indicate that green finance is a critical enabler in the low-carbon transition, reinforcing the findings to add the topic of green finance for sustainable airports to airport training programs. Without adequate knowledge and skills in this area, aviation professionals may find it difficult to utilize financial instruments that are increasingly recognized as essential tools for achieving sustainability goals.

Table 4 illustrates the process of topic generation based on emerging keywords. The determination of the course topics was based solely on the top keywords that emerged from the airport sustainability reports. The course topics represent the overall *top keywords* in each topic.

Table 4: Course Topics

Topic.	Course Topics
1	*Emission Reduction Strategies for Airports*.
2	*Green Finance for Sustainable Airports*.

3. Gap Analysis

In the gap analysis, an analysis was conducted to compare what was produced by LDA, namely topics that match the current market demand related to sustainability courses, with the currently available courses offered by leading organizations influential in the aviation industry. Table 5 illustrates the gap analysis between the LDA results and the available courses.

Table 5: *Gap Analysis*

No.	Training Provider	Topic 1	Topic 2
		Emission Reduction Strategies for Airports	Green Finance for Sustainable Airports
1	ACI	√	-
2	IATA	-	-
3	ICAO	√	-
4	UK CAA	√	-

In Table 5, it is known that the 1st topic is a course themed on "*Emission Reduction Strategies for Airports*" where there are 3 out of 4 course providers who have this course theme. The three courses that already have the topic are Airport Council International (ACI), International Civil Aviation Organization (ICAO), and United Kingdom Civil Aviation Authority (UK CAA). Only the International Air Transport Association (IATA) does not have a course theme related to *Emission Reduction Strategies for Airports*. As for topic 2, which is a course on "*Green Finance for Sustainable Airports*", none of the providers has a course theme.

4. Limitations

This study has several limitations that should be acknowledged. First, the dataset consisted only of 38 airports that have ACA accredited reports (level 4+ and 5), which may limit generalizability to airports at earlier stages of sustainability implementation. Second, only English language reports were included, which may introduce language bias and overlook perspectives from non-English reports. Third, variations in report quality and structure were not systematically controlled, potentially affecting topic extraction. Finally, the study applied only LDA for topic modeling. For future research could compare results with alternative approaches to enhance robustness.

Conclusion

Based on the results of data processing and analysis that have been discussed and explained in this study, it can be concluded as follows the market size of the sustainability course is 2110 airports for TAM, 575 airports for SAM, and 68 airports for SOM, The implementation uses NLP methods by preprocessing text, tokenization, lowercasing, stop-words removal, and lemmatization. After completing text preprocessing, proceed with the process using LDA. Using LDA, two sustainability course topics were identified that are in line with current market demand. The two topics are *Emission Reduction Strategies for Airports* and *Green Finance for Sustainable Airports*. These topics are based on the *top keywords* from the overall data that emerged from the *sustainability* documents. From the results of the gap analysis and LDA conducted, there is 1 topic that is desired by the market but is not currently used as a theme for courses by international organization course providers related to sustainability, namely *Green Finance for Sustainable Airports*. It will help course providers to consider opening new classes related to the subject.

Bibliography

Airport Carbon Accreditation (2024): *Accredited Airports Across the World. Available at*: https://www.airportcarbonaccreditation.org/accredited-airports/

Agustina, I., Pradesa, H. A. (2024). *Praktek Pelaporan Keberlanjutan Di Indonesia: Sebuah Telaah*

Kritis Atas Literatur Terdahulu, Semarang: Jurnal Ekonomi, Manajemen, Akuntansi, dan Perpajakan (Jemap), https://doi.org/10.24167/jemap.v7i1.10947

Bastian, E. I. (2023). *Unsupervised Framework Design for Text-Based Safety Analysis, Undergraduate Thesis*, Institut Teknologi Bandung.

Climate Now by Copernicus (2024): *Climate Bulletins. Available* at: https://climate.copernicus.eu/climate-bulletins

Davalas, A. (2023). *The Importance of the TAM-SAM-SOM Model And How Big Data and AI Help. International Journal of Social Science and Economic Research*, 08(12), 3936–3944. https://doi.org/10.46609/ijsser.2023.v08i12.016

Dimitriou, D., & Karagkouni, A. (2022). *Due Diligence of Transport Infrastructure Operators Sustainability: A Circular Economy Driven Approach. Frontiers in Sustainability*, *3*. https://doi.org/10.3389/frsus.2022.916038

Kang, H., and Kim, J. (2022). *Analyzing and Visualizing Text Information in Corporate Sustainability Reports Using Natural Language Processing Methods. Applied Sciences* (Switzerland), 12 (11). https://doi.org/10.3390/app12115614

Kurtagić, S., and Kadric, D. (2022). *Improvement of Airports Energy Efficiency, Case Study of Airport Sarajevo. Proceedings of the 33rd International DAAAM Symposium* 2022, 133 – 142. https://doi.org/10.2507/33rd.daaam.proceedings.019

Liu, W., & Zhu, P. (2024). *The Impact of Green Finance on the Intensity and Efficiency of Carbon Emissions: The Moderating Effect of the Digital Economy. Frontiers in Environmental Science*, 12, 1362932. https://doi.org/10.3389/fenvs.2024.1362932

Perez, C. L. and Solano, H. J. L. (2025). *A Comprehensive Guide to Latent Dirichlet Allocation: Building a Solid Foundation with Key Statistical Concepts*. https://doi.org/10.13140/RG.2.2.20062.34883

Mayola, L., Hafizh, M., and Putra, D. M. (2024). *Algoritma Jaccard Similarity untuk Deteksi Kemiripan Judul Disertasi dengan Pendekatan Variasi Stop Word Removal*. Jurnal Media Informatika Budidarma, 8(1), 477. https://doi.org/10.30865/mib.v8i1.7109

Mee, A., Homapour, E., Chiclana, F., and Engel, O. (2021). *Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit.* Knowledge-Based Systems, 228. https://doi.org/10.1016/j.knosys.2021.107238.

Rishu, & Kukreja, V. (2024). *Comic exploration and Insights: Recent trends in LDA-Based recognition studies. Expert Systems with Applications*, 255. https://doi.org/10.1016/j.eswa.2024.124732

Rose, R. L., Puranik, T. G., & Mavris, D. N. (2020). *Natural language processing based method for clustering and analysis of aviation safety narratives. Aerospace*, *7*(10), 1–22. https://doi.org/10.3390/aerospace7100143

Sustainability Course Demand Data 2016 - 2024 on Google Search Engine (2024). Available at: https://trends.google.fr/trends/explore?date=2016-01-01%202024-07-14&q=Sustainability%20Course&hl=id

Tresnasari, N. A., Adji, T. B. and Permanari, A. E. (2020): *Social-Child-Case Document Clustering based on Topic Modeling using Latent Dirichlet Allocation*, Indonesian Journal of Computing and Cybernetics Systems, 14 (2), 179-188.

Vestermark, H. (2024). *A Tool for Optimizing Regular Expressions*. https://doi.org/10.13140/RG.2.2.31652.90244

World Economic Forum (2011). *Policies and Collaborative Partnerships for Sustainable Aviation, World Economic Forum*, Geneva, 20-21. Available at: https://www3.weforum.org/docs/WEF_ATT_SustainableAviation_Report_2011.pdf

Yang, C., & Huang, C. (2023). *Natural Language Processing (NLP) in Aviation Safety: Systematic Review of Research and Outlook into the Future. Aerospace*, 10(7), 600. https://doi.org/10.3390/aerospace10070600