aerospace

MDPI

Article

Natural Language Processing of Aviation Safety Reports to Identify Inefficient Operational Patterns

Ayaka Miyamoto, Mayank V. Bendarkar * and Dimitri N. Mavris

Aerospace Systems Design Laboratory, Georgia Institute of Technology, Atlanta, GA 30332, USA
* Correspondence: mbendarkar3@gatech.edu

Abstract: With the growth in commercial aviation traffic and the need for improved environmental performance, strategies to lower emissions that can be implemented in the near term are necessary. Since novel technology takes time to enter the market, operational improvements that employ existing aircraft and require no new infrastructure are fit for this goal. While quantified data collected throughout aviation, such as arrival/departure statistics and flight data, have been well-utilized, text data collected through safety reports have not been leveraged to their full extent. In this paper, a methodology is presented that can use aviation text data to identify high-level causes of flight delays and cancellations, using delays as a metric of operational inefficiency. The dataset is extracted from the Aviation Safety Reporting System (ASRS), which includes voluntary safety incident reports in text narrative and metadata formats. The methodology uses natural language processing tools, K Means clustering, and dimensionality reduction by t-Distributed Stochastic Neighbor Embedding (t-SNE) to categorize and visualize narratives. The method identified 7 major clusters and a total of 23 sub-clusters. A comparison between the subclusters' topics and the causes of flight delays revealed by the quantified data shows that the ASRS database provides a unique safety perspective to delay cause identification, as illustrated by the method's identification of maintenance as the main cause of delays, rather than weather.

Keywords: natural language processing; text mining; aviation safety reporting system; flight delay; clustering

Citation: Miyamoto, A.; Bendarkar, M.V.; Mavris, D.N. Natural Language Processing of Aviation Safety Reports to Identify Inefficient Operational Patterns. *Aerospace* **2022**, *9*, 450. https://doi.org/10.3390/aerospace9080450

Academic Editor: Álvaro Rodríguez-Sanz

Received: 27 June 2022
Accepted: 11 August 2022
Published: 17 August 2022

1. Introduction and Background

In the face of rapid climate change, commercial aviation must mitigate its environmental impact. The European Commission's Flight Path 2050 has set emission reduction targets of 75% CO_2 reductions, 90% NOX reductions, and 65% perceived noise reductions [1]. Similar emission reduction targets have been put forth by ICAO and IATA, such that the net aviation CO_2 emissions from 2020 would be carbon-neutral and a 50% reduction of CO_2 emission would be sustained until 2050 [2]. Improvements in airframe and engine, the deployment of sustainable aviation fuels, the introduction of government regulations and economic measures to discourage activities that produce CO_2 emissions, and improved operations and air traffic management are the four main research areas regarding attempts to attain net-zero CO_2 emissions by 2050 [3]. Improvements in aircraft fuel efficiency can be made by designing more-electric architectures [4,5] or hybrid-electric powertrains [6,7]. However, these concepts are likely to take decades, due to issues with battery energy densities [8] and safety and reliability considerations [9,10]. Due to the urgency with which humanity must address global warming, waiting for new ultra-efficient aircraft or novel propulsion technology to enter the market is not sufficient. As improvements in operations and ATM can be implemented more quickly, requiring no new infrastructure and employing existing aircraft, research in this area has the potential to improve aviation's environmental performance within a shorter timeframe. The present work develops a new

approach to identify areas of operational inefficiency, which may reduce emissions and costs while improving safety and passenger satisfaction.

Ongoing operations and ATM research include arrival profile optimization, improvements in navigational technology, and weather-related improvements. For example, continuous descent arrival procedures encourage a continuous descent with no intermediate level-offs upon arrival to the destination airport. As a result, longer portions of the flight are at cruise altitude, which is more efficient and requires lower power levels, thus leading to lower emissions. Navigational improvements include the deployment of Automatic Dependent Surveillance-Broadcast (ADS-B) technology, which enables the more precise control of aircraft. Finally, the FAA's Next-Generation Air Transportation System's (NextGen) network-enabled weather technology provides advanced, real-time weather data to reduce weather-related delays, with the potential to take advantage of existing weather conditions to improve efficiency [11]. Flight delays is one metric that could be used to identify such operational inefficiencies. The mitigation of flight delays is interesting from an environmental and financial perspective. Environmentally, a reduction in flight delays will optimize and minimize aircraft operation time, which will reduce operational emissions from aviation. From a financial perspective, a study that created a statistical cost estimation model of the effect of delays on airline cost found that about 20% of airline flights arrive more than 15 min late, which costs billions of dollars annually [12]. For example, in 2007, the cost of flight delays for the U.S. economy was estimated to be $32.9 billion, for which more than half was charged to passengers [12]. To compound this problem, it is expected that delays will increase as air traffic demands grow, as it is recognized that delays nonlinearly increase as demand approaches the capacity of the air transportation system [12]. Growing delays will impair the passenger experience of air travel, worsen aviation's environmental impact, and place a financial burden on airlines, customers, and the global economy.

The aviation industry collects vast amounts of data daily, both qualitative and quantitative, from various sources and in multiple formats, including voice recordings of air traffic control, written reports from aviation personnel, and flight data from on-board measurement devices. Research utilizing these aviation quantitative data to analyze aircraft operations abounds [13,14]. Relatively little research has been conducted on aviation text data for delay-cause-identification, as they are incompatible with conventional computational methods of analysis [15]. The use of aviation text data is a novel approach to the investigation and mitigation of flight delays. With the use of machine learning methods such as Natural Language Processing (NLP), text data can be used with minimal manual labor, using computational data analysis. This study addresses the gap in the application of machine learning to aviation text data to identify the causes of flight delays.

The present work answers two research questions. First, what are causes of operational inefficiency, specifically flight delays? Second, do textual data hold information that cannot be observed through quantified data? The final products of this study are a repeatable NLP-clustering methodology, which can be expanded to any text-based data source, and the identification of causes of flight delays based on information that is available in the selected text-based dataset. The aviation database chosen for this project was selected based on accessibility, the inclusion of narrative-style text data, and the inclusion of operational information. The operational information desired for this study is a detailed report of the events leading to a flight delay. Narrative accounts of aviation events can be found in databases such as the FAA Accident and Incident Data System, NTSB Report Database, and Aviation Safety Reporting Systems (ASRS). The ASRS database is selected for this study (Available online: https://asrs.arc.nasa.gov/ (accessed on 16 June 2022)). ASRS holds voluntary, confidential safety information from frontline aviation personnel, including pilots, controllers, mechanics, flight attendants, and dispatchers in the form of text narratives, along with various metadata characterizing the flight. These narratives are collected for the purpose of policy development, human factors, education, and training. ASRS offers an extensive filtering system, allowing the user to extract the reports which fit the user's criteria of date, environment, aircraft, reporter, event assessment, etc. This

includes a filter for all events resulting in "Flight Cancelled/Delayed". Additionally, ASRS reports are publicly available and the selected ASRS reports can be extracted in HTML, Word, Excel, and CSV formats.

2. Literature Review

Known causes of flight delays have been identified using quantitative metadata or surveillance records. A study of flight delays in the U.S. found bad weather, carrier equipment, or technical airport problems to be the main causes [16]. Delay statistics for the National Airspace System (NAS), based on data from the Post Operations Evaluation Tool database, found that weather accounted for 69% of arrival and departure delays causes in 2000, followed by traffic volume, runway delays, ATC equipment problems, and other causes [17]. A study focusing on the delay characteristics at Newark International Airport similarly found that convective weather was the main cause for delays averaging longer than 15 min per arrival [18]. A European study found that longer delays were due to technical maintenance issues or aircraft defects while shorter delays were due to operational control, crew duty norms, ATC, and airport limitations, with the chain effects of delays seen as delays to previous flights of the same plane resulted in delays to later flights [19]. Considering the pre-COVID pandemic departure cause analysis from Eurocontrol [20], the largest causes for delays were reactionary, followed by airline, ATC, and weather. Mid-pandemic, the total number of delays went down, but delays due to government causes went up [21]. A study conducted by the Civil Aviation Administration of China found that nearly 50% of delays are caused by severe weather, while roughly 20% of delays are caused by air route problems [22]. The studies used a statistical analysis of quantitative metrics such as average departure delay per flight and percentage of en-route and airport delays or creating predictive models of flight delays using aviation big data and machine learning. The flight delay prediction models were produced using quantitative data including ADS-B surveillance data, date and time, air route, airport. The model was used to identify the most influential factors determining if a flight would experience a delay.

In addition to the causes of delays, there has been interest in the research of delay propagation. For example, if flight A's arrival is delayed at its destination airport, flight B, which uses the same aircraft as flight A, may also be delayed. This allows for a chain of delays to occur, leading to congestion in the airspace. Congestion can spread through airline fleets and airport systems, and congestion is expected to grow as leisure and business flying demands grow in parallel with economic growth. The Department of Transport (DOT) in the US reported that American Airlines experienced 5.6 million minutes of delays in the month of March 2018 due to traffic growth, resulting in increased costs to cover longer working hours, host passengers during the delay, and other issues [23]. Subject matter experts studying the challenges that the FAA's NextGen system must address have identified improvements in system efficiency and robustness for the reduction in flight delays as a major focus for NextGen. Limited runway capacity has been identified as a cause of flight delays, which has the potential to be mitigated, in addition to disruptions by weather, which have been identified in previous studies [24]. Runway capacity is the main limiting factor to consider when regulating the queues of departing aircraft and ensuring the orderly and well-timed landings of arriving aircraft. Of the main delay causes that were identified, carrier equipment and maintenance issues are operations-related issues that may be mitigated.

A review of the literature on aircraft maintenance planning and maintenance-related delays suggests that unscheduled maintenance can lead to costly delays or flight cancellations. Of the different types of maintenance, line maintenance, which includes routine checks, post-flight inspections, and malfunction ratification, has the power to decide GO/NOGO for the next flight, making line maintenance an operational phase of interest [25]. In the case of unscheduled maintenance due to issues found in line maintenance, the probability of delays caused by maintenance strongly correlates with the availability of resources at each airport. Research into aircraft maintenance records and maintenance

employee interviews to identify the causes of delays in line maintenance corroborate the findings on the importance of resource availability. Maintenance reports identified that poor logistics processes (i.e., availability of spares) accounted for nearly 70% of causes of maintenance delays during line maintenance, followed by unscheduled maintenance and defects found during pilot reports, accounting for 29% of maintenance delays, and poor planning, accounting for 2% [26]. Improved logistics processes were suggested to ensure the availability of spares and better planning of line maintenance activities to organize work packages and manage human resources.

The implementation of TF-IDF and K Means algorithms for document clustering has been investigated by the computing industry, providing insights into the effectiveness of this methodology in non-aviation-specific applications. For the clustering of text data, the text can be represented as a binary vector or a more refined weighted method such as TF-IDF [27]. Singh et al. [28] compared the clustering results of K Means, Heuristic K Means and Fuzzy C Means for documents represented in TF, TF-IDF, and Boolean representations with different feature schemes. For a dataset of standard newswire articles, the study found that the use of TF-IDF with stemming resulted in the most successful clustering, and fuzzy clustering was found to perform better than K Means and heuristic K Means. Khan et al. [29] investigated the efficacy of a TF-IDF/K Means methodology to summarize texts and found that the careful preprocessing of all unnecessary characters, keywords, tags, and punctuations is vital. Gowtham et al. [30] applied TF-IDF and Boolean methods to documents and clustered these using K Means for document classification. Similar to the findings of Khan et al., the importance of proper pre-processing was stressed. The study concluded that the combination of TF-IDF with K Means achieved the best results, as TF-IDF performed better than Boolean and K Means was found to be more efficient than other clustering algorithms.

Finally, the literature on the use of NLP in the ASRS dataset is limited in its scope and volume. Tanguy et al. [31] presented a support vector-machine-based automatic classification task as well as a topic modeling task using ASRS reports to identify safety needs for experts. Robinson [32] implemented topic modeling on fourteen years of ASRS reports to identify changes in safety trends over time. Subramanian and Rao [33] analyzed and classified the key factors contributing to go-around and missed-approach reports from the ASRS database and trained a model to forecast the number of incidents over a given period of time. Ghaoui et al [34] utilized a term-frequency, inverse-document frequency (TF-IDF) technique to discover the four main tasks that pilots perform during flight, which can aid in understanding the causal and contributing factors to runway incursions and other drivers for aviation safety incidents. Ref. [35] applied the Bidirectional Encoder Representations from Transformers (BERT), an attention-based language model, to determine the answer to the question "When did the incident happen?" for a set of ASRS reports. As noted earlier, little to no literature was found by the authors on the utilization of NLP on aviation text data to determine the causes of flight delays.

In summary, general flight delay causes in the literature were found to be weather, runway capacity, and maintenance issues. Within maintenance-related delays, most delay causes stem from line maintenance, in which resource availability and unscheduled maintenance are key factors. Preventing one delay-causing event can prevent multiple flights from being delayed, as delay propagation from one flight to another is a known problem in aircraft maintenance routing. However, most of the above studies utilize different types of quantitative data to analyze flight delays, leaving a gap in the analysis of qualitative text-based data. The present work seeks to address this gap by analyzing aviation text data.

3. Methodology

To answer the research questions, this study applies NLP and computational data analysis tools to analyze aviation text data and compares the insights drawn from the text data to the causes of flight delays identified in the literature review. While deep learning techniques that better utilize contextual information exist (e.g., BERT [36]), the present

work focuses on utilizing a Bag-of-Words (BoW) and a TF-IDF technique to process the ASRS text narratives. Transformer-based deep-learning techniques that better capture the context will be considered in future work. An overview of this methodology is found in Figure 1.

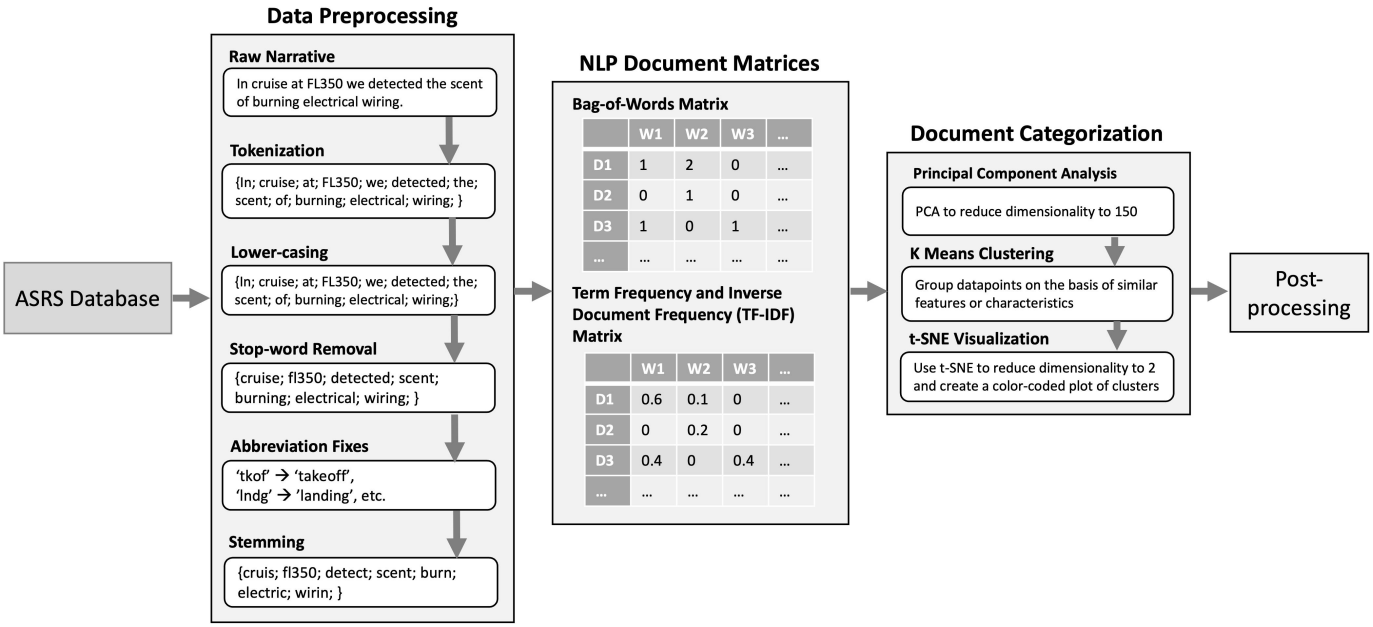


Figure 1. Methodology overview for data preprocessing, NLP implementation, and ASRS report classification leading up to post-processing (adapted from Ref. [37]).

3.1. Data Preprocessing

The dataset is selected by filtering for FAR part 121 and events resulting in "Flight Cancelled/Delayed", resulting in 4195 reports from January 1999 to January 2022 from ASRS. This work focuses on commercial operations only by filtering for FAR part 121 operations, as the financial motivations for reducing flight delays are based on the commercial aviation market. The narratives are pre-processed to address two points. The first is to clean this in preparation for NLP and the second is to address the inconsistent use of abbreviations. The cleaning of the narratives is completed in four steps: tokenization, lower-casing, stop-word removal, and stemming. Tokenization takes the sentences contained in the narrative and separates them by word, so that each word is a token used to vectorize the narrative. Lower-casing ensures that the capitalization of letters does not interfere with identical words, which are considered distinct. For example, tokens such as "Flight" and "flight" are combined, ensuring that word counts are not affected by differing the capitalization and reducing the dimensionality of the dataset. Stop-word removal removes words that provide little insight into the characterization of the document, such as "the" or "and". The final step reduces words to their root forms to combine similar words, such as different tenses of the same verb, for the same objectives as lower-casing. This step can be accomplished using stemming or lemmatization. Stemming trims a word to its absolute root, resulting in a token that is not a real English word, while lemmatization converts this to the most closely related existing English word. As the narratives in the ASRS dataset lack speech tags which lemmatization often requires, stemming is chosen as the final preprocessing step for NLP [38]. ASRS narratives often include abbreviations, but the abbreviations are not constant across multiple reports. For example, the word "takeoff" may be written as "tkof" or "takeoff" depending on the reporter. As the NLP process used in this methodology checks for identical words, "tkof" and "takeoff" are clustered separately. A preprocessing step to identify and fix the 17 most common abbreviations is implemented to minimize the effect of this issue. After tokenization, lower-casing, and removing stop words, the word or the root form of each word is checked against a list of English words in Python's NLTK

library. If a word, or the word with suffixes such as 's', 'ed', 'er', etc., removed, is not recognized by the list of English words, it is flagged as an abbreviation. The average frequency of abbreviations per report was calculated to be 11%. The list of 17 edited abbreviations are in Table 1. Abbreviations that held more significance than the full word were kept in their original form. This included terms such as air traffic control (ATC), Quick Reference Handbook (QRH), auxiliary power unit (APU), and minimum equipment list (MEL).

Table 1. Preprocessed abbreviations list

Abbreviation	Full Word
acft	aircraft
eng	engine
flt	flight
rptr	reporter
capt	captain
lndg	landing
rwy	runway
emer	emergency
kt	knot
tkof	takeoff
gnd	ground
apch	approach
chk	check
pwr	power
evac	evacuation
hyd	hydraulic
mech	mechanic

3.2. Natural Language Processing

NLP is a machine learning technique that allows for a computer to handle human language. The purpose of employing NLP is to sort large quantities of text data without manual sorting. The preprocessed narrative vectors are used to generate a bag-of-words (BoW) matrix, in which each row represents a narrative, and the columns are stemmed words. For each narrative, the word count for each word in the columns are created, resulting in a numerical representation of the words of which it is composed. The numerical matrix form reduces dimensionality and can be manipulated in the subsequent clustering algorithm step [37]. The Term Frequency and Inverse Document Frequency (TF-IDF) method is employed to identify the characterizing words of each report. The TF-IDF method is a two-step process. First, in the TF step for a given report, the frequency of each term is normalized using the word count of the most frequent word in the given report. In the IDF step, the inverse document frequency is calculated for each word of each report based on Equation (1) [38]. The result is a matrix of a TF-IDF score for each word in each report, such that words that are commonly seen across all reports are given scores close to 0, while unique words that frequently appear in select reports are given higher scores. The high-scoring words can be interpreted as the key words that characterize the report. The inherent normalization when computing the TF-IDF score allows for the scores to be compared across reports, regardless of the varying lengths of the reports. Readers interested in better understanding the TF-IDF algorithm are directed to Ref. [39].

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,d) \tag{1}$$

where,

$$tf(t,d) = \frac{f_{t,d}}{max\{f'_{t,d} : t' \in d\}} \tag{2}$$

$$idf(t,D) = log\frac{N}{|\{d \in D : t \in d\}|} \tag{3}$$

3.3. Clustering

To categorize the reports based on the characteristic words of each report, we employ K Means clustering. Clustering is a machine-learning technique to categorize data based on similar features. In this case, reports with similar language and words are grouped. While there are many clustering techniques, K Means was chosen as prior work using TF-IDF and K Means found success in an aviation-based application [37]. The dimensionality of this task was reduced to 150 components using principal component analysis (PCA), a linear mathematical algorithm which reduces the dimensionality of a dataset while retaining variation in the dataset [38]. The number of clusters that the reports will be grouped into is user-defined and determined by visually checking a color-coded scatterplot visualization of the clusters and determining if the boundaries are precise and the colored groups are distinct groupings. This visualization was created using t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is an algorithm developed by van der Maaten and Hinton in 2008 to scale high-dimensional data to lower dimensions [40], and will be used to reduce the high-dimensional TF-IDF matrix to two dimensions. Traditional metrics of success for K Means clusters, the Silhouette score and Calinski–Harabasz (CH) score, both encountered issues reflecting the success of clustering text-based data in previous studies [37,41], and therefore will not be reported in this work.

3.4. Observations

For each cluster, a bar chart is produced of the 20 most frequent words among the reports. This is used to manually determine the topic of each cluster that describes the characteristics of the flight delays represented in the cluster. The topic of a cluster was found by identifying the most characteristic or unique words describing a cohesive story. In this process, common words such as 'aircraft' or 'flight', which are not useful in distinguishing one cluster from another, are disregarded. The categorized flight delay events described in the text data were compared to the known causes of flight delays found in the literature. If the findings match, we will have shown the validity of using aviation text data and ASRS for determining flight delay causes. If the findings do not fully match, we will have shown the potential for text data to provide new insights to the sources of operational inefficiency.

4. Results

4.1. Overall Clustering

The 4195 event narratives extracted from ASRS were first preprocessed to make the use of abbreviations consistent. This step resulted in 1.78% of the total words across all reports being edited. The NLP, TF-IDF, and K Means clustering yielded the seven clusters shown in Figure 2, illustrated using t-SNE. Seven clusters were chosen after manually evaluating different numbers of clusters and weighing the costs of increasing model complexity against the benefits of obtaining more unique clusters. Each narrative corresponds to a point in the t-SNE plot. The axes of the plot, 'Dimension 1' and 'Dimension 2', are a result of reducing the high-dimensional matrix of the number of words and number of reports to two dimensions, and thus have no physical meaning.

The topics of each cluster can be discerned from the most frequent words in each cluster, as shown in Figure 3. The main cluster topics were as follows:

0. Engine;
1. Hydraulics;
2. Taxi/Pushback;
3. Landing/Approach;
4. Takeoff;
5. Cabin odor;
6. Maintenance.

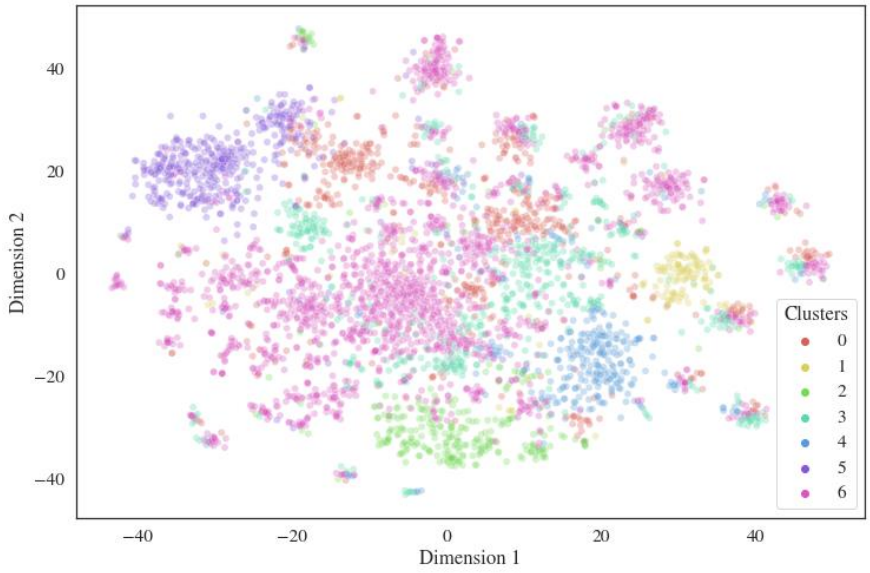


Figure 2. Main clusters identified by K-Means clustering.

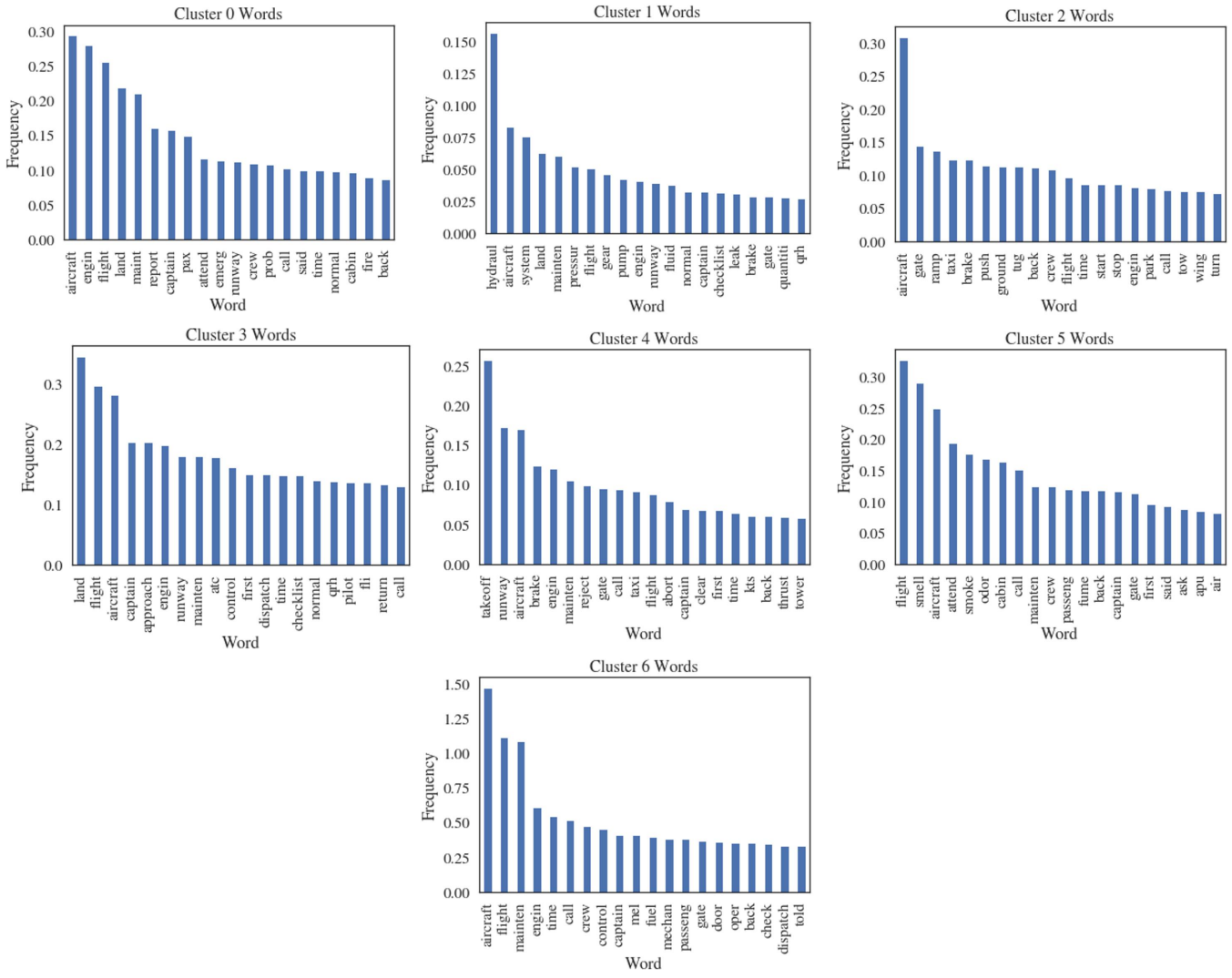


Figure 3. Most Frequent 20 Words of Each Main Cluster.

Note that the cluster numbers range from zero to six. The proportion of reports placed in each cluster is shown in Figure 4. Distinct cluster themes were manually identified based on the most frequently seen words in each cluster, but further investigation is necessary to identify more specific event causes, which may be specifically prevented. The clustering algorithm was applied again to each cluster for the subclustering effort.

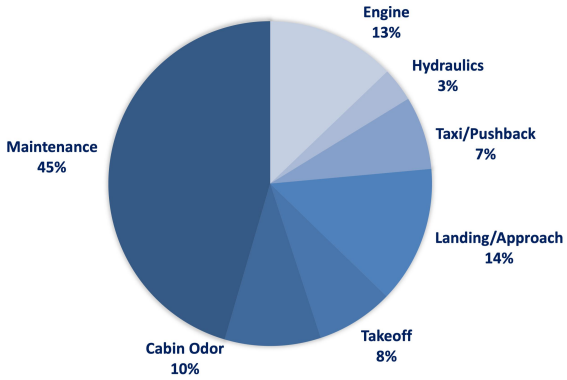


Figure 4. Proportion of Reports in the 7 Main Clusters.

4.2. Sub-Clustering

The seven main clusters were further subclustered, providing a total of 23 clusters. The hierarchical structure of the main clusters and the subclusters they were divided into is illustrated in Figure 5 with the number of event narratives in each cluster/subcluster shown on the bottom right.

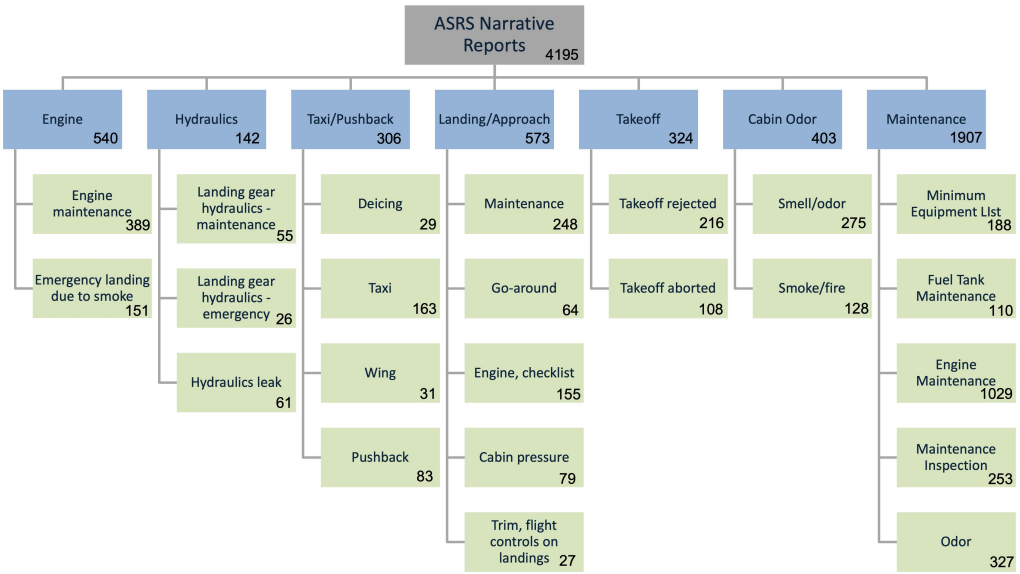


Figure 5. All clusters and subclusters with identified cluster topic and number of reports.

Subclustering yielded more specific safety events that lead to flight delays. However, Cluster 6, the largest main cluster with the main cluster topic of 'Maintenance', resulted in vague subcluster topics. The largest subcluster, accounting for 54% of the reports in Cluster 6, had a generic topic of "Engine Maintenance", as seen in Figure 6. The most frequent words in that subcluster are given in Figure 7. Due to the most frequent words in the subcluster being commonly found words, such as 'aircraft', 'mainten', and 'flight', it appears that Cluster 6 is the catch-all cluster, which captures reports with less defining characteristics or outliers. Other subcluster topics, such as "Odor", have some specificity but raise questions regarding why these reports were placed in Cluster 6 rather than Cluster

5, with the topic of "Cabin Odor". The inherent limitation in identifying cluster topics from individual words is the method's inability to identify key functions, such as dispatchers, pilot, or technicians, or the specific aircraft parts involved in the scenario. Such information would be necessary to use these insights to prevent future flight delays.

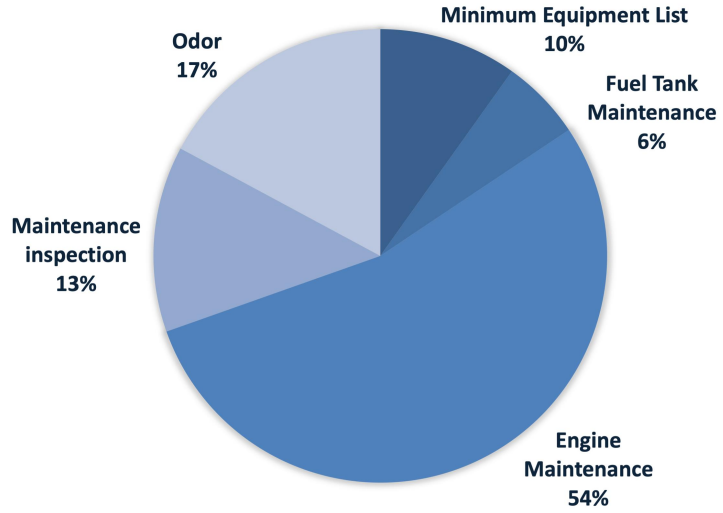
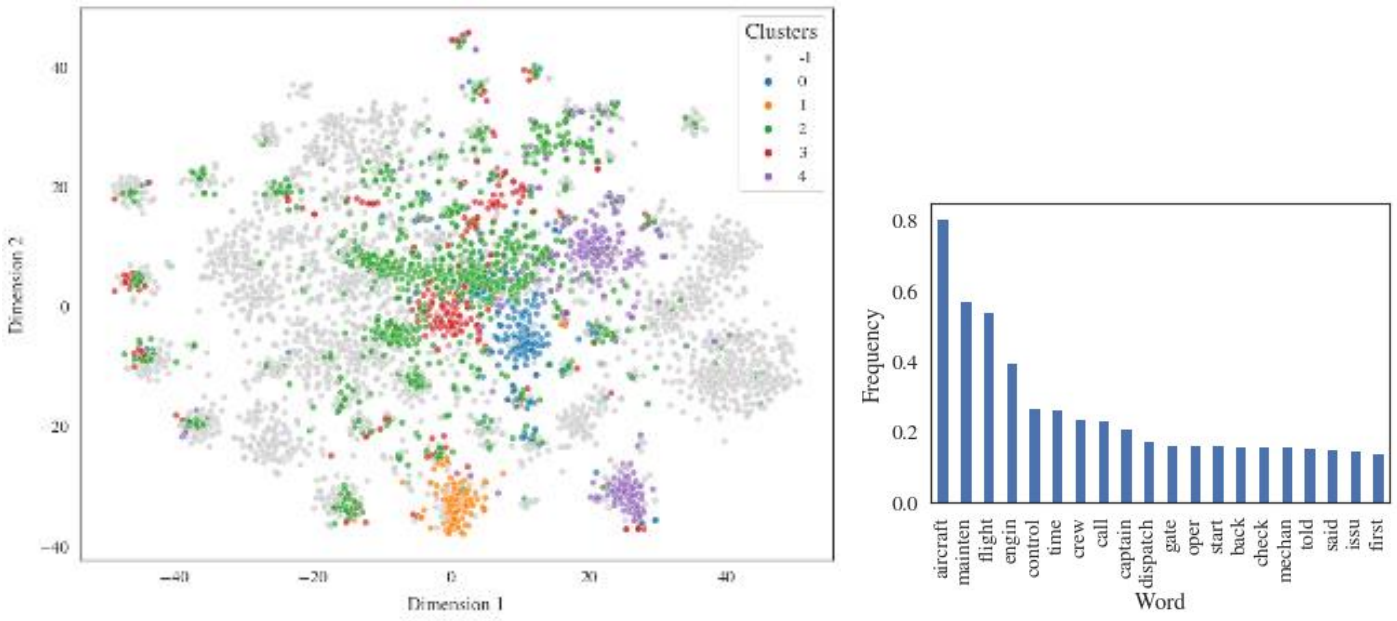


Figure 6. Proportion of reports per subcluster for Cluster 6 'Maintenance'.



(a) (b)

Figure 7. (**a**) t-SNE visualization of the Cluster 6 subclusters. (**b**) Most frequent words in Cluster 6.2 with the topic 'Engine Maintenance'.

Another limitation of this method was its dependency on the use of a consistent vocabulary across all narratives to achieve the best results. This was illustrated in the subclusters found in Cluster 4. The overall cluster topic was identified as "Takeoff", with two subclusters: "Takeoff rejected" and "Takeoff aborted". Rejecting and aborting takeoff have the same meaning but were clustered separately due to different writers using varying vocabulary to describe the same event. The abbreviation removal step in Section 3.2 aimed to address this issue, which is inherent to the TF-IDF approach to employing NLP.

An additional observation was made regarding the effect of ASRS's voluntary reporting structure on an aviation safety study. The reports filed to ASRS are voluntary, meaning that reports are submitted only when the reporter deems the event to be report-worthy. The determination of report-worthiness can differ between reporters, suggesting that the database is not a comprehensive collection of all instances of a given event. Furthermore, the determination of which details of an event should be included or omitted is made by the reporter, which can be affected by factors such as the reporter's function or background. Reports in ASRS are written by pilots, cabin crew, ATC, dispatchers, technicians, UAS crew, and more. Due to the diversity in the function of writers, having an event described from multiple perspectives would produce a comprehensive understanding of the event, but having only one report of an event would result in a biased or partial description.

4.3. Correlation of Clustering Results to Metadata

An investigation between the a metadata type provided in ASRS was conducted to identify trends in the clusters that are not observable through the t-SNE visualizations and the most frequent words in the cluster. ASRS reports include metadata such as date, location, flight phase, aircraft make model, primary problem, contributing factors, etc., in addition to the text-based narratives. As the cluster topics that were identified were often part-specific, such as engine, landing gear or wing, the "Aircraft Component" metadata was selected for further study. Trends in the clusters were identified by plotting bar charts of the "Aircraft Component" metadata input for each cluster and comparing the results to the cluster's most frequent words and other clusters. While such metadata may be added as features to the clustering algorithm in addition from NLP outputs, they are not always available in text databases. Hence, they are used to validate the clustering outcomes in the present methodology.

At both the cluster and subcluster levels, the manually identified cluster topics were consistent with the most frequently associated aircraft components. For example, the topic for Cluster 5 was identified as "Cabin Odor" (Table 2). At the cluster level, the top aircraft component was the Air Conditioning and Pressurization Pack. The subcluster topics were identified as "Smell/Odor" for Subcluster 0 and "Smoke/Fire" for Subcluster 1. For Subcluster 0, the top aircraft components were (1) Air Conditioning and Pressurization Pack, (2) APU, and (3) Coalescer Bag, as seen in Figure 8. The coalescer bag is a filter in the air-conditioning system that collects mist from the air, which can emit a strong odor if it is not cleaned, directly relating to the subcluster topic of smell. For Subcluster 1, the top 2 aircraft components were the same but the third was Electrical Distribution, which relates to the subcluster topic of smoke, as electrical wiring can be a source of smoke or fire. The consistency of the event topic found through NLP and analysis of the most frequently associated aircraft components verifies that the NLP methodology can identify relevant and important information from the safety narratives.

Table 2. Most frequently associated aircraft component metadata for each major cluster.

Cluster	Topic	Top Aircraft Component
0	Engine	Turbine Engine
1	Hydraulics	Hydraulic Main System
2	Taxi/Pushback	Nosewheel Steering
3	Landing/Approach	Turbine Engine
4	Takeoff	Turbine Engine
5	Cabin Odor	Air Conditioning and Pressurization Pack
6	Maintenance	Turbine engine

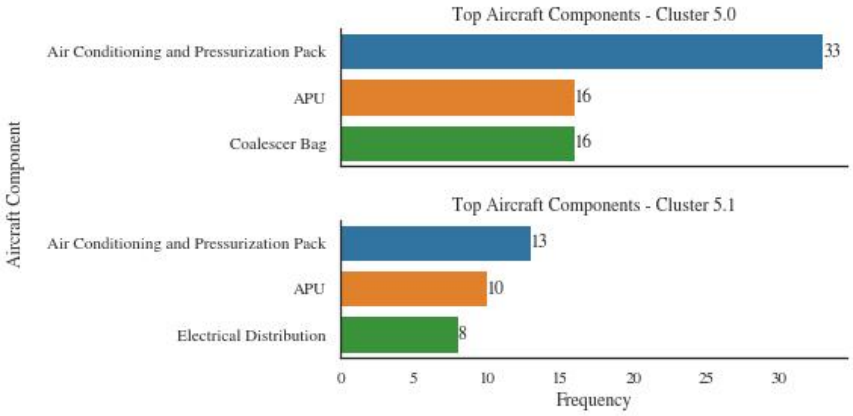


Figure 8. Cluster 5 Subclusters' Aircraft Component Metadata Trends.

While the observations made in Section 4.2 found that the NLP methodology used in this study cannot identify specific aircraft components that relate to the safety event described by itself, it can provide context that cannot be gleaned from the selected metadata. This is well-illustrated by Cluster 1: "Hydraulics". The subcluster topics were found to be (1) landing gear hydraulics—maintenance, (2) landing gear hydraulics—emergency, and (3) hydraulics leak. Determining the reason for the three subclusters would be difficult based on the aircraft component metadata alone, as the metadata for the subclusters are similar, as seen in Figure 9. Despite the vastly different circumstances in which the landing gear incidents occurred for subclusters 1.0 and 1.1, the contextual information would be lost in this metadata type, giving value to the NLP's ability to successfully identify the circumstances in which the safety events took place.

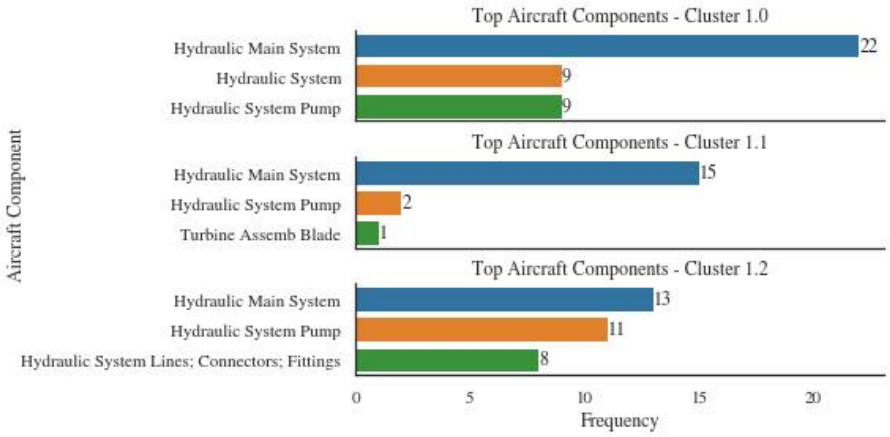


Figure 9. Cluster 1 Subclusters' Aircraft Component Metadata Trends.

4.4. Comparison to Insights from Quantitative Data

The literature review on the known causes of flight delays revealed that weather accounts for a significant portion of delays. However, the NLP-clustering methodology used on the ASRS database resulted in there being little mention of weather, aside from a deicing subcluster under Cluster 2: Taxi/Pushback. This illustrates the unique safety perspective that ASRS narratives bring to delay-cause identification. The events reported in ASRS are safety-oriented events, which exclude normal operations that are captured well by quantified statistical data. This observation is confirmed by comparing the number of reports related to flight delays and cancellations to the reported number of flight delays and cancellations. The Bureau of Transportation Statistics [42] provides the number of flights which were reported as delayed/cancelled or on time per year. While not all airlines report their statistics to BTS and, therefore, the values do not fully represent all flight operations, this value provides a notional understanding of how many delayed/cancelled

flights are reported to ASRS. Comparing the number of relevant ASRS reports that we extracted each year to the statistics in BTS for 2013 to 2021, ASRS captures roughly 0.06% of all flight delay/cancellation instances. As a result of ASRS's specialization in safety-related incidents, non-safety related causes, such as weather and runway capacity, which have received attention in the literature, are not present in the results. Instead, the reports from ASRS emphasize the maintenance aspect of delays. These findings suggest that improvements in NAS safety could also prevent delays and thereby improve emissions and operational costs.

Of the 23 subclcusters, 12 included 'mainten' in the top 20 frequently seen words. Due to the abundant mention of maintenance in the ASRS records, maintenance-related delay causes were investigated. The literature review revealed that unscheduled maintenance and defects found during pilot reports and the unreliable availability of spares account for many maintenance delays during A-check maintenance. To search for clusters related to unscheduled maintenance and spares, the frequency of key words such as 'mechan', 'mainten', 'spare', 'wait', and 'inspect' in each cluster was calculated and ranked. The results of the rankings for each keyword are shown in Table 3. The Maintenance–Minimum Equipment List subcluster being most highly ranked for the keywords 'mechan' and 'mainten' suggests that maintenance requests relating to the minimum equipment list are a frequent cause of unscheduled maintenance, leading to delays.

Table 3. Subclusters ranked by frequency of maintenance-related keywords

Keyword	Cluster–Subcluster Topic with Most Frequent Use of Keyword	Cluster–Subcluster Topic with Second Most Frequent Use of Keyword
Inspect	Maintenance–Maintenance Inspection	Taxi/Pushback–Deicing
Mechan	Maintenance–Minimum Equipment List	Hydraulics leak
Mainten	Maintenance–Minimum Equipment List	Hydraulics leak
Spare	Term not used in dataset	
Wait	Taxi/Pushback–Deicing	Maintenance–Odor

5. Conclusions

This work implemented a framework to identify high-level causes of flight delays by applying NLP and clustering algorithms to aviation text data. Seven major categories and a total of 23 more-detailed topics were identified for ASRS reports of events resulting in flight delays or cancellations, each of which represent a circumstance in which the delay occurred. These seven major circumstances were engine, hydraulics, taxi/pushback, landing/approach, takeoff, cabin odor, and maintenance. A comparison between the 23 ASRS narrative topics and the causes of flight delays identified in the literature revealed the unique benefits and limitations of using an aviation safety database for operations research. While ASRS narratives failed to capture the characteristics of normal operations, such as the prevalent effect of weather on the flight delays highlighted in the literature, it provides a safety perspective to the identification of causes of delays. The usefulness of ASRS in cause identification suggests that improvements in NAS safety can also prevent flight delays. The present work has shown the an NLP-clustering method's ability to identify high-level causes and the circumstances in which delays occur. The methodology can also be extended to any narrative-style data. However, more specific insights, such as specific aircraft parts that require attention or the procedure that directly leads to delays, cannot be identified from this method alone. Future work will include an exploration of other clustering algorithms such as Agglomerative Hierarchical to compare the clustering results to that of the present work. Further investigation of the subclusters by looking for correlations with more metadata types, such as reporter, primary problem, and contributing factors, could lead to more specific insights into clusters with vague topics, such as the Maintenance cluster in Cluster 6. Additionally, a comparison of the frequency of words across clusters could shed light on the value of words that were commonly used and served little purpose in the determination of cluster topics in the present methodology. Finally,

subdividing the reports by flight phase before clustering to reflect the five categories of delays that the FAA uses (gate, taxi-out, enroute, terminal, and taxi-in) or the four categories of delays that the DOT uses (gate, taxi-out, airborne, taxi-in) may help to obtain clarity regarding the cluster topics.

Author Contributions: Conceptualization and Methodology, A.M. and M.V.B.; Data curation, Implementation, Formal analysis, and Investigation, A.M.; Writing—original draft preparation, A.M.; Writing—review and editing, M.V.B.; Supervision, M.V.B.; Resources, D.N.M.; Project Administration, D.N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author would like to thank Tejas Puranik and Rodrigo Rose for their feedback and comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Aircraft Technology Roadmap to 2050*; Report; IATA: Montreal, QC, Canada, 2020.
2. *Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis*; Report; ICAO: Montreal, QC, Canada, 2020.
3. *Destination 2050—A Route to Net Zero European Aviation*; Report NLR-CR-2020-510; NLR—Royal Netherlands Aerospace Centre and SEO Amsterdam Economics: Amsterdam, The Netherlands, 2021.
4. Bendarkar, M.V.; Rajaram, D.; Cai, Y.; Mavris, D.N. Optimal Paths for Progressive Aircraft Subsystem Electrification in Early Design. *J. Aircr.* **2022**, *59*, 219–232. [CrossRef]
5. Jones, R. The More Electric Aircraft: The past and the future? In Proceedings of the IEE Colloquium on Electrical Machines and Systems for the More Electric Aircraft, London, UK, 9 November 1999; pp. 1–4.
6. Antcliff, K.R.; Capristan, F.M. Conceptual Design of the Parallel Electric-Gas Architecture with Synergistic Utilization Scheme (PEGASUS) Concept. In Proceedings of the 18th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Denver, CO, USA, 5–9 June 2017. [CrossRef]
7. Cinar, G.; Cai, Y.; Bendarkar, M.V.; Burrell, A.I.; Denney, R.K.; Mavris, D.N. System Analysis and Design Space Exploration of Regional Aircraft with Electrified Powertrains. In Proceedings of the AIAA SCITECH 2022 Forum, San Diego, CA, USA, 3–7 January 2022. [CrossRef]
8. Bills, A.; Sripad, S.; Fredericks, W.L.; Singh, M.; Viswanathan, V. Performance Metrics Required of Next-Generation Batteries to Electrify Commercial Aircraft. *ACS Energy Lett.* **2020**, *5*, 663–668. [CrossRef]
9. Bendarkar, M.V.; Sarojini, D.; Mavris, D.N. Off-Nominal Performance and Reliability of Novel Aircraft Concepts during Early Design. *J. Aircr.* **2022**, *59*, 400–414. [CrossRef]
10. Papathakis, K.V.; Burkhardt, P.A.; Ehmann, D.W.; Sessions, A.M. Safety Considerations for Electric, Hybrid-Electric, and Turbo-Electric Distributed Propulsion Aircraft Testbeds. In Proceedings of the 53rd AIAA/SAE/ASEE Joint Propulsion Conference, Atlanta, GA, USA, 10–12 July 2017. [CrossRef]
11. *Aviation and Climate Change: Aircraft Emissions Expected to Grow, but Technological and Operational Improvements and Government Policies Can Help Control Emissions, Report to Congressional Committees, GAO-09-554*; Report; US Government Accountability Office: Washington, DC, USA, 2009.
12. Ball, M.; Barnhart, C.; Dresner, M.; Hansen, M.; Neels, K.; Odoni, A.; Peterson, E.; Sherry, L.; Trani, A.; Zou, B. Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States. Available online: https://rosap.ntl.bts.gov/view/dot/6234 (accessed on 16 June 2022).
13. Basora, L.; Olive, X.; Dubot, T. Recent Advances in Anomaly Detection Methods Applied to Aviation. *Aerospace* **2019**, *6*, 117. [CrossRef]
14. Gavrilovski, A.; Jimenez, H.; Mavris, D.N.; Rao, A.H.; Shin, S.; Hwang, I.; Marais, K. Challenges and Opportunities in Flight Data Mining: A Review of the State of the Art. In Proceedings of the AIAA Infotech @ Aerospace, San Diego, CA, USA, 4–8 January 2016; Available online: https://arc.aiaa.org/doi/pdf/10.2514/6.2016-0923 (accessed on 16 June 2022).
15. Madeira, T.; Melício, R.; Valério, D.; Santos, L. Machine Learning and Natural Language Processing for Prediction of Human Factors in Aviation Incident Reports. *Aerospace* **2021**, *8*, 47. [CrossRef]
16. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. Using scalable data mining for predicting flight delays. *ACM Trans. Intell. Syst. Technol. TIST* **2016**, *8*, 1–20. [CrossRef]

17. Mueller, E.; Chatterji, G. Analysis of Aircraft Arrival and Departure Delay Characteristics. In Proceedings of the AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum, Los Angeles, CA, USA, 1–3 October 2002. [CrossRef]
18. Allan, S.; Beesley, J.; Evans, J.; Gaddy, S. Analysis of delay causality at Newark International Airport. In Proceedings of the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 3–7 December 2001; pp. 1–11.
19. Zámková, M.; Prokop, M.; Stolín, R. Factors influencing flight delays of a European airline. *Acta Univ. Agric. Silvic. Mendel. Brun.* **2017**, *65*, 1799–1807. [CrossRef]
20. *All-Causes Delay and Cancellations to Air Transport in Europe for 2019*; Report CDA_2019_004; Eurocontrol: Brussels, Belgium, 2020.
21. *All-Causes Delay and Cancellations to Air Transport in Europe for 2021*; Report CDA_2021_04; Eurocontrol: Brussels, Belgium, 2022.
22. Gui, G.; Liu, F.; Sun, J.; Yang, J.; Zhou, Z.; Zhao, D. Flight delay prediction based on aviation big data and machine learning. *IEEE Trans. Veh. Technol.* **2019**, *69*, 140–150. [CrossRef]
23. Eltoukhy, A.E.; Wang, Z.; Chan, F.T.; Chung, S.H.; Ma, H.L.; Wang, X. Robust aircraft maintenance routing problem using a turn-around time reduction approach. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *50*, 4919–4932. [CrossRef]
24. Balakrishnan, H.; Clarke, J.P.; Feron, E.M.; Hansman, R.J.; Jimenez, H. Challenges in Aerospace Decision and Control: Air Transportation Systems. In *Advances in Control System Technology for Aerospace Applications*; Feron, E., Ed.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 109–136. [CrossRef]
25. Papakostas, N.; Papachatzakis, P.; Xanthakis, V.; Mourtzis, D.; Chryssolouris, G. An approach to operational aircraft maintenance planning. *Decis. Support Syst.* **2010**, *48*, 604–612. [CrossRef]
26. Mofokeng, T.J.; Marnewick, A. Factors contributing to delays regarding aircraft during A-check maintenance. In Proceedings of the 2017 IEEE Technology & Engineering Management Conference (Temscon), Santa Clara, CA, USA, 8–10 June 2017; IEEE: Santa Clara, CA, USA, 2017; pp. 185–190. [CrossRef]
27. Allahyari, M.; Pouriyeh, S.A.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K.J. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv* **2017**, arXiv:1707.02919.
28. Singh, V.K.; Tiwari, N.; Garg, S. Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means. In Proceedings of the 2011 International Conference on Computational Intelligence and Communication Networks, Washington, DC, USA, 7–9 October 2011; pp. 297–301. [CrossRef]
29. Khan, R.; Qian, Y.; Naeem, S. Extractive based text summarization using k-means and tf-idf. *Int. J. Inf. Eng. Electron. Bus.* **2019**, *11*, 33.
30. Gowtham, S.; Goswami, M.; Balachandran, K.; Purkayastha, B.S. An Approach for Document Pre-processing and K Means Algorithm Implementation. In Proceedings of the 2014 Fourth International Conference on Advances in Computing and Communications, Kochi, India, 27–29 August 2014; pp. 162–166. [CrossRef]
31. Tanguy, L.; Tulechki, N.; Urieli, A.; Hermann, E.; Raynal, C. Natural language processing for aviation safety reports: From classification to interactive analysis. *Comput. Ind.* **2016**, *78*, 80–95. [CrossRef]
32. Robinson, S. Temporal topic modeling applied to aviation safety reports: A subject matter expert review. *Saf. Sci.* **2019**, *116*, 275–286. [CrossRef]
33. Subramanian, S.V.; Rao, A.H. Deep-learning based Time Series Forecasting of Go-around Incidents in the National Airspace System. In Proceedings of the 2018 AIAA Modeling and Simulation Technologies Conference, Atlanta, GA, USA, 25–29 June 2018. [CrossRef]
34. El Ghaoui, L.; Li, G.C.; Duong, V.A.; Pham, V.; Srivastava, A.N.; Bhaduri, K. Sparse machine learning methods for understanding large text corpora. In Proceedings of the CIDU, Mountain View, CA, USA, 19–21 October 2011; pp. 159–173.
35. Kierszbaum, S.; Lapasset, L. Applying Distilled BERT for Question Answering on ASRS Reports. In Proceedings of the 2020 New Trends in Civil Aviation (NTCA), Prague, Czech Republic, 23–24 November 2020; pp. 33–38. [CrossRef]
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
37. Rose, R.L.; Puranik, T.G.; Mavris, D.N. Natural Language Processing Based Method for Clustering and Analysis of Aviation Safety Narratives. *Aerospace* **2020**, *7*, 143. [CrossRef]
38. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef] [PubMed]
39. Qaiser, S.; Ali, R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **2018**, *181*, 25–29. [CrossRef]
40. Devassy, B.M.; George, S. Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. *Forensic Sci. Int.* **2020**, *311*, 110194. [CrossRef] [PubMed]
41. Kauffmann, J.; Esders, M.; Montavon, G.; Samek, W.; Müller, K.R. From clustering to cluster explanations via neural networks. *arXiv* **2019**, arXiv:1906.07633.
42. Bureau of Transportation Statistics, U.D.o.T. On-Time Performance—Reporting Operating Carrier Flight Delays at a Glance. Available online: https://www.bts.gov/ (accessed on 16 June 2022).