aerospace

MDPI

Review

Natural Language Processing (NLP) in Aviation Safety: Systematic Review of Research and Outlook into the Future

Chuyang Yang [1,*] and Chenyu Huang [2]

[1] School of Technology and Professional Services Management, Eastern Michigan University, Ypsilanti, MI 48197, USA
[2] Aviation Institute, University of Nebraska Omaha, Omaha, NE 68182, USA; chenyuhuang@unomaha.edu
* Correspondence: cyang14@emich.edu; Tel.: +1-909-327-7241

Abstract: Advanced digital data-driven applications have evolved and significantly impacted the transportation sector in recent years. This systematic review examines natural language processing (NLP) approaches applied to aviation safety-related domains. The authors use Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) to conduct this review, and three databases (Web of Science, Scopus, and Transportation Research International Documentation) are screened. Academic articles from the period 2010–2022 are reviewed after applying two rounds of filtering criteria. The sub-domains, including aviation incident/accident reports analysis and air traffic control (ATC) communications, are investigated. The specific NLP approaches, related machine learning algorithms, additional causality models, and the corresponding performance are identified and summarized. In addition, the challenges and limitations of current NLP applications in aviation, such as ambiguity, limited training data, lack of multilingual support, are discussed. Finally, this review uncovers future opportunities to leverage NLP models to facilitate the safety and efficiency of the aviation system.

Keywords: aviation safety; natural language processing; human factors; aircraft accident investigation

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

Academic Editors: Julius Keller, Dimitrios Ziakkas and Abner Flores

Received: 22 May 2023
Revised: 16 June 2023
Accepted: 28 June 2023
Published: 30 June 2023

1. Introduction

With the rapid development of digital data storing and transmission in the aviation industry, large contextual data such as incident/accident reports [1,2], air traffic communication transcripts [3], passengers' and customers' reviews [4], and manufacturer and maintenance domain-language-based documents [5] have played substantial roles in ensuring safe and efficient operations [6]. However, the analysis of such textual-based data usually requires significant human resource investment [7]. Considerable efforts in terms of time and money are needed for work, such as voice recognition [8], text mining and identification [9,10], topic categorization [11,12], and semantic reasoning [13]. Therefore, automatic and practical analytic approaches are needed to overcome these labor-intensive challenges. Along with the increasing computation power, artificial intelligence (AI)-based NLP approaches have captured industry and academia's attention in recent years [11].

A typical procedure to develop a basic NLP model includes several steps, as detailed here. (1) Data cleaning and tokenization. They are essential processes in NLP that involve cleaning and processing textual data to enable further analysis. Data cleaning typically involves removing stop words, lowering the case, and reducing words into a single form. This process aims to standardize words with the same meaning but slightly different representations and unify them for easier grouping such as "Bag-of-Words (BoW)" [9,10]. Following data cleaning, the words are split into phrases or smaller units for more precise analysis. (2) Vectorization (or word embedding). Computers are not human beings that can understand and process strings directly; instead, they require numerical inputs. Therefore, individual strings must be mapped or labeled into real numbers to enable computer

processing [9]. (3) Model training. It is commonly completed through deep learning, specifically through recurrent neural networks (RNN) and long short-term memory (LSTM) networks [7,14]. RNN is a type of deep learning artificial neural network (ANN) with recurrent connections on the hidden state that ensures sequential information is captured. This is particularly important in understanding human languages, which involves capturing sequential information presented in the input data, i.e., the dependencies between the words in the text while making predictions. During RNN constructions, information of a node is passed back to immediate previous nodes, unlike in ANN, where it is only forwarded.

Academia has been at the forefront of developing and testing NLP-based solutions for improving aviation safety and efficiency. Researchers have developed related theoretical frameworks and algorithms that analyze pilot reports [15], air traffic control (ATC) communications [16,17], and other aviation-related textual data [14,18–20] to identify patterns and trends related to human error. These algorithms can also be used to develop predictive models that anticipate potential human errors and provide real-time feedback to pilots and air traffic controllers. The industry has also recognized the potential of NLP applications, with companies investing in new NLP tools and systems [21]. By providing insights into the root causes of human error, these tools can help aviation professionals take proactive measures to prevent future incidents/accidents [22,23]. NLP-based chatbots and virtual assistants can provide pilots and ATC real-time assistance, reducing workload and enhancing situational awareness [24]. By leveraging NLP techniques, academia can provide valuable insights into the cause of human errors in aviation and develop innovative solutions to address these issues.

Regulators are also taking notice of the potential of NLP in aviation safety, with agencies such as the International Civil Aviation Organization (ICAO), U.S. Federal Aviation Administration (FAA), National Transportation Safety Board (NTSB), and the National Aeronautics and Space Administration (NASA) exploring the use of NLP-based applications to address safety issues in the aviation industry [25–28]. There exists an urgent need for aviation stakeholders to understand the status of these state-of-the-art research and applications in the aviation industry.

A previous work indicates that there is a need for an in-depth review of NLP applications in aviation safety [6]. Therefore, the authors performed a systematic review to understand the following research questions:

1. What is the performance of NLP applications on aviation safety-related subdomains?
2. What are the challenges and limitations of these NLP applications?

This systematic literature review examines the worldwide NLP applications in analyzing incident/accident safety reports and air traffic communication data from the period 2010–2022. The specific NLP approaches, AI training methods, and model performance are summarized. The limitations and challenges of each study are discussed. A list of used terminologies is included in Table 1.

Table 1. Abbreviation list.

Acronym	Full Name	Acronym	Full Name
ADS-B	Automatic Dependent Surveillance-Broadcast	LDA	Latent Dirichlet Allocation
AM	Acoustic model	LM	Language model
ANN	Artificial neural networks	LoS	Losses of separation
ASR	Automatic speech recognition	LSA	Latent semantic analysis
ASRS	Aviation Safety Reporting System	LSTM	Long short-term memory
ATC	Air traffic control	MCNN	Multiscale CNN
BERT	Bidirectional Encoder Representations	MLP	Multilayer perceptron
BLSTM	Bidirectional long short-term memory	NASA	National Aeronautics and Space Administration
CAAC	Civil Aviation Administration of China	NB	Naïve Bayes

Table 1. *Cont.*

Acronym	Full Name	Acronym	Full Name
CER	Character error rate	NER	Name entity recognition
CFR	Code of Federal Regulations	NTSB	National Transportation Safety Board
CNN	Convolutional neural networks	OC-POS	Occurrence position
CRF	Conditional random field	PCA	Principle component analysis
CTC	Connectionist temporal classification	PM	Pronunciation model
DGAC	Directorate General for Civil Aviation	ResNet	Residual network
EASA	European Union Aviation Safety Agency	RNN	Recurrent neural network
FAA	Federal Aviation Administration	RTF	Real-time factor
FC	Fully connected layers	SMEs	Subject matter experts
GAU	Gated attention unit	SRL	Semantic role labeling
GMM	Gaussian mixture model	STM	Structural topic modeling
HFACS	Human factors analysis and classification system	SVD	Singular vector decomposition
HMI	Human–machine interface	SVM	Support vector machine
HMM	Hidden Markov models	TF-IDF	Term frequency and inverse document frequency
IATA	International Air Transport Association	t-SNE	T-distributed stochastic neighbor embedding
ICAO	International Civil Aviation Organization	UAS	Unmanned aerial system
k-NN	K-nearest neighbors algorithm	WER	Word error rate
LAN	Label attention network		

2. Materials and Methods

The study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method to identify related literature [29,30]. Three databases were searched: Scopus, Web of Science (WoS), and Transport Research International Documentation (TRID). These databases provide diverse sources to facilitate the identification of studies across multiple interdisciplinary domains. In addition, a limit of publication date from the period 2010–2022 was used as one of the applied search strategies.

2.1. Study Selection

The search keywords in this study were drawn from two categories: natural language processing (NLP) and aviation (and its sub-domains). NLP techniques were drawn from a recent systematic review of NLP methods by Pons et al. [31], Kreimeyer et al. [32], and Dreisbash et al. [33]. Aviation (and its sub-domains) are drawn from the review by Ginieis et al. [34]. The search syntax used in each search engine and database followed the expression:

(NLP techniques) AND (Aviation sub-domains),

where specific search syntax listed inside of each parenthetical phrase was selected using an "OR" Boolean operator. Table 2 presents all NLP techniques and aviation sub-domains found in the literature above and used to construct the search syntax for this PRIMSA-based study.

Table 2. Search terms for studies applying natural language processing in aviation safety.

Natural Language Processing Techniques		Aviation (and Its Sub-Domains)	
Variable	**Acronym**	**Variable**	**Acronym**
Natural language processing	NLP	Air transportation	-
Text mining	-	Air transport	-
Text classification	-	Air traffic control	ATC
Latent semantic analysis	LSA	Aerospace	-

Table 2. *Cont.*

Natural Language Processing Techniques		Aviation (and Its Sub-Domains)	
Variable	**Acronym**	**Variable**	**Acronym**
	-	Airport	-
	-	Airline	-
		Airplane	-
		Aircraft	-

An initial screening was done to assemble all potential studies based on their titles, keywords, and abstracts. RefWorks executed the removal of duplicate studies in the first round, and the authors still identified several duplicates that existed and needed to be removed manually during the second round (Figure 1). After removing duplications, two reviewers independently screened all papers' titles, keywords, and abstracts, resulting in 30 articles being selected for full-text review. The inclusion criteria for the full-text review are:

1. At least one NLP technology is applied;
2. At least one sub-domain of aviation is related;
3. Study must be related to safety;
4. Study must be published in a peer-reviewed journal.

Following the full-text review criteria, 30 articles were included in the final list.

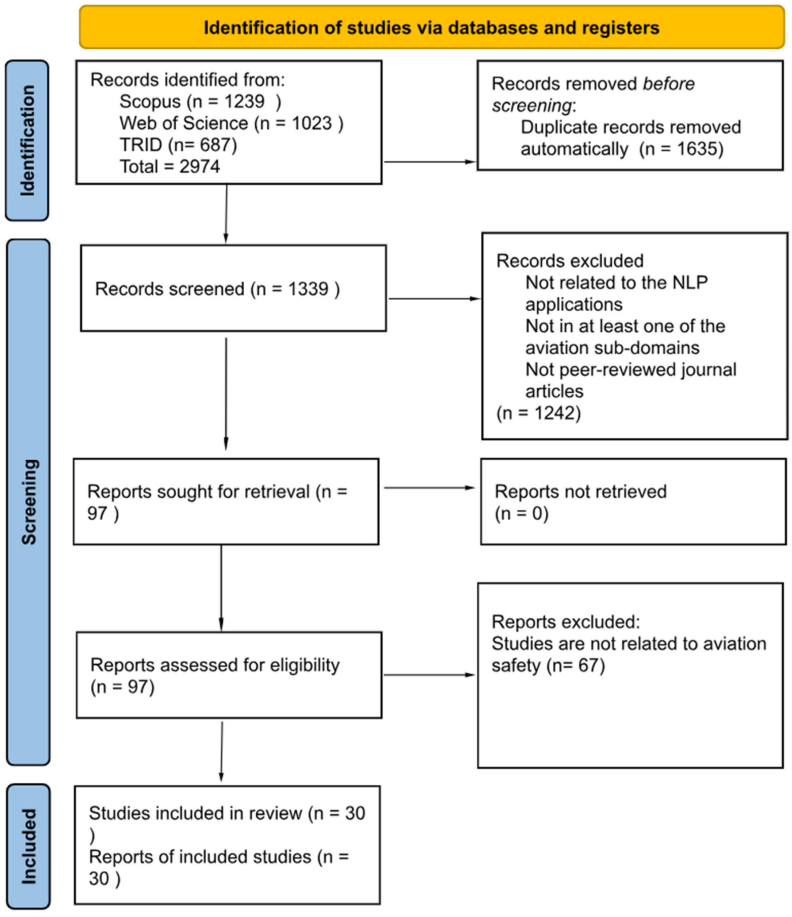

Figure 1. PRISMA process used to identify and select relevant studies. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses; TRID = Transport Research International Documentation [29,30].

2.2. Reported Factors

For this systematic review, the following information was extracted from each article:

1. Objective;
2. The target database and language;
3. Sample size;
4. Model(s), including the NLP model(s) and any additional model(s);
5. Performance of model(s).

3. Results

A total of 30 papers satisfied our inclusion criteria and were included in the final review. Based on the research questions of this study, the papers are categorized based on the sub-domains of aviation. Overall, 20 out of 30 articles relate to aviation incident/accident safety report analysis, while the rest of the included studies relate to air traffic control (ATC) communication. Therefore, the following part of this section contains a detailed synthesis of each group.

3.1. NLP Applications on Incident/Accident Reports

A total of 20 papers analyzing aviation incident/accident safety reports are included in this section. Table 3 summarizes reviewed studies' detailed information such as author, year of publication, objective of study, data source, sample size, and language of data source. A total of 13 out of 20 studies are based on the U.S. National Aeronautics and Space Administration (NASA) Aviation Safety Reporting System (ASRS) database. Two studies are based on the U.S. National Transportation Safety Board (NTSB) safety records [2,14]. In addition to 18 studies that primarily focused on English textural content, one study [9] is based on Chinese accident reports, and another study [20] used a French database.

Table 3. Summary of NLP Applications in Incident/Accident Report Analysis.

Authors, Year	Objective(s)	Data Source	Sample Size	Language
Abedin et al., 2010 [35]	Identify the potential causes of aviation incidents.	Aviation Safety Reporting System (ASRS)	1333	English
Shi et al., 2018 [13]	Identify risk factors in safety management systems.	ASRS	168,227	English
Andrzejczak et al., 2014 [15]	Identify human factors contributing to anomalies.	ASRS	127,776	English
Ahadh et al., 2021 [36]	Identify the stage of flight when an aviation accident occurs.	ASRS	37,681	English
Zhang and Mahadevan, 2019 [12]	Quantify the risk relating to the consequences of hazardous events for aviation incident risk prediction.	ASRS	64,573	English
Perboli et al., 2021 [37]	Identify human factors in the causes of aviation accidents.	Deloitte experts' reports	24	English
Jiao et al., 2022 [9]	Identify and classify causes in Chinese civil aviation incident reports.	Chinese accident reports	20,000	Chinese
Robinson, 2019 [23]	Identify the temporal trends of factors affecting safety in commercial airline operations.	ASRS	64,776	English
Tanguy et al., 2016 [20]	Identify tendencies of abnormality during a civil air flight.	ASRS and French DGAC *'s database	136,861	English and French
Dong et al., 2021 [7]	Identify the primary factor and multiple contributing factors of each incident from six most causal factors.	ASRS	181,651	English
Kuhn, 2018 [11]	Identify latent topics and trends in incident reports.	ASRS	01/2010 to 04/2015	English
Zhang et al., 2021 [14]	Automate the prognosis of aviation safety accidents.	NTSB	1673	English

Table 3. *Cont.*

Authors, Year	Objective(s)	Data Source	Sample Size	Language
Andrzejczak et al., 2012 [38]	Identify human factors of self-reported anomalies.	ASRS	Not indicated	English
Miyamoto et al., 2022 [10]	Identify inefficient operational patterns that cause flight delays and cancellations (from a safety perspective).	ASRS	4195	English
Robinson et al., 2015 [39]	Map primary causal factors in self-reported safety narratives.	ASRS	4497	English
Irwin et al., 2017 [22]	Visualize human errors for detailed analysis of text-based narratives.	ASRS	4547	English
Rose et al., 2022 [2]	Identify themes within technical datasets.	ASRS and NTSB	13,336 (ASRS) and 386 (NTSB)	English
Koteeswaran et al., 2019 [18]	Predict the topmost causes from an aircraft accident database.	Aviation Accident Dataset (AAD)	1379	English
Rose et al., 2020 [1]	Extract underlying trends from narratives.	ASRS	13,336	English
Madeira et al., 2021 [19]	Identify and classify human factors from aviation incident reports.	ASN database	1674	English

* Direction générale de l'aviation civile.

3.1.1. NLP Models

As mentioned in the Introduction section, text pre-processing is the first stage for most NLP studies, which intends to clean the raw textual data and resolve the abbreviations inconsistency issues. As presented in Table 4, common methods include tokenization [1], lower-casing [1], stop-word removal [10], stemming or lemmatization [2], and Bag-of-Words (BoW) and term frequency–inverse document frequency (TF-IDF) matrices [10]. Four studies [1,7,10,14] used the Python library Natural Language Toolkit (NLTK) to facilitate these pre-processing tasks. An unsupervised, domain-independent, and language-independent keywords extraction method, Yet Another Keyword Extraction (YAKE), was used to identify keywords by Ahadh et al. [36]. Abedin et al. [35] applied two approaches to identify causes based on a semantic lexicon, which is automatically constructed via Thelen and Riloff's Basilisk framework, and both methods outperformed the baseline system significantly. Tanguy et al. [20] used a custom rule-based normalizer, which CFH/Safety Data developed in this stage. Madeira et al. [19] tested two feature extraction models Word2Vec and Doc2Vec, concluding that they were superior to the TF-IDF.

Several topic modeling approaches have been proposed in the reviewed articles, and one popular method is Latent Dirichlet Allocation (LDA). In general, LDA is based on the assumption that each document is composed of topics, and each topic is, in turn, composed of different words. Five out of twenty studies applied LDA-based methods to model the topics. Ahadh et al. [36] proposed GuidedLDA techniques to overcome the drawback that topics generated are not human-interpretable due to the randomness of the initial assignment of words to topics. As a generalization of more commonly used LDA and correlated topic models, structural topic modeling (STM) has gained prominence in relatively recent years [11]. STM further extends the LDA framework to allow for covariates to be incorporated while selecting more prevalent topics in specific documents [2]. Kuhn [11] and Rose et al. [2] applied STM and demonstrated the feasibility of an STM-based method for classifying aviation safety narratives.

Support vector machine (SVM) is one of the most common approaches to classifying factors [12,35]. Zhang and Mahadevan [12] use a hybrid SVM and DNN model to quantify the risks associated with the consequence of each hazardous cause, yielding an average score of 81% in precision, 3% higher than the scores of SVM, and 6% higher than DNN ensemble models. The recall rate and F1 score also indicated that the hybrid model outperformed the rest of the models.

Table 4. Detail on NLP applications in incident/accident report analysis.

AUTHORS, YEAR	Models		Evaluation
	NLP Model(s)	Reasoning Model(s)	
Abedin et al., 2010 [35]	Weakly supervised lexicon learning with SVMs	Not Applicable	• F-score is 53.7%
Shi et al., 2018 [13]	Latent semantic analysis with NB, VFDT, and OBA	Not Applicable	OBA yields the best performance in all four scenarios with mean accuracies of 76.5%, 76.8%, 77.0% (human factor classifier), and 88.3%, 87.0%, 88.45, and 88.55 (aircraft classifier), respectively.
Andrzejczak et al., 2014 [15]	IBM SPSS Modeler 13: Text Analytics	HFACS	This method reveals the relationship between human factors and reported anomalies.
Ahadh et al., 2021 [36]	GuideLDA	Not Applicable	The weighted average accuracy is 77%.
Zhang and Mahadevan, 2019 [12]	A hybrid SVM and DNN model	A risk-based event outcome categorization	The hybrid model yields better performance in precision, with an average score of 0.81, which is 3% higher than the SVM and 6% higher than DNN.
Perboli et al., 2021 [37]	Word2vec and Doc2vec	SHEL	TFw2v_model has the best performance with a total precision of 88.89%.
Jiao et al., 2022 [9]	TF-IDF, Word2vec, and OC-POS withLR, L-SVM, KNN, DT, NB, SVM, RF, AdaBoost, GBoost, and XGBoost	A rule-based system to identify the related factors	XGBoost classifier and OC-POS methods have the best performance, where F1-score is above 0.90 when identifying 25 causes from the target dataset.
Robinson, 2019 [23]	LDA	Subject matter experts (SMEs)	All three SMEs were able to identify a cohesive theme from each topic.
Tanguy et al., 2016 [20]	LDA with SVMs	Not Applicable	Result: 85.96% precision for ten iterations in the DGAC corpus and 46.49% in the ASRS corpus.
Dong et al., 2022 [7]	Averaged Stochastic Gradient Descent Weight-Dropped (AWD) LSTM	Not Applicable	The proposed model yields an average accuracy of 82% on the six common factors and about 89% on the two most common factors on average.
Kuhn, 2018 [11]	LDA with STM	Not Applicable	The results need to be verified by SMEs.
Zhang et al., 2021 [14]	LSTM	Damage and injury level	The accident vs. incident model has an accuracy of 73% on validation data, while the sensitivity and specificity of the trained model are 75% and 72.14%, respectively.
Andrzejczak et al., 2012 [38]	Diffusion Maps (DM)	Not Applicable	The proposed model yields an average accuracy of 82% on the six common factors and about 89% on the two most common factors on average.
Miyamoto et al., 2022 [10]	BoW with TF-IDF	t-SNE and K-Means Clustering	The present work shows the ability to identify high-level causes and the circumstances in which delays occur.
Robinson et al., 2015 [39]	LSA with SVD	Not Applicable	An unsupervised categorization accuracy of 44% for primary cause within the existing taxonomy based on a small sample.
Irwin et al., 2017 [22]	LSA	Isometric Mapping and GIS	The present study confirms that the proposed approach is useful for reducing, interpreting, and organizing narrative data.
Rose et al., 2022 [2]	LDA with STM	Not Applicable	This study demonstrates the feasibility of an STM-based approach for classifying aviation safety narratives.
Koteeswaran et al., 2019 [18]	Improved oscillated correlation feature selection (IOCFS) withNB, SVM, ANN, k-NN, and J48	Not Applicable	k-NN yields the best performance (accuracy of 99.03%), with the value of k = 5
Rose et al., 2020 [1]	BoW with TF-IDF	t-SNE and K-Means Clustering	The method identified 10 major clusters and 31 sub-clusters.
Madeira et al., 2021 [19]	Word2Vec and Doc2Vec_Models with SVM and Bayesian optimization	HFACS	The best predictive models achieved a Micro F-score of 90%, 77.9%, and 87.5%.

3.1.2. Latent Factor Reasoning and Labeling

Among NLP applications in aviation incident/accident report analysis, latent factor reasoning and labeling are significant characteristics. In the reviewed studies, different causality models are carried out to facilitate the automatic identification and classification process [9,12,15,19,37]. Two out of twenty studies used the Human Factor and Classification System (HFACS) model to classify self-reported anomalies based on ASRS pilot reports [15,19]. Andrzejczak et al. [15] found that only 4% of the examined ASRS reports were identified as 'violation-related,' which is inconsistent with Wiegamann and Shappell's statement that 25% of aviation accidents contained some violation [40]. The anonymity and confidentiality of the reporting mechanism might help shed light on pilots' hesitation in reporting their violations [15]. A novel HFACS-ML framework proposed by Madeira et al. [19] showed that favorable predicting performance can be achieved. According to a study by Perboli et al. [37] that identifies human factors in aviation accident causes, NLP techniques are adapted to the Software–Hardware–Environment–Liveware (SHEL) standard accident causality model on field tests, yielding a precision of over 86% and a practical manner cost and time reduction of 30% for the whole investigation process. Zhang and Mahadevan [12] categorized all the events into five groups based on the level of risk and its consequence in their hybrid model, yielding an effective means of quantifying the risk relating to the consequences of hazardous events. Based on the type of incidents, Jiao et al. [9] labeled Chinese aviation incident reports into eleven categories, resulting in an F1-score above 0.90 when identifying 25 causes from the target database. Zhang et al. [14] adopted types of incidents/accidents, aircraft damage levels, and types of fatality to label events from NTSB reports, demonstrating their approach with an accuracy of 70% on validation data.

In addition, Rose et al. [1] and Miyamoto et al. [10] developed and tested a framework combining a k-means clustering and a 2D mapping with t-Distributed Stochastic Neighbor Embedding (t-SNE) was created to categorize and visualize the narratives.

3.1.3. Performance Comparison Based on Application Scenarios

The reviewed application scenarios in incident/accident reports could be categorized as the (1) identification and classification of causal factors in incident/accident reports [7,12–14,18,19,35,37,38] and the (2) identification of trends/latent topics of incidents/accidents [1,2,9–11,15,18,20,22,23,36,39].

Among the studies focused on the identification and classification of causal factors, Andrzejczak et al. [15] reported the frequencies of each identified factor that contributed to self-reported civil aviation anomalies, while the numerical performance metrics are unavailable in the study. Shi et al. [13] tested three different NLP models and concluded that the OBA model yields the best performance (in terms of accuracy) when targeting on human factors (76.5%) and aircraft classification (88.3%) in ASRS, while Dong et al. [7] concluded their proposed LSTM models yield better performance on both human factors (84.8%) and aircraft classification (85.1%) but also have a greater coverage with six topics. Regarding the LSTM application in NTSB accident report analysis, Zhang et al. [14] concluded that one (accident vs. incident based on narratives) of their models yields a rate of 77.9%, 79.7%, and 78.0% in precision, accuracy, and F1-score, respectively. Tanguy et al. [20] indicated that the performance of their proposed SVM-based model varied depending on the topics and language of corpus. For example, a precision rate of 85.96% and an F1-score of 87.59% were achieved in 'bird strikes' (DGAC corpus), while it was only 46.49% and 47.49%, respectively, in "confusion" (ASRS corpus).

Regarding the studies on identifying trends and latent topics, one study by Robinson et al. [39] pointed out that the coding results were different due to the significantly different backgrounds of SMEs. Therefore, they suggested a further investigation, on evaluating the consistency of LSA cosine values when compared to a qualitative coding process, is needed. In another study by Robinson [23], SMEs were able to independently identify themes when providing LDA-modeled topics in a structured manner. In addition, the SMEs were able to

identify the reporter's qualification as well as other factors (environment and regulatory) consistent with observed temporal trends of topic usage. Four of the five topics examined correlated significantly based on the numerical results. Kuhn [11] used STM to identify latent topics and trends from ASRS and suggested that verification by SMEs is needed for future work.

In addition, five out of twenty also visualize causal factors/topics from incident/accident reports [1,2,10,22,39]. As mentioned in Section 3.1.2, a combination of k-means clustering and t-SNE are the common tools in visualizing causal factors in incident/accident reports [10]. For instance, Miyamoto et al. [10] identified 7 major categories and a total of 23 more-detailed topics resulting in flight delays or cancellations, which indicated that ASRS narratives could be potentially leveraged to provide a safety perspective to identify causes of delays. Another visualization method is using a latent semantic analysis-based projection of narrative data into a geographic information system (GIS) [22,39].

3.2. NLP Applications on ATC

Air traffic control is another field where applications of NLP have been studied often to improve flight safety by preventing communication mistakes [6]. A total of 10 publications were identified based on the selection criteria illustrated in Section 2. A summary of the included studies is shown in Table 5.

Table 5. Summary of NLP applications on air traffic communication.

Authors, Year	Objectives	NLP Models	Data Sources	Sample Size	Language
Badrinath & Balakrishnan, 2022 [8]	ASR for ATC communication	• Mozilla's implementation of Deep Speech (RNN), • Python library—Spacy (NER), • N-gram language model	Transcripts of ATC communications from the U.S. and Europe	84 h of audio transcription	English
Zhang et al., 2022 [3]	Mandarin speech recognition for ATC	• ResNet34 • GAU • CTC	The Aishell open-source Mandarin corpus and ATC voice recordings	178 h of Aishell corpus and 67 h of ATC corpus	Chinese
Lin et al., 2021 [41]	Multilingual speech recognition in ATC systems	• MCNN • BLSTM • CTC	Raw ATC speech recorded at Chengdu, Shanghai, and Kunming Airports in China	1148 h of Chinese speech And 281 h of English speech	Chinese, English
Sun & Tang, 2021 [42]	Automated ATC communication error detection to prevent loss of separation (LoS)	• ASR to extract features (IBM Watson Speech to Text) • Communication errors characterizing (LinguaKit and Cortical.io) • Bayesian Network (BN) modeling to predict communication errors and LoS	ATC communication from simulated approach control scenarios	75 min simulation (234 clearances)	English
Jia et al., 2017 [16]	Aviation radiotelephony read-back verification	• LSTM-RNN	Experimental civil aviation radiotelephony corpus built from original ATC communication recordings and books for training	800 pairs of instruction and readback	Chinese
Wang et al., 2021 [43]	Trajectory prediction	• BiLSTM-LAN-CRF	The Mandarin-based 5000 control instructions	N/A	Chinese

Table 5. *Cont.*

Authors, Year	Objectives	NLP Models	Data Sources	Sample Size	Language
Lin et al., 2019 [44]	ATC ASR and CIU-based method to convert speech into ATC-related elements	• Two-dimensional convolution and average-pooling layers • An encoder–decoder architecture-based neural network • BLSTM-based CIU joint model	Raw ATC speech from ZUUU in China	578 h ATC speech for modeling training (481 h Chinese and 97 h English)	Chinese, English
Lin et al., 2020 [17]	Automatic Speech Recognition as a component of the ATC safety monitoring system	• CNN • RNN • BLSTM • FC	ATC communication speech recorded at civil airports in China	342 h of Chinese speech and 47 h of English speech	Chinese, English
Vukoic et al., 2021 [45]	Cognitive load estimation from speech using spectral features	• CNN	Recorded speech from human–machine interaction experiment	4.8 h of speech	English
Tan et al., 2022 [46]	Speech emotion recognition for autonomous vehicle	• Multi-model combining Spectrogram and Text (BERT)	Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set	N/A	English

Across the ten studies identified from the literature screening process, machine learning techniques were most used, mainly neural networks-based algorithms, in over half of all studies. The use of proprietary software packages (Mozilla Deep Speech, IBM Watson, LinguaKit, and Cortical.io) was also mentioned in two studies as NLP tools. Concerning the field data, 9 out of 10 studies were developed on actual ATC communication recordings, and one study collected audio information from an experimental environment. English and Chinese Mandarin were the languages observed in the selected studies relating to ATC applications; studies in other languages were excluded because of irrelevance to ATC or not being published in English. It is noticeable that studies relating to NLP applications on ATC generally include two components—methods of automatic speech recognition (ASR) and strategies of extracting information from communication for further use in aviation operations. For example, the first three studies in Table 5 are examples of studies focusing on strategies of ASR for ATC communication. In contrast, the rest of the studies consider both ASR and the additional use of ASR, such as error detection, trajectory prediction, and cognitive load estimation. The findings were summarized by the two aspects of automatic speech recognition and the additional use of data from communication.

3.2.1. Automatic Speech Recognition

Automatic speech recognition (ASR) has been widely studied in non-aviation fields. However, given the special characteristics of aviation radiotelephony procedures, such as jargon and short sentences, different language and dialects, readback, and background noise, additional studies are necessary and conducted for ASR in aviation. All identified studies include ASR as a study component for English, Chinese Mandarin, and multilingual communication in ATC, while five studies include the ASR method as a primary study objective. Descriptions of ASR in each study are summarized in Table 6.

Table 6. Detailed models on NLP applications in air traffic communication.

Authors, Year	ASR as a Primary Objective	Information Extraction	Models		Evaluation
			ASR	Information Extraction	
Badrinath & Balakrishnan, 2022 [8]	×	Call sign and runway number	• Mozilla's implementation of Deep Speech (an end-to-end speech recognition model with an RNN for acoustic model) • N-gram language model	• Name entity recognition (NER) from a Python Library-Spacy for call-sign extraction • A rule-based grammar for runway number extraction	• A word error rate (WER) of 0.17 was obtained for the ASR. • A real-time factor (RTF) of 0.3 was achieved. • NER yields an F1 score of 0.95 on the actual transcript and 0.69 on the ASR-generated transcript for call-sign extraction; rule-based grammar yields an F1 score of 1 on the actual transcript and 0.95 on ASR generated transcript for runway number extraction
Zhang et al., 2022 [3]	×		• ResNet-GAU-CTC framework (an end-to-end speech recognition model) • ResNet (a deep residual network) • GAU (a gated attention unit) • CTC (connectionist temporal classification)		The proposed model's character error rate (CER) was 11.1% on the expanded Aishell corpus and 8% on the ATC corpus.
Lin et al., 2021 [41]	×		• CNN • RNN • CTC • MCNN		A 3.95% label error rate (LER) on Chinese characters and English words
Sun & Tang, 2021 [42]		Communication features and communication errors	• ASR to extract features (IBM Watson Speech to Text)	• Communication errors Characterizing (LinguaKit and Cortical.io) • Bayesian Network (BN) modeling to predict communication errors and LoS	No evaluation of ASR; study findings indicate a high correlation between read-back errors and LoS.
Jia et al., 2017 [16]		Semantic characteristics of ATC instructions and pilot readback	• Manual transcription by professional ATC	• LSTM-RNN for semantic information extraction	The proposed semantic consistency verification scheme with K-nearest neighbors (k-NN) and random forest (RF) as classifiers is more stable and accurate (83.8% and 83%)
Wang et al., 2021 [43]		Semantic characteristics of ATC instruction		BiLSTM-LAN-CRF (a deep neural network-based algorithm) to extract the entities of ATC instruction	The percentage of wrong tags was used as metrics for performance evaluation; BiLSTM-LAN-CRF yields the best result over the other three models.
Lin et al., 2019 [44]	×	Controlling intent and parameters	• Two-dimensional convolutional operation (CNN + BLSTM + CTC-based neural network and average pooling layers) • An encoder-decoder architecture-based neural network	An RNN-based joint model for detecting the controlling intent and labeling the controlling parameters	A 4% WER with an average of 0.147 RTF was achieved.

Table 6. *Cont.*

Authors, Year	ASR as a Primary Objective	Information Extraction	Models		Evaluation
			ASR	Information Extraction	
Lin et al., 2020 [17]	×	Repetition check, flight confirmation verification, and conflict detection	• CNN + BLSTM + FC + CTC (an end-to-end model)	• DNN	The proposed model decoding with the RNN-based language model yields the best result with a 5.07% and 5.99% WER for Chinese and English.
Vukoic et al., 2021 [45]		Cognitive load		• CNN	The method yields 83.7% accuracy with CNN classifiers, which outperformed SVM and k-NN by 13.2% and 10.5%, respectively.
Tan et al., 2022 [46]		Speech emotion	• AlexNet network model	• BERT model • SoftMax classifier	The proposed method yields the best result over other methods, with a 74% weighted accuracy and 65.4% unweighted accuracy.

The selected articles show that the end-to-end speech recognition framework is a study trend for better ASR strategies over traditional HMM-based models, especially given the challenges of multilingual recognition, special use of language, and high speech rate. Among the five articles which include ASR for aviation as a primary research objective, three articles adopted an end-to-end framework [3,8,17]. Neural networks-based methods are also widely used in ASR. All ten articles used or proposed new neural networks-based methods for ASR, and in particular the combined uses of different neural networks were commonly studied. The integration of the convolutional neural networks (CNN), recurrent neural networks (RNN), connectionist temporal classification (CTC), and bidirectional long short-term memory (BLSTM) are examples. Recent ASR-related studies indicate that RNN and LSTM can effectively model the long-range temporal dependencies in the audio sequences, and CTC shows advantages in automatic alignment and fast convergence.

From the findings of the studied articles, the integrative use of those methods yielded superior results in aviation applications too. The word error rate (WER) and character error rate (CER) were commonly used to evaluate the ASR accuracy in five of the ten papers. The real-time factor (RTF) is another parameter for evaluating ASR efficiency. From the findings of studies, the WER/CER of ASR ranges from 4% to 17%. Lin et al. [44] achieved the lowest WER of 4% and the highest RTF of 0.147 s of an ASR strategy for a real-time controlling dynamic in air traffic systems. Two-dimensional convolutional operation (Conv2D), average-pooling layers (APL), and BLSTM were adopted in their work [44].

3.2.2. Operational Information Extraction

In the overall process of applying NLP in ATC, ASR serves as the critical first step to converting aviation communication speech into textual data. Operational information extraction from communication text creates great value for various practical purposes such as communication error detection, pilot readback check, traffic conflict detection, and cognitive load estimation. Eight of the ten studies include information extraction from different perspectives, such as aircraft call signs, runway numbers, ATC controlling intent, and semantics of speech. Descriptions of information extraction in each study are summarized in Table 6.

For operational information extraction for ATC, machine learning-based methods dominate the selected studies, particularly neural networks-based methods, which were widely used in six out of eight relevant papers [16,17,43–46].

The combined uses of LSTM, RNN, CNN, and DNN were explored in different studies; the bidirectional encoder representation from transformers (BERT) model was adopted by one study to check speech emotion [46]. In addition, name entity recognition, rule-

based grammar, and Bayesian networks were also used [8,42]. In terms of evaluation, no common method is observed from selected studies, potentially because of the diversity of applications. However, the F score was used to assess the accuracy of aircraft call-sign extraction and runway number extraction [8]. An important observation from this study [8] is that the F-scores decrease because of the errors from ASR, which reemphasized the importance of ASR accuracy when integrating ASR and information extraction in the same system. Other evaluation methods adopted comparisons with different classifiers, such as support vector machine (SVM), k-nearest neighbors (k-NN), and random forest (RF) [16,45].

4. Discussion

4.1. Challenges and Limitations

While NLP has many potential applications in aviation, several challenges and limitations also need to be addressed. Based on this systematic review, the authors list NLP approaches' main challenges and limitations when applying them to text analysis in aviation safety-related domains.

4.1.1. Ambiguity and Context

NLP struggles to understand the context and meaning of words and phrases, especially in aviation-specific jargon. For example, the term "runway" can refer to both the physical runway and the takeoff/landing procedures associated with it.

Reviewing NLP-based studies on incident/accident report analysis, Rose et al. [2] highlighted that there are more easily identifiable topic labels in NTSB reports, which can be clearly differentiated from another, than in ASRS reports. Hence, a structural topic modeling (STM) approach performs better when applied to NTSB reports [2]. NTSB aviation accident reports have non-standard abbreviations that add noise to the model-learning process [3,14]. Similar issues also exist in non-English databases, such as a study on Chinese civil aviation incident reports from the period 2007–2021 by Jiao et al. [9] indicates that the experiment was interfered with due to invalid information, such as inconsistent writing norms and standards from different airlines. One way to improve the training model performance is to develop a specialized corpus for handling aviation accident reports only [3,14].

Automatic speech recognition (ASR) is one of the essential elements for NLP applications in aviation communication. It attempts to transcribe voice information into textual data for relevant analyses and knowledge discovery. However, aviation communication differs from many other daily dialogues with its uniqueness and a variety of specificities; ambiguity, high speech rate, and jargon are examples of those challenges [3,44,47]. Relevant studies explored support vector machines (SVM)-based methods, conditional random fields, and maximum entropy Markov models-based algorithms to extract ATC controlling intent and parameters [48–50]. Cordoba et al. [51] also proposed an ASR system for cross-task and adapting speaking features of air traffic controllers. This systematic review of recent studies also finds that many existing well-developed ASR models for non-aviation fields fail to solve those issues with acceptable performance; ASR and NLP for aviation communication should be specifically investigated. Recent studies show that deep learning-based models, such as LSTM-based, CNN-based, and RNN-based architectures, appear superior in modeling and understanding aviation language and could be the direction for future studies [17,41,44,52–54].

4.1.2. Multilingual Support

Aviation involves passengers and crew from different parts of the world speaking diverse languages. NLP must support multiple languages and dialects to be helpful in this industry.

A study [9] on a Chinese dataset used different preprocessing steps and text analysis since word embedding and TF-IDF cannot capture important information from long texts [9]. In conclusion, Jiao et al. [9] indicated that a rule-based approach coupled with human intervention is a powerful tool to explore in future studies.

Multilingualism appears more challenging in ASR for air traffic communication [41,43,44]. Though international flights are required to maintain English communication, the diverse accents of English communication create significant challenges for developing an effective ASR for air traffic communication. Domestic flights in many non-English speaking countries still use the corresponding native language. From the literature review, we found that English and Chinese were two primary languages studied separately or for multilingual ASR strategies, and deep learning-based end-to-end ASR shows good performance for English and Chinese [3,8,16,17,41,43,44]. To develop a generic framework for different languages, Lin et al. [17,41,44] proposed a multilingual framework to integrate ASR with a subsystem-controlling intent inference (CII) to recognize and organize textual information into a predefined data structure for further uses. Provided the training data in different languages, this framework is expected to be applied in the corresponding environment.

Although recent studies significantly improve the accuracy of new ASR models by innovatively integrating machine learning techniques into end-to-end ASR models, a dedicated ASR model is still needed for each language; few studies are observed to explore a single ASR model that is capable of interpreting multiple languages for air traffic communication. Further studies on this aspect may promote better field applications of such technology in a multilingual aviation operation environment.

4.1.3. Noise and Background Sounds

Background noise and distractions are common in aviation communication. This can make it difficult for NLP systems to interpret spoken commands or queries accurately.

Background noise is a particular factor affecting the accuracy of ASR in air traffic communications. The VHF (very high frequency) radio transmission commonly contains noise caused by static, radio frequency interference, or thermal noise; inferior speech quality is considered one of the main challenges for implementing ASR in air traffic control (ATC) operations [3,44]. Reviewing the previous studies, neural network-based strategies were explored to address the noise problem; for instance, the CNNs were used to improve the encoder architecture to reduce the impact of background noise [55], and the average pooling layer was adopted to filter the noise based on the analysis of ATC speech [17,41,44].

However, challenges from diverse and complicated noise in different application scenes still exist and should be considered and addressed in future studies. To develop an accurate ASR strategy for field applications, future efforts must be taken to eliminate the impact of background noise.

4.1.4. Limited Training Data

Training NLP models requires large amounts of data, but the aviation industry may have limited data available due to privacy concerns and safety regulations.

Thirteen out of twenty studies in the first group only investigated NLP applications on ASRS, given the accessibility and richness of ASRS data [6]. However, the limitations of ASRS data should not be ignored when compared to NTSB aviation accident reports:

1. It is usually considered an incomprehensive report regarding the whole process of an incident/accident [12] and is considered less formal than the NTSB reports, including official investigation results;
2. Objectiveness is hindered due to the nature (anonymity and confidentiality) of the reporting procedure [15].

It would be beneficial to leverage more comprehensive reports on severe accidents in NTSB to shed light on pattern identification and risk mitigation solutions [12].

The reviewed NLP-based studies demonstrate the potential to classify the causes of incident/accident textual data. However, since evaluating the subject matter experts (SMEs) is essential in training models, human resource limitations lead to the training dataset's limited sample size [7,14,19,22,35,38]. One solution to overcome the limited training dataset is to propose a better deep-learning architecture that requires fewer data or a data-augmentation model [7]. Several augmentation models are proposed, such as

Google N-gram [35]. Perboli et al. [37] also proposed to leverage the increasing knowledge base where every new report is processed as a more structured training dataset.

The accuracy of ASR depends on the amount of available labeled data for model training. However, the amount of publicly available transcribed air traffic communication voice data is limited compared with other regular dialogues [3,8,44]. Reviewing previous studies, a cross-lingual knowledge transfer learning method and a semi-hidden layer cross-lingual DNN architecture were proposed to overcome the small sample caused problems [56], and a combination of unsupervised pre-training and supervised transfer learning was also proposed [57]. A combination of unsupervised pre-training and supervised transfer learning was also proposed [57]. In addition, semi-supervised learning [58], context-aware ASR models [59], and integrating standard ATC phraseology were explored to overcome those limitations [60], but the overall performance in terms of word error rate needs more significant improvements in the domain of ATC communication. In general, end-to-end ASR architectures with deep neural networks were studied. They seemed promising to improve ASR accuracy under the impact of limited training data, which might be further explored. Above all, the development of a large, publicly available, and multilingual annotated air traffic communication data repository seems necessary for further studies.

4.1.5. Safety-Critical Systems

ASR shows great potential in developing automatic safety-critical systems in aviation operations, such as the detection of communication incurred by human errors, deviations from voice instructions, and operator status monitoring [61]. Six out of ten selected studies explored ASR and NLP for safety support from different perspectives; automated ATC communication error detection to prevent loss of separation, aviation radiotelephony readback verification, operator cognitive functions load estimation, and speech emotion recognition are examples [16,17,42,43,45,46]. Many other safety-supporting functions could be explored in further studies by leveraging the fast-developing ASR/NLP strategies. However, many challenges still need to be overcome for safety-supporting applications in aviation, such as the accuracy of ASR and information extraction, processing speed, and multilingual recognition [17]. The use of fast-developing deep learning-based techniques seems promising for practical solutions to those challenges.

The current NLP techniques might not entirely replace safety-critical systems that require high accuracy and reliability, such as air traffic control or collision avoidance systems. However, NLP appears promising and effective to be utilized as a support tool for decision-makers [62].

4.1.6. Real-Time Processing

Decisions must be made quickly in air traffic management and control scenarios. NLP systems might need help to process information fast enough to keep up with the pace of aviation operations.

ASR could be time sensitive for certain types of applications in air traffic communications, for example, real-time air traffic control safety monitoring, air traffic communication feedback verification, and air traffic conflict detection. Those tasks require real-time processing capability for ASR solutions with high accuracy. For the popular neural networks-based ASR models in recent studies, one of the primary factors determining the computational time is the number of hidden layers and neurons in each layer of a large neural network model. In contrast, a large neural network model usually yields better accuracy. The findings from this literature review project indicate that a few studies have practically achieved the level of real-time ASR and data processing with the fastest real-time factor (RTF) of 0.147 s [8,44]. For future research and applications, the trade-off between processing time and accuracy might be of interest.

4.1.7. Cost

Developing and implementing NLP systems in the aviation industry can be expensive, which may limit the adoption of these technologies.

The reviewed studies mainly focused on forecasting the final adverse event outcomes, while the intermediate event propagation process is ignored [14]. As a potential improvement, a more sophisticated end-to-end model might need to investigate the evolution process of aviation incidents/accidents [14].

4.2. Future Opportunities

Near the end of this review, several new NLP applications in the aviation field have caught researchers' attention, which are discussed in this section.

Several new studies were published to extend NLP applications in aviation by adopting the Bidirectional Encoder Representations from Transformers, also known as BERT. BERT is developed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers [63]. BERT is a "transformer-based" large language model, which is openly available but can be adapted to domain-specific tasks such as aviation. In 2020, Kierszbaum and Lapasset [64] first employed BERT to extract information from ASRS to answer the following question, "When did the incident happen?", and they yielded roughly 70% correct answers. More applications such as aeroBERT-Classifer [65], Aviation-BERT [62,66], and SafeAeroBERT [67] have been proposed and tested in up-to-date studies regarding the analysis of aviation text corpora. Based on the findings of those new studies, BERT-based techniques can generally help yield better accuracy in aviation natural language data analysis.

As a trend in the NLP field, ChatGPT/GPT4 has succeeded in several domains as an up-to-date AI language model. GPT also has the potential to improve safety and efficiency in aviation safety-related domains in the following ways: First, similarly to other NLP tools, GPT can analyze and understand human language in aviation-related texts such as incident/accident reports, safety notices, technical manuals, and pilot reports to identify patterns and trends in human error, providing valuable insights into the root causes of aviation incidents/accidents. A recent study indicated that NLP has successfully identified and classified NOTAM reports [68]. Second, GPT can be used to develop predictive models that anticipate potential human errors and provide real-time feedback to pilots and air traffic controllers. Third, GPT can automatically identify subtle changes in language or communication patterns that may imply an increased risk of human error, allowing aviation professionals to take proactive measures to prevent accidents. Finally, GPT can develop chatbots and virtual assistants to provide pilots and ATC real-time assistance, reducing workload and enhancing situational awareness. These chatbots and virtual assistants can help aviation professionals make informed decisions and respond quickly to changing conditions, ultimately improving the safety and efficiency of operations.

On the other hand, there is still plenty of work to be done before such NLP applications can be used in aviation. As this review discussed earlier in this section, a specific domain-based Corpus is required to train the language model better since the aviation industry uses many professional acronyms and abbreviations. In addition, redundancy is needed when applying NLP applications to safety-critical systems such as ATC communications. Such applications should support the human-in-the-loop (HITL) decision-making process rather than entirely taking over human tasks. Addressing these challenges and limitations is essential for successfully implementing NLP in aviation safety-related domains. With continued research and development in AI software and hardware, NLP can improve safety and efficiency in the aviation industry. Future directions include further investigation into how NLP can be applied to the operations of other aviation stakeholders, such as airports, airlines, and manufacturers.

5. Conclusions

From academia to industry to regulators, there is a growing recognition of the value of NLP-based solutions in enhancing situational awareness, reducing workload, and improving decision capabilities in aviation. This systematic review took steps toward bridging the research gap related to synthesizing knowledge on how NLP techniques impact aviation safety domains. It presents a systematic review of NLP applications in aviation safety-related domains. Specifically, the reviewed studies were grouped into incident/accident safety report analysis and ATC communications. Important factors such as the NLP model and its corresponding performance, targeted database, and applied language were summarized. Finally, the authors discussed challenges and limitations such as ambiguity, lack of multilingual support, limited training data, etc. Finally, the corresponding opportunities and future directions were included. In summary, as AI technology continues to advance, it is expected that NLP will play an increasingly substantial role in improving the safety and efficiency of the aviation system.

Author Contributions: Conceptualization, C.Y. and C.H.; methodology, C.Y.; software, C.H.; formal analysis, C.Y. and C.H.; investigation, C.Y. and C.H.; resources, C.Y. and C.H.; data curation, C.Y. and C.H.; writing—original draft preparation, C.Y. and C.H.; writing—review and editing, C.Y. and C.H.; visualization, C.Y.; supervision, C.Y.; project administration, C.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This review was funded by Eastern Michigan University's Provost Research Support Award (003037). The APC was funded by Eastern Michigan University's Faculty Open Access Publishing Fund (003387), administered by the Associate Provost and Associate Vice President for Graduate Studies and Research, with assistance from the EMU library.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. 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]
2. Rose, R.L.; Puranik, T.G.; Mavris, D.N.; Rao, A.H. Application of structural topic modeling to aviation safety data. *Reliab. Eng. Syst. Saf.* **2022**, *224*, 108522. [CrossRef]
3. Zhang, S.Y.; Kong, J.G.; Chen, C.; Li, Y.B.; Liang, H.J. Speech GAU: A single head attention for mandarin speech recognition for air traffic control. *Aerospace* **2022**, *9*, 395. [CrossRef]
4. Xu, X.; Liu, W.; Gursoy, D. The impacts of service failure and recovery efforts on airline customers' emotions and satisfaction. *J. Travel Res.* **2019**, *58*, 1034–1051. [CrossRef]
5. Falessi, D.; Cantone, G.; Canfora, G. Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques. *IEEE Trans. Softw. Eng.* **2011**, *39*, 18–44. [CrossRef]
6. Amin, N.; Yother, T.; Johnson, M.; Rayz, J. Exploration of Natural Language Processing (NLP) Applications in Aviation. *Coll. Aviat. Rev. Int.* **2022**, *40*, 203–216.
7. Dong, T.; Yang, Q.; Ebadi, N.; Luo, X.R.; Rad, P. Identifying incident causal factors to improve aviation transportation safety: Proposing a deep learning approach. *J. Adv. Transp.* **2021**, *2021*, 5540046. [CrossRef]
8. Badrinath, S.; Balakrishnan, H. Automatic Speech Recognition for Air Traffic Control Communications. *Transp. Res. Rec.* **2022**, *2676*, 798–810. [CrossRef]
9. Jiao, Y.; Dong, J.; Han, J.; Sun, H. Classification and causes identification of Chinese civil aviation incident reports. *Appl. Sci.* **2022**, *12*, 10765. [CrossRef]
10. Miyamoto, A.; Bendarkar, M.V.; Mavris, D.N. Natural Language Processing of Aviation Safety Reports to Identify Inefficient Operational Patterns. *Aerospace* **2022**, *9*, 450. [CrossRef]
11. Kuhn, K.D. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 105–122. [CrossRef]
12. Zhang, X.G.; Mahadevan, S. Ensemble machine learning models for aviation incident risk prediction. *Decis. Support Syst.* **2019**, *116*, 48–63. [CrossRef]

13. Shi, D.H.; Guan, J.; Zurada, J.; Manikas, A. A data-mining approach to identification of risk factors in safety management systems. *J. Manag. Inf. Syst.* **2017**, *34*, 1054–1081. [CrossRef]
14. Zhang, X.G.; Srinivasan, P.; Mahadevan, S. Sequential deep learning from NTSB reports for aviation safety prognosis. *Saf. Sci.* **2021**, *142*, 105390. [CrossRef]
15. Andrzejczak, C.; Karwowski, W.; Thompson, W. The identification of factors contributing to self-reported anomalies in civil aviation. *Int. J. Occup. Saf. Ergon.* **2014**, *20*, 3–18. [CrossRef]
16. Jia, G.M.; Lu, Y.J.; Lu, W.B.; Shi, Y.H.; Yang, J.F. Verification method for Chinese aviation radiotelephony readbacks based on LSTM-RNN. *Electron. Lett.* **2017**, *53*, 401–403. [CrossRef]
17. Lin, Y.; Deng, L.J.; Chen, Z.M.; Wu, X.P.; Zhang, J.W.; Yang, B. A real-time ATC safety monitoring framework using a deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4572–4581. [CrossRef]
18. Koteeswaran, S.; Malarvizhi, N.; Kannan, E.; Sasikala, S.; Geetha, S. Data mining application on aviation accident data for predicting topmost causes for accidents. *Clust. Comput.* **2017**, *22*, 11379–11399. [CrossRef]
19. 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]
20. 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]
21. Carvalho, T. Natural Language Processing in Airline Maintenance Operations. In Proceedings of the Presented at Aerospace IT 2022, Chicago, IL, USA, 5 October 2022.
22. Irwin, W.J.; Robinson, S.D.; Belt, S.M. Visualization of Large-Scale Narrative Data Describing Human Error. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2017**, *59*, 520–534. [CrossRef] [PubMed]
23. Robinson, S.D. Temporal topic modeling applied to aviation safety reports: A subject matter expert review. *Saf. Sci.* **2019**, *116*, 275–286. [CrossRef]
24. OpenAI. Available online: https://openai.com/ (accessed on 14 March 2023).
25. Groff, L. Applying Natural Language Processing Tools to Occurrence Reports. ICAO. Available online: https://www.icao.int/safety/iStars/Documents/IUG%20Meeting%201/Presentations/Applying%20Natural%20Language%20Processing%20Tools%20to%20Occurrence%20Reports%20-%20Loren%20Groff.pdf (accessed on 7 April 2023).
26. ICAO. Available online: https://www.icao.int/safety/Pages/Artificial-Intelligence-(AI).aspx (accessed on 7 April 2023).
27. Kopald, H. Automatic Speech Recognition and Understanding of ATC Voice Communications. In Proceedings of the Air Transportation Information Exchange Conference (ATIEC) 2021, Virtual Event, 16 September 2021.
28. NTSB. Available online: https://www.ntsb.gov/safety/safety-studies/Documents/SRR2201.pdf (accessed on 12 February 2023).
29. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160. [CrossRef] [PubMed]
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Int. J. Surg.* **2021**, *88*, 105906. [CrossRef]
31. Pons, E.; Braun, L.M.M.; Hunink, M.G.M.; Kors, J.A. Natural Language Processing in Radiology: A Systematic Review. *Radiology* **2016**, *279*, 329–343. [CrossRef]
32. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [CrossRef]
33. Dreisbach, C.; Koleck, T.A.; Bourne, P.E.; Bakken, S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int. J. Med. Inform.* **2019**, *125*, 37–46. [CrossRef]
34. Ginieis, M.; Sánchez-Rebull, M.V.; Campa-Planas, F. The academic journal literature on air transport: Analysis using systematic literature review methodology. *J. Air Transp. Manag.* **2012**, *19*, 31–35. [CrossRef]
35. Abedin, M.A.U.; Ng, V.; Khan, L. Cause identification from aviation safety incident reports via weakly supervised semantic lexicon construction. *J. Artif. Intell. Res.* **2010**, *38*, 569–631. [CrossRef]
36. Ahadh, A.; Binish, G.V.; Srinivasan, R. Text mining of accident reports using semi-supervised keyword extraction and topic modeling. *Process. Saf. Environ. Prot.* **2021**, *155*, 455–465. [CrossRef]
37. Perboli, G.; Gajetti, M.; Fedorov, S.; Giudice, S.L. Natural Language Processing for the identification of Human factors in aviation accidents causes: An application to the SHEL methodology. *Expert Syst. Appl.* **2021**, *186*, 115694. [CrossRef]
38. Andrzejczak, C.; Karwowski, W.; Mikusinski, P. Application of diffusion maps to identify human factors of self-reported anomalies in aviation. *Work* **2012**, *41*, 188–197. [CrossRef] [PubMed]
39. Robinson, S.D.; Irwin, W.J.; Kelly, T.K.; Wu, X.O. Application of machine learning to mapping primary causal factors in self-reported safety narratives. *Saf. Sci.* **2015**, *75*, 118–129. [CrossRef]
40. Wiegmann, D.A.; Shappell, S.A. Human error analysis of commercial aviation accidents: Application of the human factors analysis and classification system (HFACS). *Aviat. Space Envion. Med.* **2001**, *72*, 1006–1016.
41. Lin, Y.; Guo, D.Y.; Zhang, J.W.; Chen, Z.M.; Yang, B. A unified framework for multilingual speech recognition in air traffic control systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 3608–3620. [CrossRef] [PubMed]

42. Sun, Z.; Tang, P. Automatic communication error detection using speech recognition and linguistic analysis for proactive control of loss of separation. *Transp. Res. Rec. J. Transp. Res. Board* **2021**, *2675*, 1–12. [CrossRef]
43. Wang, X.; Mao, Y.; Wu, X.Y.; Xu, Q.C.; Jiang, W.Y.; Yin, S.W. An ATC instruction processing-based trajectory prediction algorithm designing. *Neural Comput. Appl.* **2021**, 1–14. [CrossRef]
44. Lin, Y.; Tan, X.; Yang, B.; Yang, K.; Zhang, J.; Yu, J. Real-time controlling dynamics sensing in air traffic system. *Sensors* **2019**, *19*, 679. [CrossRef]
45. Vukovic, M.; Stolar, M.; Lech, M. Cognitive Load Estimation From Speech Commands to Simulated Aircraft. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1011–1022. [CrossRef]
46. Tan, L.; Yu, K.; Lin, L.; Cheng, X.; Srivastava, G.; Lin, J.C.-W.; Wei, W. Speech Emotion Recognition Enhanced Traffic Efficiency Solution for Autonomous Vehicles in a 5G-Enabled Space–Air–Ground Integrated Intelligent Transportation System. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 2830–2842. [CrossRef]
47. Biadsy, F. Automatic Dialect and Accent Recognition and its Application to Speech Recognition. Ph.D. Thesis, Columbia University, New York, NY, USA, 2011.
48. Haffner, P.; Tur, G.; Wright, J.H. Optimizing SVMs for complex call classification. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03), Hong Kong, China, 6–10 April 2003; pp. I-632–I-635.
49. Yao, K.; Peng, B.; Zweig, G.; Yu, D.; Li, X.; Gao, F. Recurrent conditional random field for language understanding. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4077–4081.
50. Bonnisseau, J.-M.; Lachiri, O. On the objective of firms under uncertainty with stock markets. *J. Math. Econ.* **2004**, *40*, 493–513. [CrossRef]
51. Cordoba, R.D.; Ferreiros, J.; San-Segundo, R.; Macias-Guarasa, J.; Montero, J.M.; Fernandez, F.; D'Haro, L.F.; Pardo, J.M. Air traffic control speech recognition system cross-task and speaker adaptation. *IEEE Aerosp. Electron. Syst. Mag.* **2006**, *21*, 12–17. [CrossRef]
52. Yao, K.; Peng, B.; Zhang, Y.; Yu, D.; Zweig, G.; Shi, Y. Spoken language understanding using long short-term memory neural networks. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014.
53. Xu, P.; Sarikaya, R. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 78–83.
54. Guo, D.; Tur, G.; Yih, W.; Zweig, G. Joint semantic utterance classification and slot filling with recursive neural networks. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 554–559.
55. Zhou, K.; Yang, Q.; Sun, X.S.; Liu, S.H.; Lu, J.J. Improved CTC-Attention Based End-to-End Speech Recognition on Air Traffic Control. In Proceedings of the 9th International Conference on Intelligence Science and Big Data Engineering (IScIDE), Nanjing, China, 17–20 October 2019.
56. Wang, J.; Liu, S.H.; Yang, Q. Transfer learning for air traffic control LVCSR system. In Proceedings of the 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 10 December 2017.
57. Lin, Y.; Li, Q.; Yang, B. Improving speech recognition models with small samples for air traffic control systems. *Neurocomputing* **2021**, *445*, 287–297. [CrossRef]
58. Srinivasamurthy, A.; Motlicek, P.; Himawan, I.; Szaszák, G.; Oualil, Y.; Helmke, H. Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20 August 2017.
59. Oualil, Y.; Klakow, D.; Szasza'k, G.; Srinivasamurthy, A.; Helmke, H.; Motlicek, P. A context-aware speech recognition and understanding system for air traffic control domain. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 404–408.
60. Nguyen, V.N. Using Linguistic Knowledge for Improving Automatic Speech Recognition Accuracy in Air Traffic Control. Master's Thesis, Østfold University College, Halden, Norway, 2016.
61. Kopald, H.D.; Chanen, A.; Chen, S.; Smith, E.C.; Tarakan, R.M. Applying automatic speech recognition technology to Air Traffic Management. In Proceedings of the 2013 IEEE/AIAA 32nd Digital Avionics Systems Conference (DASC), East Syracuse, NY, USA, 5–10 October 2013.
62. Xiao, J.; Chennakesavan, A.; Chandra, C.; Bendarkar, M.V.; Kirby, M.; Mavris, D.N. BERT for aviation text classification. In Proceedings of the AIAA Aviation 2023 Forum, San Diego, CA, USA, 12–16 June 2023.
63. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
64. Kierszbaum, S.; Lapasset, L. Applying distilled BERT for question answering on ASRS reports. In Proceedings of the 2020 IEEE New Trends in Civil Aviation (NTCA), Prague, Czech Republic, 23–24 November 2020; pp. 33–38.
65. Andrade, S.R.; Walsh, H.S. SafeAeroBERT: Towards a safety-informed aerospace-specific language model. In Proceedings of the AIAA AVIATION 2023 Forum, San Diego, CA, USA, 12–16 June 2023.
66. Chandra, C.; Jing, X.; Bendarkar, M.V.; Sawant, K.; Elias, L.; Kirby, M.; Mavris, D.N. Aviation-BERT: A preliminary aviation-specific natural language model. In Proceedings of the AIAA AVIATION 2023 Forum, San Diego, CA, USA, 12–16 June 2023.

67. Tikayat Ray, A.; Cole, B.F.; Pinon Fischer, O.J.; White, R.T.; Mavris, D.N. aeroBERT-Classifier: Classification of Aerospace Requirements Using BERT. *Aerospace* **2023**, *10*, 279. [CrossRef]
68. Maynard, P.; Clarke, S.S.; Almache, J.; Kumar, S.; Rajkumar, S.; Kemp, A.; Pai, R. Natural Language Processing (NLP) Techniques for Air Traffic Management Planning. In Proceedings of the AIAA Aviation 2021 Forum, Virtual Event, 2–6 August 2021.