applied sciences

MDPI

Article

Classification and Causes Identification of Chinese Civil Aviation Incident Reports

Yang Jiao [1], Jintao Dong [2,*], Jingru Han [1] and Huabo Sun [1,*]

[1] Engineering and Technical Research Center of Civil Aviation Safety Analysis and Prevention, China Academy of Civil Aviation Science and Technology, Beijing 100028, China
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
* Correspondence: 2020202130043@whu.edu.cn (J.D.); sunhb@mail.castc.org.cn (H.S.)

Abstract: Safety is a primary concern for the civil aviation industry. Airlines record high-frequency but potentially low-severity unsafe events, i.e., incidents, in their reports. Over the past few decades, civil aviation security practitioners have made efforts to analyze these issues. The information in incident reports is valuable for risk analysis. However, incident reports were inefficiently utilized due to incoherence, large volume, and poor structure. In this study, we proposed a technical scheme to intelligently classify and extract risk factors from Chinese civil aviation incident reports. Firstly, we adopted machine learning classifiers and vectorization strategies to classify incident reports into 11 categories. Grid search was used to adjust the parameters of the classifier. In the preliminary experiment, the combination of the extreme gradient boosting (XGBoost) classifier and the occurrence position (OC-POS) vectorization strategy outperformed with an 0.85 weighted F1-score. In addition, we designed a rule-based system to identify the factors related to the occurrence of incidents from 25 empirical causes, which included equipment, human, environment, and organizational causes. For cause identification, we used rules obtained through manual analysis with keywords and discourse. F1-score above 0.90 was obtained on the test set using the causes identification model derived from the training set. The proposed system permits insights into unsafe factors in aviation incidents and prevents reoccurrence. Future works can proceed on this study, such as exploring the causal relationship between causes and incidents.

Keywords: incident reports; text mining; natural language processing; machine learning; civil aviation

check for updates

Citation: Jiao, Y.; Dong, J.; Han, J.; Sun, H. Classification and Causes Identification of Chinese Civil Aviation Incident Reports. *Appl. Sci.* **2022**, *12*, 10765. https://doi.org/10.3390/app122110765

Academic Editor: Rafael Valencia-Garcia

Received: 2 September 2022
Accepted: 21 October 2022
Published: 24 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Safety is a primary concern for the civil aviation industry [1]. Unsafe events during a flight pose a significant threat to flight safety and can result in damage to the aircraft or even injury to personnel. Heinrich's law suggests that the occurrence of serious accidents may be closely related to the occurrence of near-misses and minor accidents [2]. Thus, the causes of incidents deserve comprehensive analysis. For each incident, the responsible airline is obliged to record all the details of its occurrence in the reports. In the investigation, report writers utilize data from multiple sources, including conversations in the cockpit, Quick Access Recorder (QAR) data [3], and interviews with the crew. Currently, many investigation reports are stored and not further utilized. Therefore, effective and rational analysis of incident reports can be of great help in identifying risk factors [4]. In practice, incident reports are usually descriptive texts and are unstructured or semi-structured. Due to the large volume, perhaps tens of thousands, of incident reports, traditional manual analysis is far from adequate [3]. Developing a feasible and efficient technical framework for filling this gap is the primary motivation for this study.

Researchers have applied text mining techniques to areas such as public health [5], chemical engineering [6] and construction [7–10]. There are quite few studies related to the aviation industry, especially in Chinese. Tanguy et al. used the support vector

machine (SVM) technique for the automatic classification of aviation safety reports [11]. Karanikas and Nederend proposed a framework to classify aviation events for controllability [12]. In addition, Tanguy et al. also used the LDA model for topic modelling; Kuhn developed a structured thematic model for aviation safety reports to identify the underlying factors that might trigger incidents [13]. Li et al. applied the human factors analysis and classification system (HFACS) to investigate human errors in aviation accidents [14]. In general, most of the current research has focused on classification, but other applications remain undeveloped. Identifying potential risk factors from plenty of reports is another challenge to be addressed [15]. In this study, we used Chinese NLP techniques to extract key information from Chinese civil aviation incident reports. In practice, as a very different language from Western languages, the NLP of Chinese focuses on processing characters, morphemes, and words [16]. On this basis, we proposed a technical framework for intelligent classification and cause identification with incident reports.

This paper organizes as follows: in Section 2, we introduce the dataset and related works. In Section 3, we describe the method used to report the classification and present the results. In Section 4, we illustrate the principles of rule-based systems for cause identification and conduct experiments. In Section 5, we summarize this study and anticipate future works.

2. Data and Methodology

2.1. Data

Incident-reporting policies dictate that airports and airlines must record detailed information if any abnormal event occurs and forward these reports to regulatory agencies. The China Academy of Civil Aviation Science and Technology (CASTC) maintains almost all of the accident reports for China. In this study, our experiment was conducted on CASTC's 2007–2021 China Civil Aviation Incident Report Database, which contains approximately 20,000 incident reports that occurred in mainland China [17]. Each incident report briefly documents the course of the incident, including causes and results, which helps identify the source of the hazard and discern what prevented the incident from becoming an accident [18].

Complete reports are commonly composed of the following three types of information:

(1) Title: A summary of the incident, including the time of occurrence, aircraft type, and incident type;
(2) Narrative: A detailed incident description, including all the details of the incident and the losses caused by it;
(3) Analysis: The concluding results and analysis of the post-incident survey, including the liability and severity rating of the incident.

All incident reports of this study were written in Chinese, with an average word count of dozens to hundreds, while most have only one narrative. In Figure 1, we present an example of a complete aviation accident report consisting of a title, narrative, and analysis. In this example, the title documents the type of incident (incorrect altitude), aircraft type (A320), and time (2017). The narrative documents the pilot's distraction resulting in incorrect altitude settings. The narrative and analysis documented a lack of cross-checking, operational violations, inclement weather (thunderstorms), and poor communication between pilots and traffic controllers that contributed to the incident.

Title:

2017 年 xx 月 xx 日 A320 飞机飞错高度

Narrative:

2017 年 xx 月 xx 日，A320 飞机执行 xx-xx 航班，xx: xx 于合肥起飞，xx: xx 左右该航班进入 x x 管制区域，此时飞机高度 7380 米，xx 管制询问机组是否 xx 管制指挥.上 7500 米,机组凭记忆答复是，随后 xx 管制指挥飞机继续上 7800 米保持。经机组回忆，当时航路有雷雨，机组正忙于绕飞雷雨，疑似听错高度指令。经核实,事发时距离对向飞机 95 公里,两机之间的距离符合安全间距，产生冲突。由于绕飞雷雨，此次机组大部分精力集中在执行雷雨绕飞工作中，精力有所分散造成错误地设置了预选高度。其次,机组未严格 SOP 标准喊话和操作程序，管制区域交接没有证实高度指令以及交叉检查高度设置。

Analysis

(1)航路上存在雷雨天气，机组忙于申请绕飞,分散了注意力导致误调了逆向高度后没有立即向管制员证实高度。

(2)航路上雷雨绕飞飞机较多,导致无线电通讯繁忙且通讯环境嘈杂。

(**a**)

Title:

A320 flying at the wrong altitude on xx/xx/2017

Narrative:

On xx. xx. 2017, an A320 aircraft was on flight xx-xx. The aircraft took off from Hefei at xx: xx and entered the control area at around xx: xx. At this time, the aircraft was at an altitude of 7,380 m and xx control asked the crew if xx control was in command. The crew replied in the affirmative from memory, and then xx control directed the aircraft to continue to 7800 m to maintain. As recalled by the crew, the crew recalled that there was a thunderstorm on the flight path at that time and the crew was busy flying around it, so it was suspected that they had misheard the altitude instruction. Verified, the incident occurred at a distance of 95 km from the opposite aircraft and the distance between the two aircraft was in line with the safe spacing and no conflict arose. Due to the thunderstorm, most of the crew's attention was focused on performing the thunderstorm detour and was distracted from the task, resulting in the incorrect setting of the pre-selected altitude. Secondly, the crew did not strictly follow the SOP standard shouting and operating procedures, and the control area handover did not confirm the altitude instructions and cross-checking altitude settings.

Analysis

(1) There was a thunderstorm on the route and the crew was busy applying for a go-around, which distracted their attention and led to the crew was distracted by the wrong reverse altitude and did not confirm the altitude with the controller immediately.

(2) There were many detours on the route, resulting in busy radio communications and a noisy communication environment.

(**b**)

Figure 1. An example of an aviation incident report in Chinese (**a**) and its English version (**b**).

2.2. Natural Language Processing

NLP is a popular area in computer science that is closely related to technologies in several fields, including artificial intelligence, Internet technologies, and mathematics. In effect, it enables computers to derive meaningful information from natural language and communicate effectively with humans. NLP utilizes increasing computing power to process a large volume of digital information and has been applied in machine translation, knowledge graphs, and automatic abstract generation [19]. When employing NLP for downstream tasks, the text needs to be pre-processed and vectorized. Standard steps of Chinese text pre-processing include data filtering, spelling standardization, Chinese word segmentation, stop-words removal, and part-of-speech (POS) tagging. Data filtering removes nonstandard reports or reports that are impossible to analyze, which ensures that the data to be analyzed include texts available for research. Spelling standardization automatically corrects misspelled words and replaces abbreviations and synonyms with standard expressions to enhance robustness. Chinese word segmentation divides a sentence into tokens, such as single words and phrases [20]. Conjunctions and adverbs provide almost no valid information for text analysis and thus, are removed to reduce redundancy. POS tagging attaches a tag to each token for its part of speech. However, Chinese does not require the steps of stemming and capitalization normalization usually used in English.

Text vectorization is an approach that is used extensively to transform unstructured text into a structured representation. Bag-of-words (BoW) representations and word embeddings are the two commonly used models. The BoW representation regards each document as a bag of words, neglecting their order, grammar, and syntax. In effect, each document can be represented as a vocabulary-length vector, in which the values are equal to the number of occurrences of the corresponding words. However, BoW ignores the order of words and cannot reflect their importance. The employment of term frequency–inverse document frequency (TF-IDF) can address this and provides the importance of each word in a document [21], as shown in Equation (1).

$$x_{ik} = f_{ik} \times log\left(\frac{N}{n_i}\right) \tag{1}$$

where f_{ik} is the occurrence time of word i in document k, N is the number of documents in the dataset, and n_i is the number of documents in which word i appears. n-grams, also known as combinations of consecutive tokens, can capture the syntactic information of words, and n denotes the number of tokens. In practice, n is usually 4 or less to avoid overly sparse vectors [22].

Tripathy et al. adopted word embedding techniques to derive a dense vector representation of a document [23]. With word embeddings, each word can be represented as a dense vector. Likewise, documents can be represented as a combination of the vectors of their words. The dimensions of the embedded space share potential feature, so models can be trained to capture semantic and syntactic similarities and other linguistic laws. Before solving a specific task, the word embedding model can be trained on an external dataset in the domain to best initialize the vectors of words [23].

2.3. Reports Analysis

NLP has been widely used in report analysis, especially in construction, medicine, and chemical engineering, due to its enormous and easily accessible databases. In most studies, classifying incident reports in a supervised manner is the first task of report content analysis, after which unsupervised methods are used to extract what is meaningful, such as themes and keywords.

Supervised algorithms have been frequently applied to handle multiple classification tasks. Tanguy et al. [11] trained 37 SVM binary classifiers to handle 37 categories of aviation reports, and the features for classification were selected from stems, words, and n-grams. Goh and Ubeynarayana [5] evaluated six supervised classifiers on 1000 publicly available construction accident narratives obtained from the US Occupational Safety and Health

Administration (OSHA) website. Baker et al. [7] compared deep learning approaches and TF-IDF+SVM in classifying injury precursors from raw construction accident reports. Chang and Shiwu built a knowledge graph (KG) from historical railway safety reports, and applied it in hazards identification and risk assessment [24]. Abdhul et al. proposed an automated and semi-supervised text mining method to analyze accident reports, domain-keywords would be identified and classified into topics [25]. Na et al. devised a text mining framework to forming a tailored domain lexicon of workspace accident and they extracted risk factors from metro construction accident reports [25]. In the identification, they utilized the qualitative variables of accident reports like location and work details. Zunxiang et al. applied text mining and complex network theory to explore the mechanism of coal mine accident-causing [26]. Tixier et al. [25] proposed a rule-based system to extract attributes (i.e., injured parts, energy sources, and body parts) from construction accident reports. Zhang et al. [25] adopted an unsupervised approach that used part-of-speech (PoS) tags to extract causes or harmful objects in accidents from titles.

Unsupervised methods were used in analyzing accident reports. Tanguy et al. modelled 163,570 documents to capture hidden information about events [11]. Tixier et al. [27] proposed a rule-based system to extract attributes (i.e., injured parts, energy sources, and body parts) from construction accident reports. Zhang et al. [10] adopted an unsupervised approach that used POS tags, namely, chunking, to extract causes or harmful objects in accidents from titles. Hui et al. adopted the Latent Dirichlet Allocation (LDA) model to detect topic words from hot work accidents [28]. Bomi and Yongyoon used the Local Outlier Factor (LOF) algorithm to detect anomalies of the chemical process [29].

Unlike incidents in the construction industry, aviation accidents are rare and difficult for researchers to access; therefore, the analysis of aviation accident reports is infrequent. Li et al. [14] applied the human factors analysis and classification system (HFACS) to investigate human error in aviation accidents and performed experiments on 41 accidents between 1999 and 2006 in Taiwan, China. Kelly and Efthymiou [15] adopted HFACS to identify the effects of human factors in controlled flight into terrain (CFIT), and 1289 unsafe actions and preconditions that contributed to events were identified from 50 CFIT accidents. Karanikas and Nederend [12] proposed a framework to classify aviation events for controllability and evaluate the potential of an event to escalate into higher severity classes.

2.4. Metrics

In this study, we adopted four metrics, i.e., true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as shown in Table 1.

Table 1. Meaning of TP, FP, FN, and TN.

	Positive	**Negative**
Identified	True Positive (TP)	False Positive (FP)
Not identified	False Negative (FN)	True Negative (TN)

Three metrics (precision, recall, and F1-score) were used to evaluate the performance of the proposed methods. Precision refers to the ratio of correct estimations to the estimated number, as defined in Equation (2). Recall is the proportion of identifications to actual positives, as defined in Equation (3). F1-score includes the capacities of precision and recall, making it a more comprehensive measurement, as defined in Equation (4).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

3. Report Classification

As presented in Figure 2, we divided the procedure of report classification into three parts, including labelling, pre-processing, and vectorization. We present the details in the following subsections.

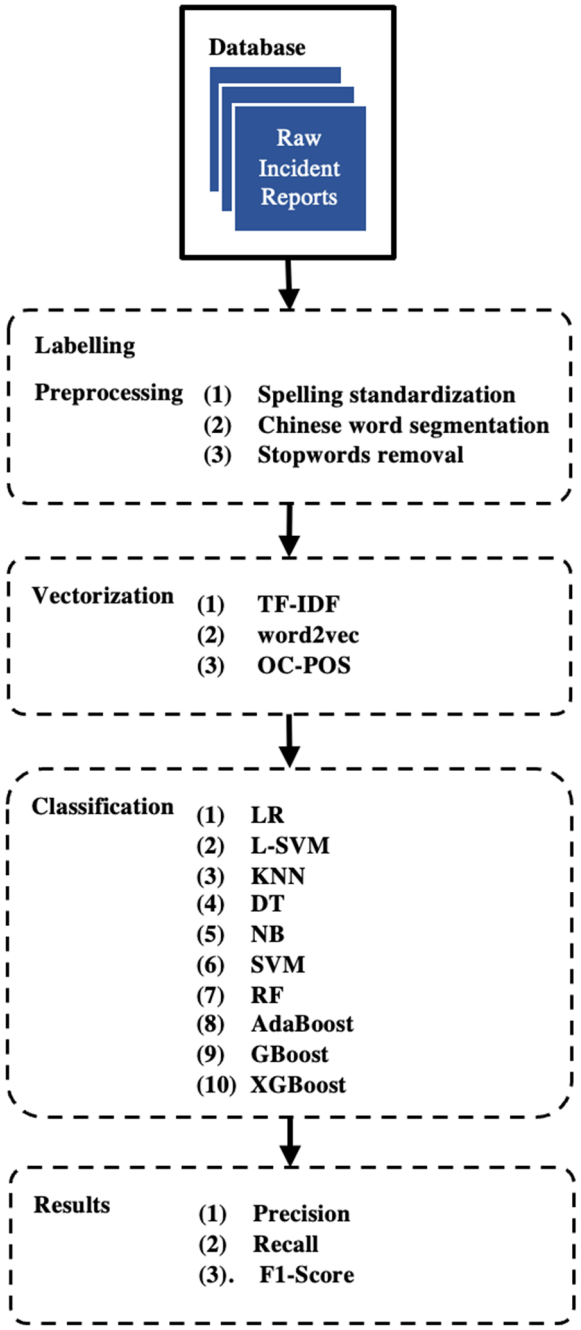


Figure 2. Framework of report classification.

3.1. Labelling and Pre-Processing Reports

In this study, the dataset for classification contained 1775 reports randomly selected from the CASTC database. Since only a small portion had titles, only narratives and analysis were used to extract features. In general, one report only has one incident label. If a report records multiple incidents (some unsafe incidents could trigger each other), it would be labelled based on the severity or causal order of the incidents.

For example, bad weather and poor communication were both recorded in the report in Figure 1. However, as the title indicated, the most crucial mistake was the Wrong height, which refers to the incident type: Deviation from the procedure. Besides, one report would be labelled as Deviation from the procedure if the incident belongs to one of the classes Deviation from departure procedure, Deviation from approach procedure, Yaw, Wrong height, and so on.

The reports in the dataset were labelled as one of 11 categories: Object Strike, Deviation from the procedure, Mechanical failure, Ground operation and maintenance, Landing problems, Engine breakdown, Environment incident, Cabin safety problems, Communication interrupt, Tail strike, or other. As shown in Table 2, the amount across categories is uneven, with Bird Strike accounting for more than one-third of cases and Tail Strike accounting for only 1.3% of the dataset, i.e., 23 reports. Table 2 shows the scopes of the categories.

Table 2. Incident categories, description and their amounts.

Category	Description	Sample	Proportion
Bird Strike (BS)	Aircraft collides with birds during take-off, landing, or cruising	611	34.4%
Deviation from procedure (DP)	Deviation from departure procedures, approach procedures, or instructed height	239	13.5%
Mechanical failure (MF)	Component failure or mechanical damage to the aircraft	228	12.8%
Ground operation and maintenance (GOM)	Aircraft damage caused by improper handling by ground service personnel	207	11.7%
Landing problems (LP)	Runway run-off, heavy landing, and other similar incidents	162	9.1%
Engine breakdown (EB)	Engine failure due to power system failure	128	7.2%
Environmental incident (EI)	Difficulty in operation or damage in harsh conditions	101	5.7%
Cabin safety problem (CSP)	Unsafe situation in the cabin because of passengers or cargo	37	2.1%
Communication interruption (CI)	Communication with the ground is interrupted during flight	25	1.4%
Tail strike (TS)	Collision between aircraft fuselage and ground due to improper operation or other factors during takeoff or landing	23	1.3%
Other	Other incidents, including crew incapacitation during flight	14	0.8%

Pre-processing steps were conducted sequentially on the dataset for classification:

(1) Spelling standardization: In this step, we removed messy codes and replaced abbreviations and irregular words with formal words. Besides, we replaced English and abbreviations with corresponding Chinese words. For example, “3发”(third engine), “左发” (left engine), and “ATC” (air traffic control), which refer to the third engine, left engine, and Air Traffic Control, respectively, were replaced with “第三发动机” (third engine), “左发动机” (left engine), and “航空交通管制” (air traffic control).

(2) Chinese word segmentation: Unlike the blank spaces used in English writing, no formal separator is adopted in Chinese, which makes separating each sentence into meaningful Chinese words a prerequisite. In this study, we used a Python package called Jieba (https://github.com/fxsjy/jieba, accessed on 15 September 2021) for Chinese word segmentation [20].

(3) Stopwords removal: Stopwords removal could reduce the dimensionality of text features. By extending the stopwords list provided by the Harbin Institute of Tech-

nology (https://github.com/goto456/stopwords, accessed on 15 September 2021), we removed stopwords with little meaning, such as "的" (of) and "以后" (afterward), and punctuations.

3.2. Vectorization

Before classification, each pre-processed report needed to be represented using a vector. Bag-of-words (BoW) representations and word embeddings are the two common models. The BoW representation regards each document as a bag of words, neglecting their order, grammar, and syntax. In effect, each document can be represented as a vocabulary-length vector, in which the values are equal to the number of occurrences of the corresponding words. However, BoW ignores the order of words and cannot reflect their importance. The employment of term frequency–inverse document frequency (TF-IDF) can be a complement by providing the importance of each word in a document [21], as shown in Equation (5).

$$x_{ik} = f_{ik} \times log\left(\frac{N}{n_i}\right) \tag{5}$$

where f_{ik} is the occurrence time of word i in document k, N is the number of documents in the dataset, and n_i is the number of documents in which word i appears. Similarly, each word can be represented as a fixed-length vector with word embeddings.

Tripathy et al. [23] adopted word embedding techniques to derive a dense vector representation of a document. Since aviation accident reports are very different from texts in other domains, pre-training on a larger dataset may allow the model to capture the meaning of words more accurately. Therefore, the word-embedding model was trained on the entire CASTC database instead of only on the labeled reports. In implementing the word-to-vector (Word2Vec) method, Gensim [30] was used to learn 128-dimensional word vectors in the database, with a window size of five and the skip-gram variant.

Additionally, considering that a long and sparse feature vector of the labelled dataset may result in overfitting, we proposed a BoW method called occurrence position (OC-POS), which originates from the occurrence of keywords and their position in the report. The value of the vector can be calculated using Equation (6).

$$M_i = \sum_{k=1}^{n} log\left(2 - \frac{p_i^k}{N}\right), 1 \le i \le L \tag{6}$$

where L is the length of the bag of words, N is the total number of words appearing in the report, and p_i^k is the relative position of the kth occurrence of word k. Keywords were collected from the reports with human intervention.

3.3. Experiments and Results

In this study, we tried the three aforementioned vectorization methods, i.e., TF-IDF, Word2Vec, and OC-POS. After vectorization, reports with labels were automatically classified by machine learning algorithms using python [31]. Given the above-introduced features, a preliminary experiment was conducted to find the classifier that produced the best results. Ten machine learning algorithms were adopted, including logistic regression (LR), linear SVM (L-SVM), k-nearest neighbor (KNN), decision tree (DT), naive Bayes (NB), SVM, random forest (RF), adaptive boosting (AdaBoost), gradient boosting (GBoost) and extreme gradient boosting (XGBoost) methods [31,32].

In the preliminary experiment, we performed 5-fold cross-validation for each combination of vectorization method and classifier. The parameters of all classifiers were set to default values. Table 3 shows the weighted F1 of classification, and the highest weighted F1 for each classifier is highlighted in bold. Consistent with Zhang et al. [11], Word2Vec could not represent reports well, which may be attributed to the overly great length, which introduced some noise from irrelevant data in vectorization. Compared to TF-IDF, the OC-POS performed better on all classifiers due to the negative effect of long feature vectors.

Table 3. Overall weighted F1 of preliminary classification.

	TF-IDF	Word2Vec	OC-POS
LR	0.792	0.685	**0.834**
L-SVM	0.807	0.730	**0.811**
KNN	0.382	0.712	**0.768**
DT	0.608	0.627	**0.720**
NB	0.641	0.606	**0.642**
SVM	0.619	0.625	**0.794**
RF	0.674	0.729	**0.814**
AdaBoost	0.275	0.423	**0.427**
GBoost	0.756	0.742	**0.825**
XGBoost	0.807	0.751	**0.835**

The SVM classifier gave average results, while the L-SVM performed better in preliminary experiments. AdaBoost, NB, and KNN performed poorly with well-below-average results. The suggestion from researchers that logistic regression performs better on small datasets while tree models perform better on larger datasets was confirmed with LR achieving great performance. It is worth noting that XGBoost achieved almost optimal results for all three feature vectors, which showed that it was reliable in all circumstances. Subsequently, a time-consuming grid-search was implemented to tune the parameters of XGBoost.

Before model training, the vectorized reports were randomly divided into a training set and a test set, and the ratio was 80% for training and 20% for testing. Table 4 lists the optimal parameters obtained by 10-fold grid search.

Table 4. Parameters of Xgboost classifier.

Parameter	Value
max_depth	15
learning_rate	0.15
n_estimators	50
min_child_weight	0
max_delta_step	1
subsample	0.8
colsample_bytree	0.6
reg_alpha	0.75
reg_lambda	0.4

Table 5 shows the results, including precision, recall, and F1 score, along with the support numbers of the test set after classification by the trained XGBoost. One finding is that the more support samples there are, the more accurate the classification. Bird Strike and Tail Strike can be classified almost correctly due to their specific tokens, such as “鸟” (bird) and “擦” (rub). Cabin safety and Other categories had the worst results, with only 0.57 in precision and 0.5 for all metrics, respectively. The classification results for the other categories were satisfactory, with F1 scores above 0.75.

As shown in Table 6, a confusion matrix evaluates the mislabeled cases. Cabin Safety incidents could be wrongly classified into Ground Operation and Maintenance since some ground events also occurred in the cabin, and the classifier did not capture the difference. The Other category is a composite of multiple incidents without consistent features, which could explain its inaccurate classifications. Any incident could be a complex process during which several unsafe events occur, and our labelling strategy focuses only on the outcome or the most severe one. NLP cannot identify the category unless all the details are organized formally and uniformly. As discussed, some reports are ambiguous regarding their labels if multiple events or unsafe factors are recorded in one report. Therefore, we recommend that human intervention be implemented after automatic classification, especially for certain incidents, such as Cabin safety.

Table 5. Final classification results.

	Precision	Recall	F1-Score	Support
Bird strike	0.99	1.00	0.99	123
Deviation from procedure	0.82	0.84	0.83	43
Mechanical breakdown	0.89	0.92	0.91	52
Ground operation and maintenance	0.81	0.73	0.77	41
Landing problems	0.84	0.79	0.82	34
Engine breakdown	0.89	0.83	0.86	29
Environmental incident	0.75	0.88	0.81	17
Cabin safety	0.57	0.80	0.67	5
Communication interruption	1.00	0.80	0.89	5
Tail strike	1.00	1.00	1.00	4
Other	0.50	0.50	0.50	2
Weighted value	0.85	0.86	0.85	355

Table 6. Confusion matrix of classification for the test dataset.

	DP	EB	GOM	EI	CS	TS	MB	LP	CI	BS	Other
DP	36	1	2	0	0	0	1	3	0	0	0
EB	0	24	1	1	0	0	2	0	0	0	1
GOM	2	1	30	0	3	0	3	1	0	1	0
EI	1	0	1	15	0	0	0	0	0	0	0
CS	0	0	1	0	4	0	0	0	0	0	0
TS	0	0	0	0	0	4	0	0	0	0	0
MB	0	0	0	3	0	0	48	1	0	0	0
LP	4	1	1	1	0	0	0	27	0	0	0
CI	1	0	0	0	0	0	0	0	4	0	0
BS	0	0	0	0	0	0	0	0	0	123	0
Other	0	0	1	0	0	0	0	0	0	0	1

4. Cause Extraction

In this section, we devised a rule-based system for extracting the causes from the report. Our system consists of a keyword dictionary and a rule set, referring to the research in construction [5]. Keywords were collected manually from texts, and the rules were designed according to the descriptions of causes. The principle of cause extraction is to scan incident reports for each cause using the corresponding keywords and rules. Here, we first introduce the dataset for performing cause extraction. Then, introduce the steps of pre-processing and the rule set. Finally, validate the robustness of the system through an experiment, and illustrate its principles with an example.

4.1. Causes Description

There are always several causes behind every unique event. In the example in Figure 1, the distraction of the pilots, lack of cross-checking, interfering movements, complex weather (thunderstorms), and poor communication between the pilots and traffic controllers were the causes to be identified. These causes are not independent but are interrelated logically and temporally. The HFACS model [33] was adopted to categorize causes structurally. In this study, we categorize causes into four categories: Equipment, Environment, Human, and Organization. Before identification, reports were labelled manually according to the content.

In implementing the labelling, we performed a content analysis on the 1775 reports of the dataset in Section 3. As shown in Table 7, 25 causes were listed, along with their coding scheme and occurrences. In detail, equipment causes are mechanical and electrical faults that occur during the flight or design phase. Environmental factors are elements that are beyond human control, such as the external environment and unexpected situations. Human attributes are unsafe actions of humans consisting of decisions, skill-based or perceptual errors, violations of pilots, inadequate supervision by traffic control, and incorrect

actions of the ground crew. Organization causes influences flights at the management level, including shortcomings in standards, supervision, or pilot training. The role of a cause may vary across incidents. For example, in the case of bird strike, engine failure, and landing problem events, engine failure is a precursor, a consequence, and both a precursor and a consequence.

Table 7. Causes and occurrences.

Hazard	Code	Frequency of Occurrence
Equipment	EI	
Dynamical system malfunction	EI1	70
Component failure	EI2	91
Landing gear and tire malfunction	EI3	17
Electric system and control system malfunction	EI4	53
Autopilot disengaged and sensor fault	EI5	44
Communication equipment disconnection	EI6	15
Flaw in design	EI7	10
Equipment breakdown in cabin	EI8	5
Environment	E	
Bird strike	E1	411
Slippery runway	E2	7
Low visibility and bad weather	E3	159
Wind shear	E4	117
Discomfort of flight crew and passengers	E5	8
Strike by other objects	E6	52
Unusual terrain	E7	25
Human	H	
Misjudgement and incorrect behaviour of crew	H1	280
Deregulation of crew	H2	216
Poor preparation and capacity of crew	H3	214
Disorder of crew resource management	H4	159
Distraction and negligence of crew	H5	231
Maloperation of ground crew	H6	93
Failure in communication with air traffic control (due to human error)	H7	88
Organization	O	
Imperfect flight standards	O1	26
Insufficient supervision of operational practices	O2	85
Insufficient safety training of staff	O3	17

4.2. Pre-Processing

Pre-processing included spelling standardization, Chinese word segmentation, and PoS tagging. Spelling standardization and Chinese word segmentation were similar to the pre-processing steps in Section 3. Since periods and commas separated each sentence and clause respectively, keywords related to actions in different sentences would not be attributed to the same subject in one sentence; consequently, commas and periods were retained. In addition, the numbers were retained, as a report may only record the parameters and not the comments in abnormal conditions. These numbers provided extra information about the operational state of aircraft or the environment. The PoS tags indicated the syntactic sequence of each sentence, which consisted of several parts, such as nouns (i.e., subject and object), verbs, pronouns, and prepositions.

4.3. System Design

Rule-based models and machine-learning algorithms are two major approaches to building a cause identification or classification system. Recently, machine-learning algorithms have been applied to text analysis on accident reports. Unfortunately, these methods have several limitations in this study. First, it is difficult to find the description of causes in reports, so the characteristics of the causes will be uncertain for these methods. In most

cases, the description of each cause appears in a sentence without a fixed location, and each report consists of dozens of sentences. Second, the identification effect of causes with low support is poor. A fairly high number of cases (e.g., 75 to 100) is the minimum needed to obtain a valid statistical model, even when learning shorter texts. However, several reasons, such as E8, C2, and C5, occur fewer than ten times in our dataset. Third, machine-learning algorithms cannot perform effective multilabel classification in our dataset. If each sample has one or more labels, 25 categories are too large a number for classifying 1775 samples. Fourth, though unsupervised machine-learning algorithms allow classification based on word distribution, they cannot meet the requirement of identifying specific causes.

Therefore, a rule-based system seemed to be a better choice because of the following advantages. First, the precondition of the dataset scale is not specified if using rules. We can design specific rules for each cause, even without high support. Second, the upgrading and iteration for rules in a rule-based system can directly improve performance and can easily determine the content related to rules. In contrast, statistical models improve their performance by blind tuning the parameters, which indicates only a better result or a worse one, without knowing the details. Third, the rule-based system enables the intervention of human knowledge and intelligence. In a specific domain, professional information matters since it goes beyond the limits of available data. While statistical models can obtain broad but shallow features, a rule-based model provides considerable insight into the incident-related context. Zhang [11] used an unsupervised approach based on grammar rules, namely, chunking, to extract common causes of construction accidents. As noted in [5], such a system is not elegant but is effective.

As mentioned above, the rule-based system is composed of cause-related keywords and rules. The rules, which can be regarded as grammar are established based on the keywords. Therefore, they would be updated at the same time. Here, we provide the details of how we build the keywords dictionary and rules set.

Keywords and rules are equivalent to the components of a sentence and grammar, respectively. In modern Chinese, the essential components, such as subject, verb, object, and the order of these components, are consistent in formal paperwork. Thus, three kinds of keywords were sought: the subject keyword (SK), attribute keyword (AK), and evaluation keyword (EK). The SK is the unit responsible for the cause, and it is always the subject in every sentence. Since the reports are written by multiple entities, the same subject may have several representations. For example, more than ten nouns could represent pilots in this context, such as “飞行员” (Pilots), PF (Pilot Flying), and “驾驶舱” (Cockpit), although their independent meanings are not exactly the same. Since H1 to H5 are all descriptions of pilot errors, they share the same SK list. AK is a property or action of SK, which is associated with a cause. Similarly, a single cause can lead to different behaviors in different situations, so each cause has dozens of AKs to be detected. For example, H1 corresponds to “预估” (estimate/estimation), “评判” (judge/judgement), “决断” (decide/decision) and so on. The SK and AK can only describe the behavior of a unit, which is not sufficient to make a judgement regarding it, so the EK is introduced. As assessments of the AK, EKs are adjectives and adverbs, corresponding to different parts of speech of AKs. Since they were all negative words, the EKs were only divided into adjectives and adverbs. For instance, “欠佳” (not good enough), “缺乏” (be short of), “过于” (too much), “过低” (too low) and “欠妥” (not proper) were the most frequent adjectives and adverbs among EKs.

Specific keywords alone are not sufficient to determine the cause, and rules are needed to guide the usage of keywords. The principle of the rules, as mentioned previously, is to determine one cause from the keywords and other information in the sentence. However, not all cases were detected with rules, and this depended on the type of cause and the written patterns. Six causes (E1, E3–E7) could be determined by only detecting keywords (SKs or AKs). Taking E1 (bird strike) as an example, the occurrence of bird-related words, such as “鸟击” (bird strike), “羽毛” (feather) and “鸟血” (bird’s blood), was enough for identification. In identifying these six causes, some cases also adopted special rules. Bird-related words (SKs) and trace-related words (AKs) with no human-related words (SKs)

in the context were also needed when identifying E1. In addition, the numbers recorded quantifiable parameters, such as visibility and lateral wind speed. Such values reached a threshold that triggered identification. In detail, visibility below 1,000 meters triggered E3, and a lateral wind speed over 10 m/s triggered E4.

For EI-type causes, the occurrence of keyword (SK and AK) combinations could be used for identification. These causes were detected using identical words, although these words were not always in the same order. For example, "链条断裂" (chain break) and "断裂的链条" (broken chain) could both trigger the identification of EI2 by detecting "链条" (chain) and "断裂" (break).

H-type and O-type cause-related rules were more complex because more than one AK and SK were distributed in different clauses of the same sentence. The key was pairing each AK with its corresponding SK. Tixier et al. [5] used fixed radii to link keywords, of which the distances from each other less than 7 could be related to each other. As this value could vary for different contents, it is not robust enough, especially in a sentence that contains multiple clauses. Instead of fixed radii, we adopted another effective distance, which depended on other information from the sentence. The scope was initialized with punctuation. Assuming that a clause was the minimal scope, commas and periods could separate clauses from each other. In addition, words like "和" (and) or "他" (he) would extend the scope of the SK to next clause.

Here, we illustrate the whole process of cause identification with an example report shown in Table 8. First, the keyword "雷雨" (thunderstorms) in sentences #2, #4, and #7 can determine the cause E3. Second, in sentence #4, only "机组" (aircrew) serves as an SK, and all the AKs in this sentence relate to pilots. "精力" (vigor) and "设置" (set) are the AKs in this sentence. "Vigor" is a noun, and the closest EK is "分散" (distracted), which is a adjective and can be the predicate of "vigor". As an adverb, "错误地" (wrongly) could be the EK of the verb "set". Therefore, H5 and H1 are the causes identified in sentence #4. In sentence #5, "机组" (pilots) and "管制" (control) could be the SK, while "管制区域" (control area) is a compound-specific word, which makes the noun "管制区域交接" (control area handover) an attribute of the pilot. The AK "喊话" (communication) and the EK combination of "未" (none) and "标准" (standard) can identify H2. When identifying H4 from AK "交叉检查" (cross-check), "以及" (as well as) expands the scope of EK "没有" (no), which makes "no" an EK of "cross-checking". The analysis is almost consistent with the narrative and can identify similar causes. However, "通讯环境嘈杂" (noisy communication environment) is mentioned as a consequence of thunderstorms.

Table 8. Example illustration of cause identification.

Part	Sentence	Context in Chinese	English Version	Identified Causes
Title	0	2017年xx月xx日A320飞机飞错高度	A320 flying at the wrong altitude on xx/xx/2017	-
Narrative	1	2017年xx月xx日, A320飞机执行 xx-xx 航班, xx: xx于合肥起飞, xx: xx左右该航班进入 xx 管制区域, 此时飞机高度7380米, xx管制询问机组是否xx管制指挥上7500米, 机组凭记忆答复是, 随后xx管制指挥飞机继续上7800米保持。	On xx. xx. 2017, an A320 aircraft was on flight xx-xx. The aircraft took off from Hefei at xx: xx and entered the control area at around xx: xx. At this time, the aircraft was at an altitude of 7380 m and xx control asked the crew if xx control was in command. The crew replied in the affirmative from memory, and then xx control directed the aircraft to continue to 7800 m to maintain.	-
	2	经机组回忆, 当时航路有雷雨, 机组正忙于绕飞雷雨, 疑似听错高度指令。	As recalled by the crew, the crew recalled that there was a thunderstorm on the flight path at that time and the crew was busy flying around it, so it was suspected that they had misheard the altitude instruction.	E3

Table 8. *Cont.*

Part	Sentence	Context in Chinese	English Version	Identified Causes
	3	经核实, 事发时距离对向飞机95公里, 两机之间的距离符合安全间距, 未产生冲突。	Verified, the incident occurred at a distance of 95 km from the opposite aircraft and the distance between the two aircraft was in line with the safe spacing and no conflict arose.	-
	4	由于绕飞雷雨, 此次机组大部分精力集中在执行雷雨绕飞工作中, 精力有所分散造成错误地设置了预选高度。	Due to the thunderstorm, most of the crew's attention was focused on performing the thunderstorm detour and was distracted from the task, resulting in the incorrect setting of the pre-selected altitude.	E3, H5, H1
	5	其次, 机组未严格50 p标准喊话和操作程序, 管制区域交接没有证实高度指令以及交叉检查高度设置。	Secondly, the crew did not strictly follow the SOP standard shouting and operating procedures, and the control area handover did not confirm the altitude instructions and cross checking altitude settings.	H2, H4
Analysis	6	(1)航路上存在雷雨天气, 机组忙于申请绕飞, 分散了注意力导致误调了逆向高度后没有立即向管制员证实高度。	(1) There was a thunderstorm on the route and the crew was busy applying for a go-around, which distracted their attention and led to the crew was distracted by the "wrong" reverse altitude and did not confirm the altitude with the controller immediately.	H5
	7	(2)航路上雷雨绕飞飞机较多, 导致无线电通讯繁忙且通讯环境嘈杂。	(2) There were many detours on the route, resulting in busy radio communications and a noisy communication environment.	E3, H7

"疑似听错" (suspected of mishearing) is not an assertion but a conjecture, which is not used in identification. An analysis of possible causes could also be recorded, even though these causes are not mentioned in the report. One observation is that certain words, such as "未", "没", and "疑似", appear in context after manual analysis. Such words should be detected if a cause is identified to filter this noise.

4.4. Experimental Results

As shown in Figure 3, the dataset was divided into a training set and a test set, with each take part of 50%. The F1-score was adopted to evaluate the performance. We conducted cross-validation on the training set first. All reports were randomly divided into five sets, and each one built keywords and rules independently. The updating of keywords and rules is an iterative process. The keywords and rules in the first version were derived from a manual analysis process for each set of reports. The rules that led to error detection were modified based on their performance in other situations until an optimal result was reached. Then, five sets of rules, which performed best in the corresponding report sets, were used for detecting other sets, namely, for performing cross-validation. Table 9 reports the F1 scores for the 5-fold cross-validation.

Although using keywords and rules from other reports, the minimum F1 score for each set was above 0.80, which is an acceptable result in terms of cause identification. This means that these rules can apply to other reports in the database. In addition, the system is sufficiently robust, and the best F1 scores were all above 0.90. After cross-validation, five sets rules were integrated and tuned to their best performance on the whole training set. The results of cause detection on the test set were the final evaluation of this system.

Considering the prevalence of synonyms, we used the Word2Vec technique to expand the keywords (mainly AKs).

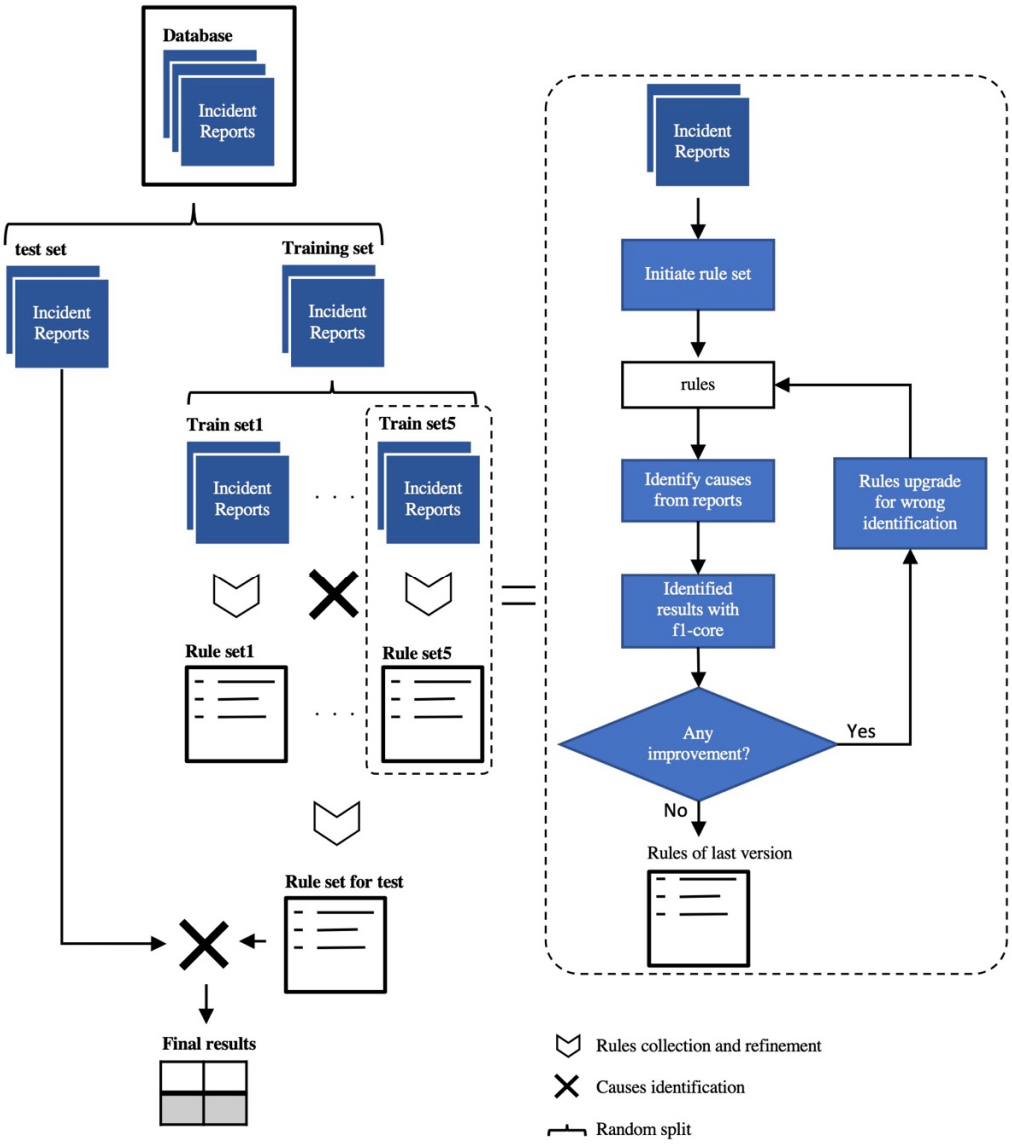


Figure 3. The framework of the experiment for cause identification.

Table 9. Cross-validation of the cause identification system.

	Rules s1	Rules s2	Rules s3	Rules s4	Rules s5
Report s1	0.91	0.88	0.88	0.84	0.88
Report s2	0.85	0.91	0.9	0.86	0.89
Report s3	0.86	0.91	0.94	0.88	0.93
Report s4	0.84	0.89	0.9	0.92	0.93
Report s5	0.85	0.88	0.88	0.9	0.9

Table 10 summarizes the precision, recall, and F1 score of the test set. In general, even if the rules were derived from other reports, the results yielded F1 scores of 0.90 or higher. However, there is a clear difference between precision and recall. Higher recall (0.95) means that almost all causes can be detected, while the relatively low precision demonstrates an unremarkable accuracy of causes identification.

Table 10. Final results of cause identification.

	Precision		Recall		F1	
	RAW	**Word2Vec**	**RAW**	**Word2Vec**	**RAW**	**Word2Vec**
EI1	0.938	0.682	1.000	1.000	0.968	0.811
EI2	0.800	0.765	0.900	0.975	0.847	0.857
EI3	1.000	0.833	1.000	1.000	1.000	0.909
EI4	0.755	0.905	0.974	1.000	0.851	0.950
EI5	0.966	0.903	1.000	1.000	0.982	0.949
EI6	0.769	0.769	0.909	0.909	0.833	0.833
EI7	1.000	1.000	0.800	1.000	0.889	1.000
EI8	1.000	1.000	0.800	0.800	0.889	0.889
EI	0.842	0.828	0.946	0.980	0.891	0.897
H1	0.883	0.886	0.945	0.980	0.913	0.931
H2	0.816	0.807	0.952	0.970	0.879	0.881
H3	0.910	0.923	0.973	0.979	0.940	0.950
H4	0.940	0.948	0.948	0.948	0.944	0.948
H5	0.878	0.931	0.946	0.970	0.911	0.950
H6	0.925	0.930	0.875	0.946	0.899	0.938
H7	0.827	0.681	0.925	0.925	0.873	0.785
H	0.878	0.875	0.946	0.966	0.911	0.918
E1	1.000	1.000	1.000	1.000	1.000	1.000
E2	1.000	0.667	1.000	1.000	1.000	0.800
E3	0.917	0.754	1.000	0.977	0.957	0.851
E4	0.932	0.932	1.000	0.986	0.965	0.958
E5	1.000	1.000	1.000	1.000	1.000	1.000
E6	1.000	1.000	0.857	0.857	0.923	0.923
E7	0.821	0.605	1.000	1.000	0.902	0.754
E	0.918	0.816	0.994	0.981	0.954	0.891
O1	0.735	0.730	0.862	0.931	0.794	0.818
O2	0.853	0.733	0.928	0.957	0.889	0.830
O3	0.778	0.778	1.000	1.000	0.875	0.875
O	0.811	0.738	0.920	0.955	0.862	0.833
Weighted	0.873	0.849	0.949	0.969	0.909	0.905

The results show that these rules are somewhat imperfect since many FPs that fit the rules are identified. In descending order of F1-score, the cause types are E-type, H-type, EI-type, and O-type; this is because the E-type is described by fewer keywords, as mentioned above. Additionally, the rules for H-type causes were the most complicated because of the complex grammar these sentences used. One observation is that the causes that share identical keywords with others have lower precision. For example, H6 and EI7 are both communication-related causes, and similar keywords will appear in both cases. EI2 can easily be confused with other EI-type causes, as damage to one component can cause other systems to fail. Dozens of keywords could trigger EI4, and the training set was not capable to contain all of them. H2 and O1 also had overlapping keywords since both they related to standards and regulations.

Applying Word2Vec, although the difference was not significant, the overall F1-score was lower. More keywords led to more false positives, and these misidentifications further reduced the precision. However, adding more keywords can also improve recall by adding related words to the identification. Using Word2Vec is not much different and the original rules give satisfactory results. On the other hand, to meet specific requirements, such as higher recall, Word2Vec, and other word-embedding techniques can be tried.

5. Conclusions

In this study, we explored a method for processing civil aviation accident reports that can be applied in practice. Using this method, we achieve the automatic classification and cause identification of Chinese civil aviation incident reports from the CASTC database. The Python language and several Python libraries were used in implementing these methods.

First, the XGBoost classifier and OC-POS vectorization methods are used to classify the text reports, as these methods perform best among ten classifiers and three vectorization strategies. As a result, the overall F1 score was 0.85, the precision for different categories was 0.50 to 0.99, the recall was 0.50 to 1.00, and the F1 score was 0.50 to 0.99, which is better than the results of other studies. This method enabled the automatic classification of reports, and the need to analyze several kinds of incidents could be satisfied. Second, we built a rule-based system to identify the causes of incidents. The proposed system obtained an F1-score above 0.90 when identifying 25 causes (8, 7, 7, and 3 equipment, human, environment, and organization causes, respectively) from our dataset. In addition, the basic rules were extended using Word2Vec. However, the precision improved, and the overall results (F1-score) worsened. Unlike most studies, our study was conducted on a Chinese dataset. It makes this study different in the pre-processing steps and text analysis.

There are some limitations of this study. First, the quality of the reports may influence the results. The database contains reports from different airlines, spanning more than a decade, and with inconsistent writing norms and standards. Invalid information might be recorded and could interfere with the experiment. In addition, writing errors (e.g., incorrect wording or missing and irregular punctuation) may occur during recording, transcription, and decoding. Although we corrected some mistakes before the experiment, the remaining errors may still compromise the robustness of cause identification. Second, these methods rely on the stability of libraries. In the classification, Jieba was adopted to process the Chinese tokenization, and the results were used to represent the reports. Third, the support samples were limited. Since the report writing is complex and conducting manual analysis is time-consuming and labor-intensive, we only selected a random portion of reports from the database for testing in the experiment. Finally, the identification of causes is not fully automatic. It still requires the manual collection of rules as well as keywords.

Further studies can be extended on this work. For classification, though the word-embedding model does not perform well in the experiments in this paper, they can try other word-embedding models. In addition, word-embedding models and TF-IDF vectorization cannot capture important information from long texts, which can be addressed by testing deep-learning methods with attention mechanisms. Furthermore, in this study, we obtained the causes and categories from the incident reports, which made it possible to analyze them further. Due to the complexity of incident reporting, especially in Chinese, a rule-based approach coupled with human intervention seems be a powerful approach to explore. In addition to causes, other crucial information can be identified in reports, such as the consequences of incidents.

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

Funding: Hubei Provincial Key Research and Development Program (No. 2021BAA185) and National Natural Science Foundation of China (U2033216, U1833201 and 42071368).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We appreciate the support of this work from Hubei Provincial Key Research and Development Program (No. 2021BAA185) and National Natural Science Foundation of China (U2033216, U1833201 and 42071368).

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

References

1. Janic, M. An assessment of risk and safety in civil aviation. *J. Air Transp. Manag.* **2000**, *6*, 43–50. [CrossRef]
2. Marshall, P.; Hirmas, A.; Singer, M. Heinrich's pyramid and occupational safety: A statistical validation methodology. *Saf. Sci.* **2018**, *101*, 180–189. [CrossRef]
3. Huang, R.; Sun, H.; Wu, C.; Wang, C.; Lu, B. Estimating Eddy Dissipation Rate with QAR Flight Big Data. *Appl. Sci.* **2019**, *9*, 5132. [CrossRef]
4. Arnaldo Valdés, R.M.; Gómez Comendador, F. Learning from accidents: Updates of the European regulation on the investigation and prevention of accidents and incidents in civil aviation. *Transp. Policy* **2011**, *18*, 786–799. [CrossRef]
5. Tixier, A.J.P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **2016**, *62*, 45–56. [CrossRef]
6. Goh, Y.M.; Ubeynarayana, C.U. Construction accident narrative classification: An evaluation of text mining techniques. *Accid. Anal. Prev.* **2017**, *108*, 122–130. [CrossRef] [PubMed]
7. Kurian, D.; Sattari, F.; Lefsrud, L.; Ma, Y. Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. *Saf. Sci.* **2020**, *130*, 104873. [CrossRef]
8. Baker, H.; Hallowell, M.R.; Tixier, A.J.P. AI-based prediction of independent construction safety outcomes from universal attributes. *Autom. Constr.* **2020**, *118*, 103146. [CrossRef]
9. Cheng, M.-Y.; Kusoemo, D.; Gosno, R.A. Text mining-based construction site accident classification using hybrid supervised machine learning. *Autom. Constr.* **2020**, *118*, 103265. [CrossRef]
10. Tixier, A.J.P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining. *Autom. Constr.* **2017**, *74*, 39–54. [CrossRef]
11. Zhang, F.; Fleyeh, H.; Wang, X.; Lu, M. Construction site accident analysis using text mining and natural language processing techniques. *Autom. Constr.* **2019**, *99*, 238–248. [CrossRef]
12. Liu, C.; Yang, S. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Syst. Appl.* **2022**, *207*, 117991. [CrossRef]
13. 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]
14. Xu, N.; Ma, L.; Liu, Q.; Wang, L.; Deng, Y. An improved text mining approach to extract safety risk factors from construction accident reports. *Saf. Sci.* **2021**, *138*, 105216. [CrossRef]
15. Song, B.; Suh, Y. Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety. *J. Loss Prev. Process Ind.* **2019**, *57*, 47–54. [CrossRef]
16. Qiu, Z.; Liu, Q.; Li, X.; Zhang, J.; Zhang, Y. Construction and analysis of a coal mine accident causation network based on text mining. *Process Saf. Environ. Prot.* **2021**, *153*, 320–328. [CrossRef]
17. Xu, H.; Liu, Y.; Shu, C.M.; Bai, M.; Motalifu, M.; He, Z.; Wu, S.; Zhou, P.; Li, B. Cause analysis of hot work accidents based on text mining and deep learning. *J. Loss Prev. Process Ind.* **2022**, *76*, 104747. [CrossRef]
18. 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]
19. Karanikas, N.; Nederend, J. The controllability classification of safety events and its application to aviation investigation reports. *Saf. Sci.* **2018**, *108*, 89–103. [CrossRef]
20. 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]
21. Li, W.-C.; Harris, D.; Yu, C.-S. Routes to failure: Analysis of 41 civil aviation accidents from the Republic of China using the human factors analysis and classification system. *Accid. Anal. Prev.* **2008**, *40*, 426–434. [CrossRef] [PubMed]
22. Kelly, D.; Efthymiou, M. An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *J. Saf. Res.* **2019**, *69*, 155–165. [CrossRef]
23. Peng, H.; Cambria, E.; Hussain, A. A Review of Sentiment Analysis Research in Chinese Language. *Cogn. Comput.* **2017**, *9*, 423–435. [CrossRef]
24. Zhou, C.; Hu, D. Research on Inducement to Accident/Incident of Civil Aviation in Southwest of China based on Grey Incidence Analysis. *Procedia Eng.* **2012**, *45*, 942–949. [CrossRef]
25. Kamla, J.; Parry, T.; Dawson, A. Analysing truck harsh braking incidents to study roundabout accident risk. *Accid. Anal. Prev.* **2019**, *122*, 365–377. [CrossRef] [PubMed]
26. Liu, J.; Wu, F.; Wu, C.; Huang, Y.; Xie, X. Neural Chinese word segmentation with dictionary. *Neurocomputing* **2019**, *338*, 46–54. [CrossRef]
27. Peng, T.; Liu, L.; Zuo, W. PU text classification enhanced by term frequency–inverse document frequency-improved weighting. *Concurr. Comput. Pract. Exp.* **2014**, *26*, 728–741. [CrossRef]
28. Tripathy, J.K.; Sethuraman, S.C.; Cruz, M.V.; Namburu, A.; Mangalraj, P.; Vijayakumar, V. Comprehensive analysis of embeddings and pre-training in NLP. *Comput. Sci. Rev.* **2021**, *42*, 100433. [CrossRef]
29. Řehůřek, R.; Sojka, P. *Software Framework for Topic Modelling with Large Corpora*; ELRA: Luxembourg, 2010; pp. 45–50.

30. Schwarz, J.S.; Chapman, C.; McDonnell Feit, E. Welcome to Python. In *Python for Marketing Research and Analytics*; Springer: Cham, Switzerland, 2020; pp. 3–7. [CrossRef]
31. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 13–17 August 2016.
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *JMLR* **2011**, *12*, 2825–2830.
33. Kaptan, M.; Sarıalioğlu, S.; Uğurlu, Ö.; Wang, J. The evolution of the HFACS method used in analysis of marine accidents: A review. *Int. J. Ind. Ergon.* **2021**, *86*, 103225. [CrossRef]