aerospace

MDPI

Article

Standardized Extraction of Air Traffic Control Hazard Features Based on Expert Knowledge

Xianghua Tan [1], Zhipeng Cai [1,2], Zhibin Quan [3] and Weili Zeng [1,4,*]

[1] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; tanxh_21@nuaa.edu.cn (X.T.); caizhipeng@comac.cc (Z.C.)
[2] Shanghai Aircraft Test Flight Engineering Co., Ltd., Shanghai 200232, China
[3] School of Automation, Southeast University, Nanjing 210096, China; sigma_quan@seu.edu.cn
[4] State Key Laboratory of Air Traffic Management Systems, Nanjing 211106, China
* Correspondence: zwlnuaa@nuaa.edu.cn

Abstract: Air traffic control (ATC) hazard feature extraction is a key information retrieval task for air traffic hazard records. While text-based feature extraction ranks term importance based solely on statistical results, we aim to use external knowledge to extract features that meet the definition of hazards. This paper proposes a feature extraction method based on expert knowledge to define hazard features and construct a hazard analysis framework. We illustrate the model training process using communication navigation and surveillance (CNS) data, which includes candidate feature generation, feature vectorization, and cluster-based standardization. The correct structure of terms in hazard records, the vector distribution of candidate features, and the clustering effect of different methods are briefly explored. The algorithm refines and accumulates expert knowledge through iteration. The experiment results demonstrate that the dataset obtained after specific linguistic processing based on expert knowledge could extract more informative candidate features to construct analysis context by k-means. The proposed model outperformed four comparative algorithms in accuracy, reaching 82% and 86% in the air traffic control operation (ATCO) dataset and the CNS dataset, respectively. Additionally, the information-rich hazard features support safety management departments' decision-making, reducing the cost of investigating hidden hazards.

Keywords: text feature extraction; hazard; air traffic control; standardization; expert knowledge

Academic Editor: Judith Rosenow

Received: 25 November 2024
Revised: 17 January 2025
Accepted: 21 January 2025
Published: 27 January 2025

Citation: Tan, X.; Cai, Z.; Quan, Z.; Zeng, W. Standardized Extraction of Air Traffic Control Hazard Features Based on Expert Knowledge. *Aerospace* **2025**, *12*, 94. https://doi.org/10.3390/aerospace12020094

1. Introduction

Safety and air traffic control operating efficiency have come into conflict as a consequence of the rapidly expanding aviation market. Global air passenger traffic increased by 7.5% from 2016 to 2019, but the number of accidents increased from 75 to 114 in 2019, with an average accident growth rate of 15% [1]. The rapid expansion of Chinese air traffic flow has exacerbated this problem. According to statistics from the Civil Aviation Administration of China in 2019, there were more than 10 million air traffic services, but the growth rate of accident symptoms was 25% higher than the growth rate of air traffic services [2]. Following COVID-19, air traffic flow is expected to increase significantly in China, posing additional safety concerns.

Safety is the cornerstone of aviation. Air traffic operating environments are dynamic and continuously present new safety hazards [3]. Learning lessons and extracting hazard information from past and ongoing hazards is essential for preventing similar hazards and improving safety management [4]. Hazard identification is the core of risk management

in civil aviation safety management. Aviation safety analysis methods based on text have either taken an accident analysis approach to disclose the relationship between hazards and accidents or an expert analysis approach to identify the causes of accidents following severe or rare accidents [5,6]. Because of the widespread adoption of safety concepts and advanced aviation data monitoring systems, management departments have easily acquired numerous text datasets from practitioner reports that can be used for in-depth safety evaluation and hazard detection [7]. How to extract key features of hazardous sources from a large number of reported unsafe incident description reports is intuitively important for safety risk identification and assessment.

To identify and classify hazard records more precisely, the International Civil Aviation Organization (ICAO), in its Safety Management Manual (Doc 9859), defines a hazard as a condition or object that has the potential to reduce the safety of personnel, equipment, or support capabilities [3]. The definition of a hazard is broadened in Chinese safety management documents to encompass environmental factors, as well as unstable energy or substances within the operating system [8]. We define hazard as either conditions or objects that have the potential to compromise safe operations (e.g., "poor visibility" and "obstacles"), deriving this definition from the aforementioned documents. Thus, hazard features are explained as words or phrases that generalize hazard information (e.g., both "poor weather" and "bad weather" are extracted as "weather").

Automatic feature extraction based on text aims to extract key information in documents. There are two categories of feature extraction methods based on text data: supervised learning and unsupervised learning [9,10]. The former methods are considered a classification task, which involves learning to differentiate between key and non-key information using labelled data. The feature extraction method based on unsupervised learning utilizes term features in the text (e.g., word frequency [11], part of speech [12], context [13], etc.) to prioritize terms and select vital information. For the hazard feature extraction task, the information that experts care about is not always statistically significant. In aviation operations, hazards are relatively dispersed, and their descriptions vary significantly based on the reporter's department or region. A substantial quantity of description variance diminishes the efficiency of extraction models based on statistical features. Our method incorporates a vast quantity of expert knowledge to classify and standardize the extraction of ATC hazards.

In response to the lack of standardized descriptions for hazard sources, this paper proposes an iterative extraction model to facilitate the accumulation of expert experience and the standardized extraction of hazard features. The model generates candidate features based on phrase co-occurrence. Domain experts, who are not natural language processing specialists, find it challenging to provide effective tuning options for hazard extraction models without concise explanations or methods for interacting with results. This paper employs Euclidean distance to assess semantic similarity between key features. Using this measurement method, candidate features are clustered and visualized with the k-means algorithm. In this study, 249 key hazard features were extracted, and 2171 similar hazard features were integrated. The paper's primary contributions are:

- Based on an iterative approach, we propose a hazard feature extraction method that integrates expert knowledge and refines it.
- We concluded that the k-means algorithm had the highest accuracy by comparing the semantic clustering effects of different algorithms. The clustering results are represented visually to enhance retrieval and analysis efficiency for experts.
- The rest of the paper is structured as follows. In Section 2, the pertinent literature on aviation safety analysis and feature extraction is reviewed. In Section 3, the algorithm of this paper is introduced in detail. In Section 4, we design some experiments and

discuss them accordingly. Finally, some conclusions and directions for future work are drawn in Section 5.

2. Literature Review

2.1. Aviation Safety Analysis

The aviation industry is data-intensive. In the process of secure production, there is an abundance of safety data, such as operational data, incident reports, and radar trajectories. The aviation industry has created a large number of databases for collection and in-depth statistical analysis in order to perpetually maintain and enhance aviation safety and investigate potential safety hazards. With an increase in the quantity or variety of data, data-driven methods are used to analyze safety situations and predict accidents in advance [14]. Text-based machine learning methods are applied to capture the process of accidents, analyze the causes of accidents, and explore the correlation between accidents [15]. Given the large amount of hazard records available, some papers have proposed the feature extraction model, which is used to extract information of hazard records from machine learning. One method involves semantic enhancement, introducing prior knowledge to improve model performance. Pimm et al. augmented basic language analysis with aviation-specific knowledge to extract structural compound terms correctly and replace abbreviations and synonyms, then extract features by association analysis [16]. Tanguy et al. [17] found the "tense problem" in the text, and improved Pimm et al.'s data processing by stemming. In addition, researchers are expanding their study of the annotation of professional terms, synonym substitution, and spelling correction.

Another method explores the relationships between metadata and hazard features through cluster or topic analysis of textual metadata. The primary challenge to be addressed in this context pertains to the process of producing metadata. Ahadh et al. [18] used statistical analysis of various textual components such as capitalization, word frequency, and position to generate metadata. Akhbardeh et al. [19] and Rose et al. [20] dealt with similar metadata by DBSCAN and k-means clustering to facilitate the capture of topic features. The above two types of methods used a large amount of labelled data to enhance precision, but they were unable to explain hazard features.

2.2. Key Feature Extraction

Key feature extraction methods are mainly divided into two categories of research directions: unsupervised and supervised. The term frequency–inverse document frequency (TFIDF) is utilized to assess the significance of terms by taking their frequency and inverse document frequency into account [21]. Campos et al. [22] proposed the YAKE model, which narrows the scope of feature extraction to a single document, using word frequency, case, position, relatedness, and occurrence as the basis for identifying and sorting keywords in the document. The utilization of embedding has garnered interest due to its effectiveness in combining statistical features with contextual data of words or phrases and mapping them into vectors [23]. Sen2vec and Doc2vec implement sentence and document representation based on contextual information [24,25]. Some academics extract important aspects based on document or phrase similarity through vector representations of sentences and documents. Traditional supervised methods are regarded as classification problems aimed at distinguishing between keywords and non-keywords. Zhang et al. [26] classified words or terms in documents into three categories, "good keywords", "indifferent keywords", and "bad keywords", and defined candidate keywords through global and local information. Meng et al. [27] proposed a method for keyword generation based on an encoder–decoder structure that captures the deep semantics of text via deep learning.

Within the professional domain, key feature extraction models focus more on salient attributes and specialized expertise. Liang et al. [28] combined keyword extraction with sampling techniques to resolve imbalanced keyword distribution in policy texts. Through semi-supervised training, Shen et al. [29] integrated expert knowledge and enhanced key feature extraction models for news texts. Another method for integrating expert experience by Zheng et al. [15] employed active learning to map the key features to the event category instead of a single text.

The primary challenges associated with the extraction of hazard features in Chinese research pertain to word segmentation and the absence of labelled datasets. Wang Jiening et al. achieved word segmentation and recognition of specialized vocabulary through the development of a customized dictionary. Additionally, they trained a hazard extraction model by means of manual labeling. As a result of the constraints associated with manual labels, it was possible to extract and recognize only nine features for hazard records [30]. Dong et al. used spelling normalization and stopping word removal to process the Chinese dataset. The research on hazards in China is deficient in hazard description standards, resulting in challenges such as difficult data processing and limited feature information [31].

Because of the stringent safety management and clear communication requirements in ATC, hazard feature extraction prioritizes the integration of expert and domain knowledge. However, standardized descriptions have rarely been considered in the existing research on feature extraction of ATC hazards. Regarding the extraction of features from hazard records, we propose a method based on domain knowledge and expert knowledge to extract standard hazard features.

3. Methodology

The ATC hazard feature extraction process is shown in Figure 1. The methodology is divided into three iterative training steps: candidate feature generation, features vectorization, and standardized description. The generation of candidate features follows the latest iteration of expert knowledge, including extraction suggestions and classification suggestions. During feature vectorization, the goal is to design a mapping method that preserves the distribution and semantics of features. Subsequently, the standardized description step utilizes the k-means algorithm for semantic clustering, and the results are visualized to enhance the transparency and readability of algorithm outcomes. The iterative method adjusts the generation of subsequent candidate features and feature vectorization based on the results of each clustering (including accuracy and error information).

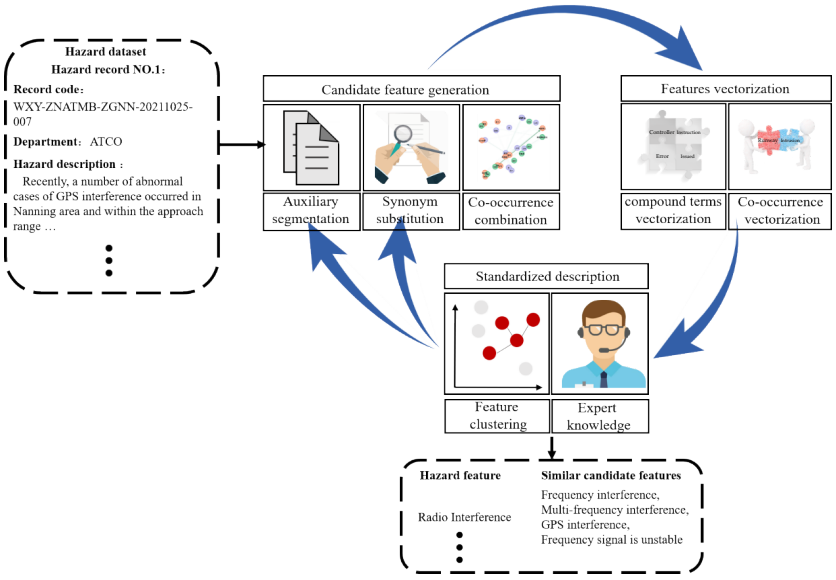

Figure 1. Flowchart of hazard features extraction model.

3.1. Candidate Feature Generation

Features of a hazard are not simply the most specific word or phrase, as features are (at least partially) intended for the topic of a hazard. Expertise informed the clarification and standardization of terminology in this section. Then, we combined the standard terms based on the co-occurrence window to generate candidate hazard features (e.g., "runway"/"invasion", "crew"/"deviation from route").

1. Auxiliary segmentation. Because of the absence of "space" between words, Chinese texts should be segmented into words. Word segmentation refers to the semantic decomposition of text into individual words or phrase sets. In order to enhance comprehension of the word segmentation model, this section uses expert knowledge to construct a segmentation dictionary, including professional terms (e.g., "runway intrusion"), compound words (e.g., "primary system"), standardized features (e.g., "complex weather"), and parts of speech (POS). Matching text and segmentation dictionary achieves word segmentation and POS tagging.

The second step of auxiliary segmentation is to remove stop words, to avoid interference from meaningless information. As shown in Figure 2a,b, stop words, which are common words that are typically filtered out during text processing because they carry little meaningful information (e.g., "and", "the", "is"), accounted for 60% of the word segmentation results. The adjacent segmentation results are recombined and compared to the segmentation dictionary to generate new segmentation results. Recombination enhances the capacity of word segmentation models to produce structurally correct compound words.

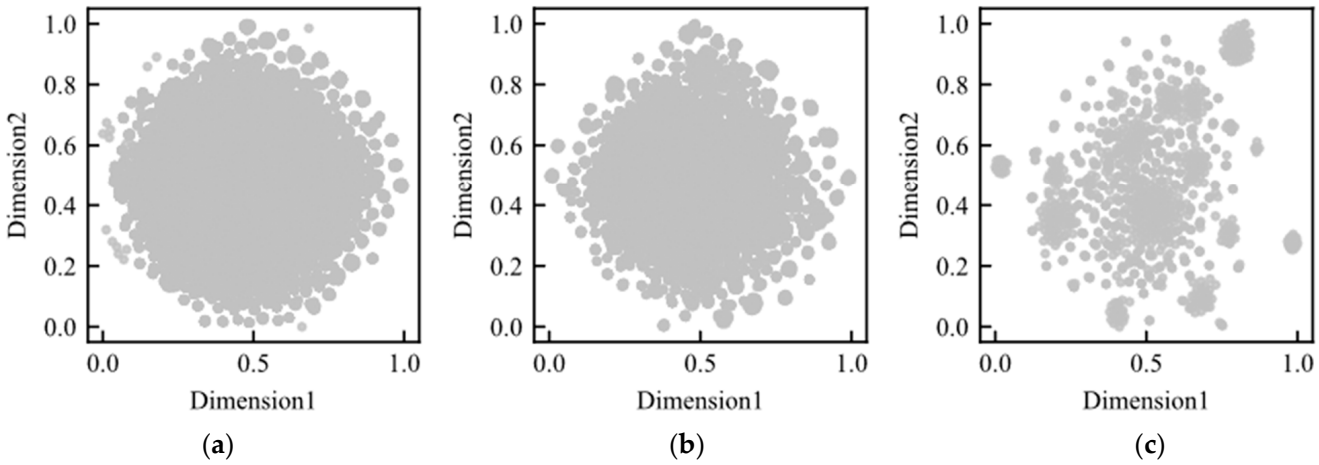

Figure 2. Visualization of specific linguistic processing in CNS. (**a**) The collection of text vectors after auxiliary segmentation; (**b**) the collection of text vectors after the stop word is deleted; (**c**) the set of text vectors after synonym substitution.

2. Synonym substitution. Hazard records are reported by non-specialists, relying on their individual comprehension. Synonym substitution relies on expert knowledge and is more effective at eliminating descriptive ambiguity and text alignment. As illustrated in Figure 2b,c, the synonymously replaced data has more concentrated distribution, enabling the model to focus on more significant and clear information.
3. Co-occurrence combination. Co-occurrence combination combines adjacent items in word segmentation results and generates candidate features. This generation method facilitates the extraction of professional nouns, compound words, and description modes from hazard records. The information content of candidate characteristics keeps growing as more expert knowledge is accumulated. We calculated the frequency of word or phrase co-occurrences within window range. Table 1 shows the

co-occurrence phrases and frequencies of the phrase "runway". The phrases in the co-occurrence window are highly correlated, and the combination contains clear hazard information (e.g., "runway crossing", "runway erroneous entry"). We kept all co-occurrence combinations and output as hazard candidate features to avoid missing low-probability hazards.

Table 1. Statistics of co-occurrence combinations of runways.

	Incursion	Cross	Erroneous Entry	Occupied	Change	Polluted	Expansion	Threshold
runway	40	8	7	5	4	4	3	2

3.2. Feature Vectorization

Word embeddings are used to convert subjective words or phrases into vectors. Self-training and pre-training [32–34] are two methods used to train word embedding models. Self-training word embedding makes use of all contextual relations in the hazard corpus to determine mapping rules between words and vectors. The pre-trained word embedding model is built by training on a large-scale corpus and then fine-tuning on a hazard-specific corpus. The performance of word embeddings trained under low resource conditions is typically inferior because of over-training. As illustrated by the comparison in Figure 3, the word vectors generated by pre-training tend to form more distinct clusters than those generated by self-training.

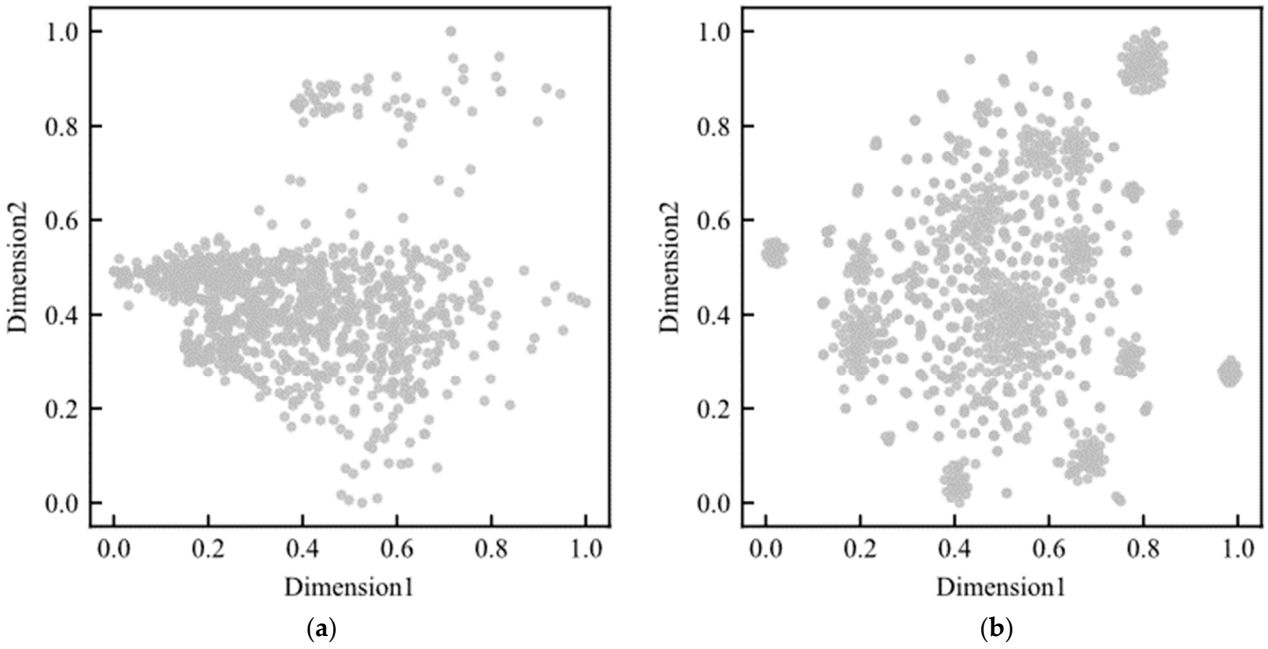

Figure 3. Visualization of word embedding with different training in CNS. (**a**) Self-training embedding; (**b**) pre-training embedding.

The pre-trained model provides initial association rules for words or phrases in the text, but compound terms (e.g., "main landing gear") lack mapping rules. Compound terms consist primarily of nouns and verbs, and their meanings are closely related to their components. We based embedding of compound terms on the semantic and global properties, as shown in Equation (1):

$$v_c = \frac{\sum_{i=1}^{p} \mathrm{tfidf_i} \times \mathrm{v_i}}{\sum_{i=1}^{p} \mathrm{tfidf_i}} \tag{1}$$

where v_c is the compound word vector and its *i-th* component is v_i. $tfidf(w)$ measures the normalized frequency $tf(w)$ and inverse document frequency $idf(w)$ of word w in the document to determine its importance:

$$tfidf(w) = tf(w) \times idf(w) \tag{2}$$

where

$$tf(w) = \frac{f(w)}{\max\{f(w')|w' \in d\}} \tag{3}$$

$$idf(w) = \log \frac{N}{|\{d \in D|w \in d\}|} \tag{4}$$

The normalized frequency $tf(w)$ of a word w in a hazard record d is computed by dividing the original frequency of the word in the hazard record $f(w)$ by its maximum frequency $\max\{f(w')|w' \in d\}$. The vector corresponding to the word w is v. Normalization is intended to ensure that numerous components v_i of lengthy text (where components v_i are more likely to feature multiple times) and short text have the same weight. Inverse document frequency $idf(w)$ measures the specificity of terms based on the ratio between the total number of hazard records N and $|\{d \in D|w \in d\}|$. Equation (4) helps reduce the importance of common words (e.g., "runway", "facilities", etc.).

Given that the distance in vector space represents semantic similarity, vectorization of candidate features must combine the semantics and distribution of co-occurrence of two terms to avoid semantic distortion. Researchers use context-based vector representation methods for learning co-occurrence combination, sentence, or document representations, which ensuring that the resulting vectors have complete semantics (including fundamental word meanings and thematic meanings). We vectorize candidate features (co-occurrence combination representations) by weighted average:

$$v_{ij} = \frac{W_{ij}}{\sum_{i=1}^{n} W_{ij}} \times v_j + \frac{W_{ij}}{\sum_{j=1}^{n} W_{ij}} \times v_i \tag{5}$$

where the candidate feature vectors v_{ij} are composed of v_i and v_j. W_{ij} is the number of co-occurrences of terms i and j. According to the distribution assumption, semantically similar words occur in similar contexts, i.e., they co-occur with the same other words. Equation (5) ensures that terms with co-occurrence distributions have similar weights through normalization, thereby facilitating the extraction of similar features.

3.3. Standardized Description

Standardized description involves experts selecting and classifying candidate hazard features. Candidate feature vectors are clustered based on semantic relationships, and the results are visualized to enhance retrieval efficiency and description precision for expert readers.

K-means is a clustering algorithm widely used in data analysis [35] that evaluates the similarity between vectors by calculating the Euclidean distance between two feature vectors. The algorithm aims at minimizing the distance between each candidate feature

vector and the specified cluster center, and it performs cluster center selection and intra-cluster vector allocation. The algorithm steps are as follows:

Step 1: Randomly select K vector from the set of all candidate features as the initial representative hazard candidate feature vector.

Step 2: Assign the remaining candidate feature vectors to the cluster with the smallest distance to the representative cluster vectors.

Step 3: Calculate E, the sum of the Euclidean distances between the non-representative vectors of each cluster and the representative cluster vectors:

$$E = \sum_{k \in K} \sum_{m \in M_k} Euclidean(v_k, v_m) \tag{6}$$

where $Euclidean(v_k, v_m)$ indicates the Euclidean distance between the representative vector v_k and the intra-cluster vector v_m. M_k represents the number of vectors contained in the cluster k.

Step 4: Calculate the average vector of each cluster and set it as the new representative cluster vector. Repeat steps 2 and 3 to redistribute the intra-cluster vectors and calculate the distance E_r between the intra-cluster vectors and the new cluster center.

Step 5: Replace the cluster center if $E_r - E < 0$.

Step 6: Repeat steps 3, 4, and 5 until the representative vector remains unchanged.

The K-means clustering results of the candidate feature vectors of the CNS dataset are shown in Figure 4. Each point represents a candidate feature, and each color represents a class of key features of the hazard. Figure 5 depicts the prospective features for "Network attack" in Figure 4 in the form of a "word cloud"; that is, the larger the font, the greater the frequency. Safety management experts were allowed to evaluate and categorize candidate features based on their visualizations. The word segmentation dictionary was expanded according to the extraction suggestions of experts (e.g., "Threatening + Network Security" in Figure 5). Security management integrates candidate features with the same semantics and forms classification suggestions (e.g., "threatening + network security" is classified as "Network attack"). Since the clustering results shown in Figure 5 had completed two iterations, some candidate features had sufficient hazard information and did not require a new extraction recommendation (e.g., "Network attacks + operational risks"). This process was iterated until no further extraction and classification suggestions were generated to terminate.

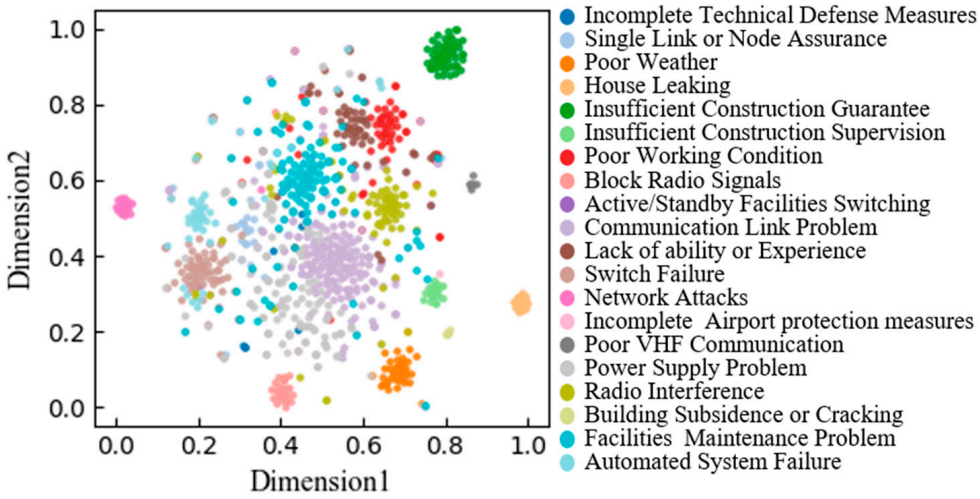

Figure 4. K-means visualization of CNS.

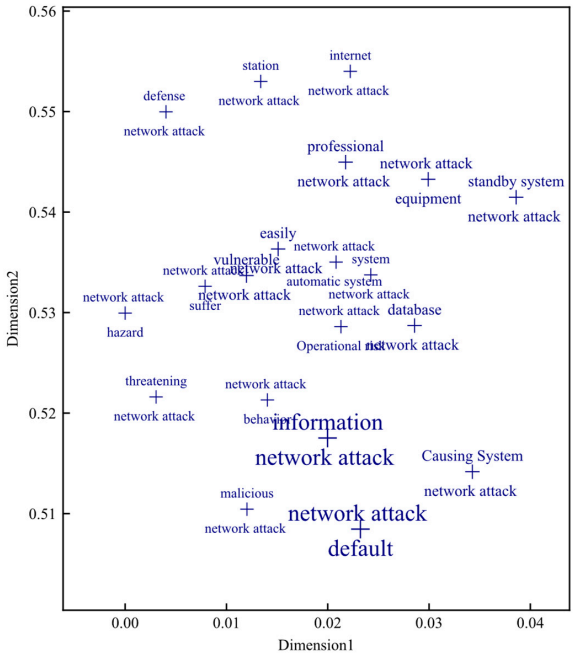

Figure 5. The word cloud of hazard features "network attack".

4. Results and Discussion

In this section, we validate the proposed key feature extraction method for hazards and compare it with other extraction methods. Section 4.1 details our dataset. Section 4.2 presents the relevant pre-training models and conducts comparative validation. Section 4.3 discusses the verification of the k-means algorithm's effectiveness by comparing it with different clustering methods and presents the feature extraction results.

4.1. Hazard Record Dataset

In this study, we extracted key features of hazards based on the 2009–2021 ATC hazard database. The database contains 9146 hazard records. Hazard records include record codes, reporting departments, and hazard information. (1) The hazard number is an index for retrieving hazard information. (2) ATCO, CNS, and meteorology comprise the reporting department, with the first two units contributing more than 85 percent of the data. (3) Hazard information is mainly written on the theme of hazard identification and trigger factors. The distribution of text length for hazard information is predominantly below 100 words, as depicted in Figure 6.

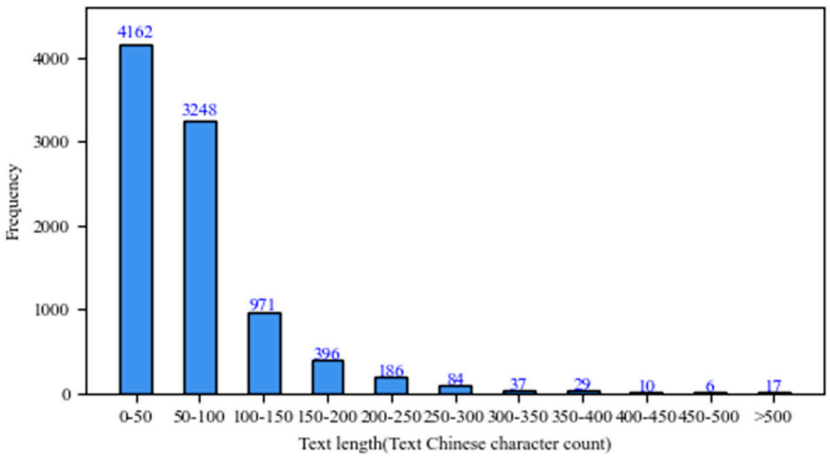

Figure 6. Statistical of hazard record length.

4.2. Cluster Analysis

The pre-trained word embedding model's capacity for semantic analysis has a greater influence on the distribution of candidate feature vectors and clustering tendencies. Figure 7 shows the two-dimensional distribution of word embedding vectors based on the hazard record corpus, Wikipedia corpus, People's Daily corpus, Baidu Encyclopedia corpus, Zhihu Q&A corpus, and merged corpus. To evaluate the feasibility of clustering, we used the Hopkins statistic [36] to analyze the clustering results of different word embedding techniques to compare the local structure of actual and random datasets and identify non-random patterns or outliers as shown in Equation (7). The calculation of the Hopkins statistic is expressed as follows:

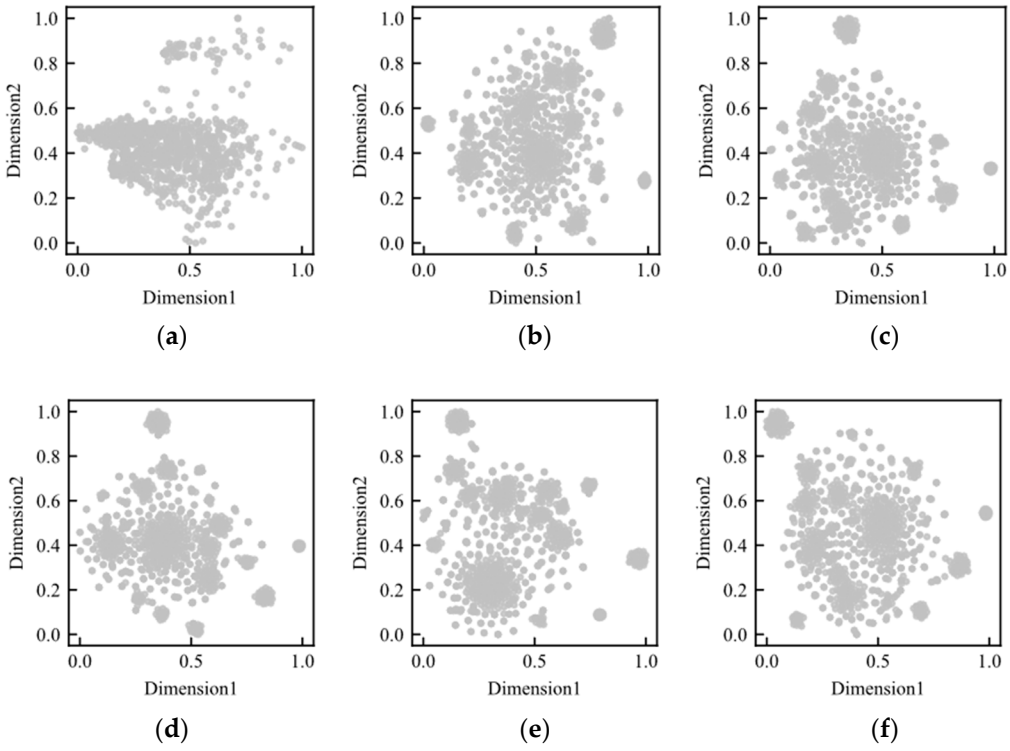

Figure 7. Visualization of word embedding in CNS based on different corpuses. (**a**) Hazard record corpus; (**b**) Wikipedia corpus; (**c**) People's Daily corpus; (**d**) Baidu Encyclopedia corpus; (**e**) Zhihu Q&A corpus; (**f**) merged corpus.

Step 1: Sample uniformly m sampling points $\{p_1, \ldots, p_m\}$ from the candidate feature vector set R.

Step 2: Calculate the distance $x_i = dist(p_i, p_j)$ from each sampling point to the nearest point p_j.

Step 3: Generate a random dataset $random_R$ drawn from a random uniform distribution with $\{p_1, \ldots, p_m\}$ and the same variation as the candidate feature vector set R.

Step 4: On the $random_R$, for each point $q_i \in random_R$, follow steps 1 and 2 to calculate $y_i = dist(q_i, q_j)$.

Step 5: For computing the Hopkins Statistics (H), the formula is defined as follows:

$$H = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} x_i + \sum_{i=1}^{m} y_i} \tag{7}$$

A dataset exhibiting clustering tendency displays a smaller average nearest neighbor distance than a uniform dataset. The meaningfulness of clustering was assessed with a confidence level of 90%, obtaining a confidence space of $H > 0.75$.

Sample datasets were generated from the six word embedding vectors mentioned above, using a sampling ratio of 20%. The Hopkins statistics for these six datasets are presented in Table 2, and the corresponding two-dimensional visualizations are illustrated in Figure 7. Evidently, the vectorization results based on the pre-trained model exhibited a greater tendency to cluster. CNS data had a more concentrated distribution of word embeddings than ATCO data. In the following, we discuss the varieties and complexity of the hazards to explain the different vector distributions in the aforementioned two departments.

Table 2. Hopkins statistic values for different word embedding models.

Profession \ Models	Hazard Record Corpus	Wikipedia Corpus	People's Daily Corpus	Baidu Encyclopedia Corpus	Zhihu Q&A Corpus	Merged Corpus
ATCO	0.678	**0.821**	0.794	0.758	0.762	0.802
CNS	0.723	**0.841**	0.807	0.789	0.812	0.794

Figure 7b shows a clustering tendency resulting from feature vectorization based on the Wikipedia corpus. Despite the fact that same-hazard descriptions are variable, word embeddings mitigate these description differences through semantic mapping, leading to feature vectors that exhibit clustering potential. Each cluster collects candidate features that follow the same semantics or topics. K is determined by the type and quantity of hazards. The accuracy of clustering achieved through the K-means method markedly surpasses that of Gaussian mixture models (GMMs) and density-based spatial clustering of applications with noise (DBSCAN). When compared with datasets labeled by domain experts, K-means clustering demonstrates an impressive range of precision, achieving a peak accuracy of 100% and a lowest observed accuracy of 85%. Figure 8 shows the clustering visualization results by highlighting in blue the correct clustering of network-attack-related hazards by each methodology and in red the misclustering. The three algorithms cluster feature vectors based on probability, density, and distance. The GMM considers the probability distributions for clustering as a combination of vector distributions within the cluster. Cluster distribution fitting results are unstable when lacking feature vectors, so there are ring-shaped error points in Figure 8a. DBSCAN clusters candidate features by setting the lowest density threshold. The uncertainty in vector distances within the cluster limited the establishment of a viable density threshold for vector clustering in Figure 8b.

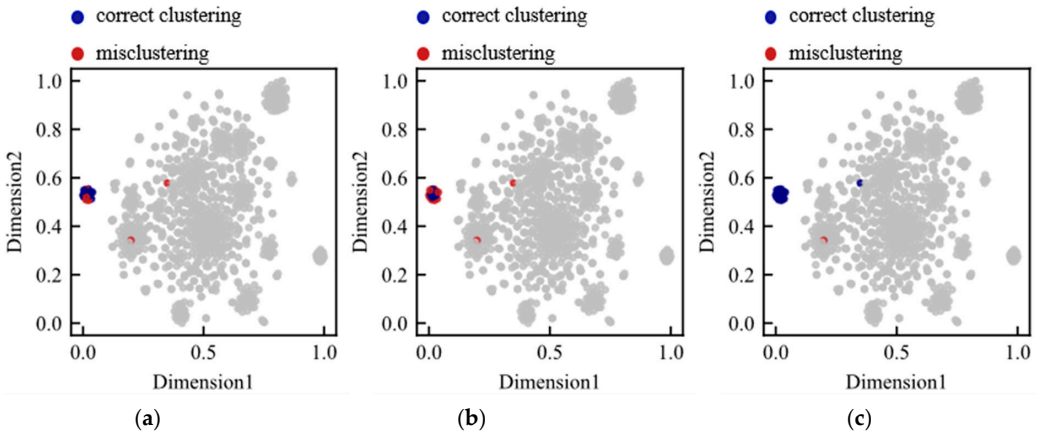

Figure 8. Comparison of hazard features for "network attack" using three different clustering algorithm. (**a**) DBSCAN; (**b**) GMM; (**c**) K-means.

As shown in Figure 8c, the clustering accuracy of the K-means algorithm was higher than that of the other algorithms, and it could discover potential semantic relationships. "worm + attack" is the blue point on the far left, which exhibits a strong semantic association with "network attack" in the high dimension. The position of "worm + attack" shifted as a result of the display of dimensionality reduction. The only incorrectly clustered candidate feature was "attack + switch failure". The switch being attacked is an equipment malfunction result. The model lacks an analysis of chain hazards, but it can generate suggestions for hazard features to prevent error clustering (e.g., semantic equivalent of "attack + switch failure" and "network attack").

4.3. Feature Extraction Results Analysis

Keywords are likely to contain topic-related text phrases, comprising over 80% of key features with more than four Chinese characters. The training corpus included 2544 ATCO hazard records and 1324 CNS hazard records. The study analyzed the efficacy of extracting key features at different levels of domain knowledge. Figure 9 summarizes the hazard feature extraction results for ATCO and CNS, considering specialists without normalization (iter0), standardized once (iter1), standardized twice (iter2), standardized twice without recombination (iter2 + nw), and using a professional dictionary.

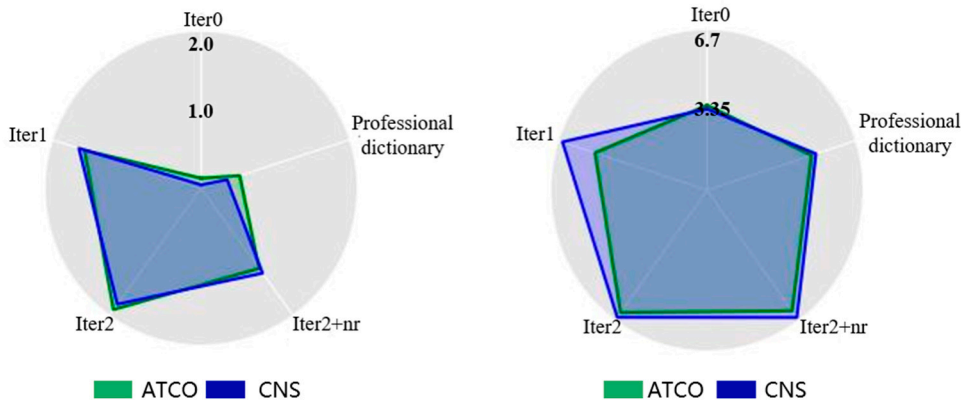

Figure 9. Feature extraction performance under different expert knowledge levels. (**a**) Average number of features; (**b**) average Chinese characters.

The standardization of hazard features was achieved through the accumulation of expert knowledge in this paper. Following standardization twice, 176 categories of hazard features for the ATCO and 72 types of hazard features for the CNS were extracted. We discovered that the most prevalent causes of hazards fell into four categories: human factors, facility factors, management factors, and environmental factors. Hazards are caused by the failure of one or more departments. For example, "illegal handover" may result from personnel violations, management negligence, or the failure of technical defense measures. Therefore, we labelled the main causes of hazards. The hazard features corresponding to the above four causes accounted for 61%, 28%, 7%, and 4%, and there were 742, 464, 162, and 93 similar hazard features, according to the annotation results. Figure 10 illustrates the top fifty features and their causative labels. Human factors exhibited greater complexity than the other three factors. From feature content, man-made hazards were mainly aimed at the behavior or ability of controllers or operation guarantee personnel, which are described in a wider and more varied perspective. The hazard records of the ATCO consist of controller errors, working performance, etc., so additional iterative training was required for the experiment depicted in Figure 9.

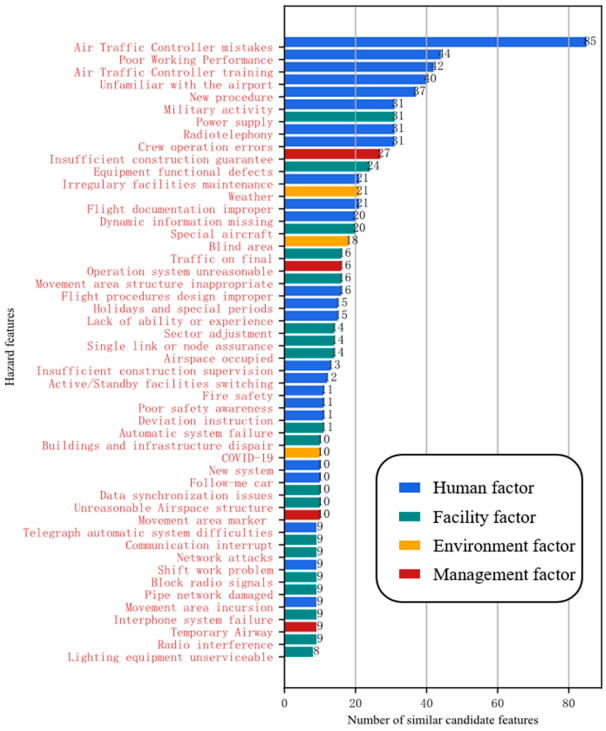

Figure 10. The top 50 features of similar candidate features and their causative factors.

Using expert knowledge, the correlation between candidate features and hazard features was clarified. Moreover, external knowledge improved the precision and scope of model identification beyond common hazard features. To reduce the reading cost of experts, we conducted semantic clustering on the candidate features; some of the results are displayed in Figure 11. Domain experts standardized the description of similar candidate features based on the clustering results; for instance, "insufficient battery capacity" and "abnormal discharge" were extracted as "power failure". Unlike other classification or topic models, our proposed extraction models for hazard features can explicitly explain the extraction basis and provide concise extraction results.

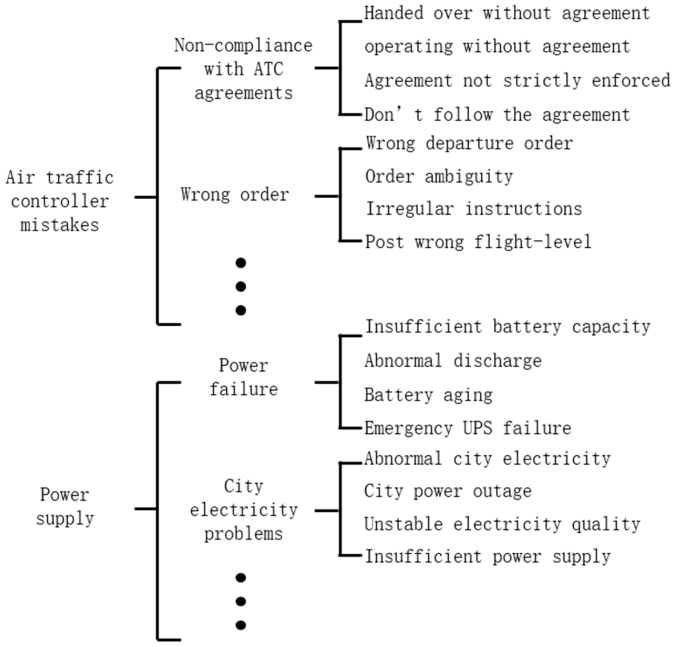

Figure 11. Text display of similar candidate features.

The performance of the algorithm was verified by 636 ATCO hazard records and 302 CNS hazard records in the test set. The longest record contained 343 Chinese characters, the smallest record consisted of 4 Chinese characters, and the mean number of Chinese characters was 67. We compared commonly used unsupervised key feature extraction algorithms as follows:

TFIDF: TFIDF is an importance ranking algorithm. Importance is calculated by word frequency and specificity [37].

Text-Rank: Text-Rank is a graph-based key information retrieval algorithm. Importance ranking is based on co-occurrence between terms [38].

YAKE: YAKE is also an importance ranking algorithm based on case, context information, position, word frequency, and occurrence. Since there is no case in Chinese, we replaced it with bold characters in the experiment [22].

Keybert: Keybert is a self-supervised algorithm. The model defines and ranks the topic similarity of each phrase through context [39].

Table 3 demonstrates that our method outperformed other methods in terms of accuracy and feature information. Hazard features were not necessarily the most statistically significant word or term, but more importantly, they fit the definition and topic. The high extraction accuracy of Keybert and YAKE! indicated the importance of contextual semantic information for hazard feature extraction. For example, the word "power" could mean an engine (e.g., auxiliary power unit) or power supply (e.g., uninterruptible power supply) in the CNS department. Our method defines and interprets terms through expert experience and domain knowledge. Similarity ranking is utilized by both our method and the Keybert algorithm. However, we compare the similarity between candidate features and hazard feature corpus, as opposed to a certain hazard record. Finally, because the complexity of controller behavior is greater than that of equipment failure, the hazard feature extraction of in ATCO requires more domain knowledge support.

Table 3. The performance results of comparison methods and our methods.

Dataset	Evaluation Index	Methods				
		TFIDF	**Text-Rank**	**YAKE!**	**Keybert**	**Ours**
ATCO	Accuracy	45.9%	52.0%	64.8%	69.3%	82.4%
CNS		58.3%	57.0%	67.5%	74.8%	86.7%
ATCO	Information (Chinese characters)	2.82	2.79	3.47	4.92	5.78
CNS		2.87	2.93	3.62	4.65	6.31

5. Conclusions

This paper proposes a feature extraction model for ATC hazards designed to effectively summarize critical hazard information. Model training employs an iterative method to integrate and accrue expert knowledge among the three processing modules of candidate feature generation, feature vectorization, and standardized description. In addition, special language processing methods, including recombination and pre-trained embeddings, are proposed to help frame the capture of correctly structured compound words and the establishment of vector mapping rules. The iterative training results yielded a list of hazard features that described both the ATCO and CNS datasets, along with the terms or candidate features most closely related to each hazard feature.

Overall, this study highlights the necessity of hazard-definition-based feature extraction in safety assurance work, which allows deeper safety assessment of ATC operations. Through expert knowledge, the candidate features can analyze the record author's understanding of the hazard in a more comprehensive manner and convert the abstract text expression into a term list containing key information and the correct structure by segment-

ing auxiliary words and replacing synonyms. The analysis of candidate features by experts is a process of delineating and explicitly standardizing hazards. These normalization proposals can regulate the generation of candidate features and eradicate the detrimental effect of subjective expressions on semantic comprehension. By analyzing co-occurring and valuable term combinations, experts are able to identify unnoticed hazards or operational hazards. Finally, the model undergoes multiple rounds of training to establish a mapping correlation between terms and vectors, utilizing pre-training embeddings and the current corpus of the database.

The experimental results showed that: (1) in the ATCO and CNS datasets, the hazard feature extraction model based on expert knowledge achieved accuracies of 82.4% and 86.7%, respectively, significantly outperforming other extraction algorithms. This demonstrates that extensive domain knowledge can effectively enhance feature extraction accuracy; (2) comparing feature extraction results across different levels of expert knowledge showed that expanding corpus knowledge allowed the model to autonomously identify more significant and comprehensive features; (3) the disparity in experimental performance between the two datasets indicates that the model requires more comprehensive expert knowledge to support the training of complex, abstract, and discrete corpora. The experimental results provide insights for subsequent accident analysis and feature extraction to increase accuracy by expanding domain knowledge or adding pertinent corpora, which coincides with the large-scale models.

Author Contributions: Conceptualization, X.T. and Z.C.; methodology, Z.Q. and W.Z.; investigation, Z.C. and X.T.; data curation, Z.Q. and X.T.; supervision, W.Z.; validation, Z.C. and X.T.; writing—original draft preparation, Z.Q. and X.T.; writing—review and editing, Z.C. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China [No.2022YFB2602403] and the Fundamental Research Funds for the Central Universities [No. NS2023036].

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author, Zeng, upon reasonable request.

Conflicts of Interest: Author Zhipeng Cai was employed by the company Shanghai Aircraft Test Flight Engineering Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Civil Aviation Administration of China. *Safety Report*; Civil Aviation Administration of China: Beijing, China, 2023.
2. Civil Aviation Administration of China. *Civil Aviation Development Statistical Bulletin*; Civil Aviation Administration of China: Beijing, China, 2021.
3. Federal Aviation Administration. *Safety Management System*; Federal Aviation Administration: Washington, DC, USA, 2018.
4. Cao, Y.; Wang, X.; Yang, Z.; Wang, J.; Wang, H.; Liu, Z. Research in marine accidents: A bibliometric analysis, systematic review and future directions. *Ocean. Eng.* **2023**, *284*, 115048. [CrossRef]
5. Li, Y.; Guldenmund, F.W. Safety management systems: A broad overview of the literature. *Saf. Sci.* **2018**, *103*, 94–123. [CrossRef]
6. Saleh, J.H.; Marais, K.B.; Bakolas, E.; Cowlagi, R.V. Highlights from the literature on accident causation and system safety: Review of major ideas, recent contributions, and challenges. *Reliab. Eng. Syst. Saf.* **2010**, *95*, 1105–1116. [CrossRef]
7. National Air and Space Administration. *Aviation Safety Reporting System*; National Air and Space Administration: Moffett Field, CA, USA, 2022.
8. Patriarca, R.; Di Gravio, G.; Cioponea, R.; Licu, A. Safety intelligence: Incremental proactive risk management for holistic aviation safety performance. *Saf. Sci.* **2019**, *118*, 551–567. [CrossRef]
9. Firoozeh, N.; Nazarenko, A.; Alizon, F.; Daille, B. Keyword extraction: Issues and methods. *Nat. Lang. Eng.* **2020**, *26*, 259–291. [CrossRef]

10. Onan, A.; Korukoğlu, S.; Bulut, H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Syst. Appl.* **2016**, *57*, 232–247. [CrossRef]
11. Witten, I.H.; Paynter, G.W.; Frank, E.; Gutwin, C.; Nevill-Manning, C.G. KEA: Practical automatic keyphrase extraction. In Proceedings of the Fourth ACM Conference on Digital Libraries, Berkeley, CA, USA, 11–14 August 1999; pp. 254–255.
12. Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 216–223.
13. Turney, P.D. Coherent keyphrase extraction via web mining. *arXiv* **2003**, arXiv:cs/0308033.
14. Ledvinka, M.; Lališ, A.; Křemen, P. Toward data-driven safety: An ontology-based information system. *J. Aerosp. Inf. Syst.* **2019**, *16*, 22–36. [CrossRef]
15. Zhang, X.; Mahadevan, S. Ensemble machine learning models for aviation incident risk prediction. *Decis. Support Syst.* **2019**, *116*, 48–63. [CrossRef]
16. Robinson, S.D. Temporal topic modeling applied to aviation safety reports: A subject matter expert review. *Saf. Sci.* **2019**, *116*, 275–286. [CrossRef]
17. 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]
18. 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]
19. 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]
20. Akhbardeh, F.; Desell, T.; Zampieri, M. NLP tools for predictive maintenance records in MaintNet. In Proceedings of the 10th International Joint Conference on Natural Language Processing: System Demonstrations, Patna, India, 18–21 December 2020; pp. 26–32.
21. Jing, L.P.; Huang, H.K.; Shi, H.B. Improved feature selection approach TFIDF in text mining. In Proceedings of the IEEE International Conference on Machine Learning and Cybernetics, Beijing, China, 4–5 November 2002; Volume 2, pp. 944–946.
22. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* **2020**, *509*, 257–289. [CrossRef]
23. Wang, B.; Wang, A.; Chen, F.; Wang, Y.; Kuo, C.C.J. Evaluating word embedding models: Methods and experimental results. *APSIPA Trans. Signal Inf. Process.* **2019**, *8*, e9. [CrossRef]
24. Li, R.; Zhao, X.; Moens, M.F. A Brief Overview of Universal Sentence Representation Methods: A Linguistic View. *ACM Comput. Surv.* **2022**, *55*, 1–42. [CrossRef]
25. Singh, K.; Devi, H.M.; Mahanta, A.K. Document representation techniques and their effect on the document Clustering and Classification: A Review. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 1780.
26. Zhang, K.; Xu, H.; Tang, J.; Li, J. Keyword extraction using support vector machine. In *Advances in Web-Age Information Management: 7th International Conference, WAIM 2006, Hong Kong, China, 17–19 June 2006*; Springer: Berlin/Heidelberg, Germany, 2006.
27. Meng, R.; Zhao, S.; Han, S.; He, D.; Brusilovsky, P.; Chi, Y. Deep keyphrase generation. *arXiv* **2017**, arXiv:1704.06879.
28. Liang, D.; Yi, B.; Cao, W.; Zheng, Q. Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and SMOTE. *Expert Syst. Appl.* **2022**, *188*, 116051. [CrossRef]
29. Shen, X.; Wang, Y.; Meng, R.; Shang, J. Unsupervised deep keyphrase generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11303–11311.
30. Xiong, M.; Wang, H.; Wong, Y.D.; Hou, Z. Enhancing aviation safety and mitigating accidents: A study on aviation safety hazard identification. *Adv. Eng. Inform.* **2024**, *62*, 102732. [CrossRef]
31. Jiao, Y.; Dong, J.; Han, J.; Sun, H. Classification and causes identification of Chinese civil aviation incident reports. *Appl. Sci.* **2022**, *12*, 10765. [CrossRef]
32. Luan, Y.; Watanabe, S.; Harsham, B. Efficient learning for spoken language understanding tasks with word embedding based pre-training. In *INTERSPEECH*; Mitsubishi Electric Research Laboratories, Inc.: Cambridge, MA, USA, 2015; pp. 1398–1402.
33. Isotani, H.; Washizaki, H.; Fukazawa, Y.; Nomoto, T.; Ouji, S.; Saito, S. Duplicate Bug Report Detection by Using Sentence Embedding and Fine-tuning. In Proceedings of the IEEE International Conference on Software Maintenance and Evolution, Luxembourg, 24 November 2021; pp. 535–544.
34. Liu, Y.; Maier, W.; Minker, W.; Ultes, S. Empathetic Dialogue Generation with Pre-trained RoBERTa-GPT2 and External Knowledge. In Proceedings of the 12th International Workshop on Spoken Dialogue System Technology, Singapore, 28 October 2022; pp. 67–81.
35. Xiong, C.; Hua, Z.; Lv, K.; Li, X. An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers. In Proceedings of the IEEE International Conference on Cloud Computing and Big Data, Macau, China, 16–18 November 2016; pp. 265–268.

36. Zhang, R.; Miao, Z.; Tian, Y.; Wang, H. A novel density peaks clustering algorithm based on Hopkins statistic. *Expert Syst. Appl.* **2022**, *201*, 116892. [CrossRef]
37. Qu, Z.; Song, X.; Zheng, S.; Wang, X.; Song, X.; Li, Z. Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification. In Proceedings of the IEEE International Conference on Big Data and Smart Computing, Shanghai, China, 15–17 January 2018; pp. 677–680.
38. Guo, W.; Wang, Z.; Han, F. Multifeature Fusion Keyword Extraction Algorithm Based on TextRank. *IEEE Access* **2022**, *10*, 71805–71813. [CrossRef]
39. Devika, R.; Vairavasundaram, S.; Mahenthar, C.S.J.; Varadarajan, V.; Kotecha, K. A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data. *IEEE Access* **2021**, *9*, 165252–165261. [CrossRef]