

中文引用格式:田泽,罗帆. 基于多源文本挖掘的飞行事故关键风险识别[J]. 中国安全科学学报, 2025, 35(8): 156–163.

英文引用格式:TIAN Ze, LUO Fan. Key risk identification of flight accidents based on multi-source text mining[J]. China Safety Science Journal, 2025, 35(8): 156–163.

基于多源文本挖掘的飞行事故关键风险识别^{*}

田泽, 罗帆 教授

(武汉理工大学 管理学院, 湖北 武汉 430070)

中图分类号:X949

文献标志码:A

DOI: 10.16265/j.cnki.issn1003-3033.2025.08.1120

基金项目:国家社会科学基金重大项目资助(23ZDA117)。

【摘要】 为提升我国民用航空飞行安全风险管控效能,精准识别引发飞行事故的关键风险因素,以微博资讯、新闻报道、航空事故调查报告为样本,开展飞行事故风险多源文本挖掘,并借助基于双向编码器的主题模型(BERTopic)识别飞行事故风险因素;运用 Word2Vec 模型和复杂网络解析风险因素的语义关联,确定关键风险因素;采用双向编码器(BERT)挖掘出引发公众最严重负面情绪的人员风险因素,并结合词频统计方法辨识低空空域的关键人员风险因素。结果表明:人员风险是导致飞行事故的关键风险,其中未严格执行操作程序、鸟击和机组成员疾病是影响飞行安全的关键风险因素;飞行机组人员未严格执行操作程序引发的公众负面情绪最为强烈,也是低空空域事故的关键人员风险因素。

【关键词】 多源文本挖掘; 飞行事故; 关键风险; 负面情绪; 低空空域

Key risk identification of flight accidents based on multi-source text mining

TIAN Ze, LUO Fan

(School of Management, Wuhan University of Technology, Wuhan Hubei 430070, China)

Abstract: In order to enhance the risk control efficiency of civil aviation flight safety in China and accurately identify the key risk factors that cause flight accidents, multi-source text mining of flight accident risks was conducted using Weibo information, news reports, and aviation accident investigation reports as samples. Flight accident risk factors were identified using the bidirectional encoder representations from transformers for topic modeling(BERTopic). The semantic correlation of risk factors was analyzed using the Word2Vec model and complex network, from which the key risk factors were determined. The bidirectional encoder representations from transformers(BERT) was adopted to mine the personnel risk factors that trigger the most serious negative emotions among the public. The key personnel risk factors in low-altitude airspace were identified through word frequency statistics. The results indicate that personnel risk is the key risk factor leading to flight accidents. Among them, failure to strictly follow operating procedures, bird strikes, and sickness of flight crew are the key risk factors affecting flight safety. The non-strict implementation of operation procedures by flight crew not only arouses the most negative public emotions but also constitutes the key personnel risk factor leading to low-altitude airspace accidents.

Keywords: multi-source text mining; flight accidents; key risk; negative emotions; low altitude airspace

0 引言

2024 年末,全球范围内连续发生多起航空事故,如阿塞拜疆客机坠毁事故、济州航空撞鸟事件^[1]。飞行事故不仅严重威胁乘客的生命财产安全,还会加剧公众对航空飞行的负面情绪反应。随着我国民航客运和货运量的稳步攀升,低空空域的开放范围逐步扩大,航空不安全事件的发生概率也在相应增加。因此,准确识别引发飞行事故的关键风险因素,对保障民用航空运营安全,具有至关重要的意义。

目前,民航事故致因研究已取得诸多成果。黄静洋等^[2]依据通航规章政策,剖析中低空航空飞行计划的风险因素;许保光等^[3]借助贝叶斯网络,探究不同风险因素对航空不安全事件的影响程度。袁乐平等^[4]基于民航事故征候报告,探寻潜在安全风险隐患。王红^[5]和刘俊杰^[6]等依据航空事故调查报告,解析事故风险的因果关系。YUAN Leping 等^[7]基于航空安全风险源,提出风险识别模型。然而,现有研究在飞行事故风险因素识别方面,多依赖于单一数据源,如事故调查报告或特定案例,难以全

面覆盖事故风险因素。

鉴于此,笔者拟融合飞行事故报告、社交媒体信息和新闻报道等文本数据进行多源文本挖掘,通过基于双向编码器的主题模型 (Bidirectional Encoder Representations from Transformers for Topic Modeling, BERTopic) 识别飞行事故的人机环管风险因素,利用 Word2Vec 模型计算风险关键词的词向量,并基于语义相似度解析风险因素间的关联关系,借助复杂网络确定关键的风险因素,结合双向编码器 (Bidirectional Encoder Representations from Transformers, BERT) 计算视频评论情感值,识别关键人员风险因素,并采用词频统计方法确定低空空域飞行事故的关键风险因素,以期降低风险控制成本,提升航空安全水平。

1 飞行事故关键风险识别模型构建

选取 2010 年—2024 年的 1970 份飞行事故报告、事故追踪报告及新闻资讯等文本数据开展多源文本挖掘,基于 BERTopic 模型、Word2vec 模型、复杂网络和 BERT 模型构建风险识别模型。基于多源文本挖掘的飞行事故关键风险识别流程如图 1 所示。

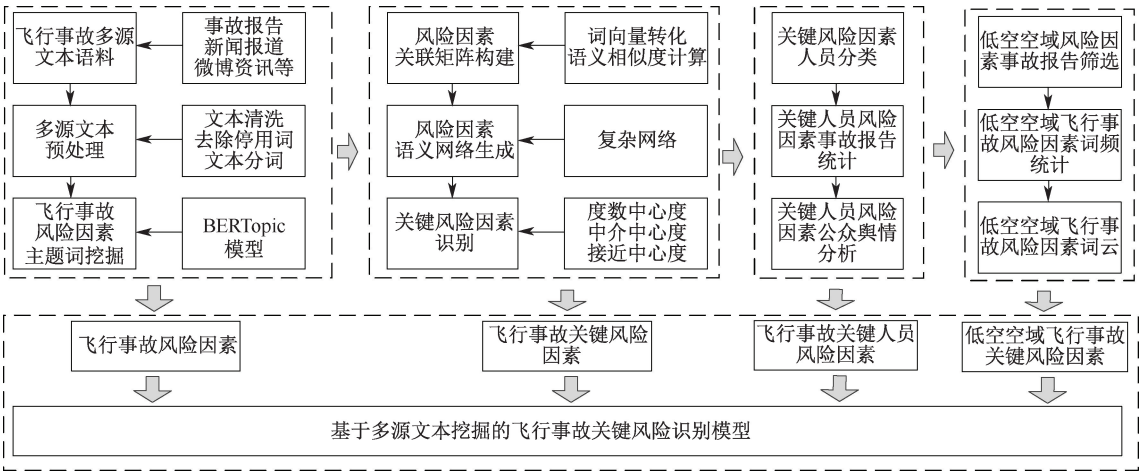


图 1 基于多源文本挖掘的飞行事故关键风险识别流程

Fig. 1 Identification process of key risk in flight accidents based on multi-source text mining

1.1 飞行事故风险因素识别模型

建立停用词库,过滤掉多源文本中的无关干扰信息,采用 BERTopic 模型提取出与飞行事故风险相关的语义信息。BERTopic 模型基于双向 Transformer 架构抽取文本特征,结合降维算法 (Uniform Manifold Approximation and Projection, UMAP) 和聚类算法 (Hierarchical Density-Based Spatial Clustering of Applications with Noise, HDBSCAN) 处理文本数据,再运用加权词频-逆文

档频率模型 (c-Term Frequency-Inverse Document Frequency, c-TF-IDF) 量化主题关键词的重要性^[8]。

BERTopic 模型难以直接揭示风险因素间的复杂关系,而 Word2Vec 模型可将风险主题词映射为词向量 λ_i, λ_j ,通过计算 d 维分布式词向量 $\lambda_i = (x_{i1}, \dots, x_{id}), \lambda_j = (x_{j1}, \dots, x_{jd})$ 之间的语义相似度,表征风险因素间的语义关联。语义相似度的计算如下:

$$\cos(\lambda_i, \lambda_j) = \frac{\sum_{k=1}^d x_{ik} x_{jk}}{\sqrt{x_{ik}^2} \sqrt{x_{jk}^2}} \tag{1}$$

式中: x_{ik} 为 λ_i 在第 k 维上的值; x_{jk} 为 λ_j 在第 k 维上的值^[9]。

构建 u 个风险因素相关度矩阵 $\boldsymbol{H}$:

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1u} \\ h_{21} & h_{22} & \cdots & h_{2u} \\ \vdots & \vdots & & \vdots \\ h_{u1} & h_{u2} & \cdots & h_{uu} \end{bmatrix} \quad (2)$$

式中: $\boldsymbol{H}$ 为不同事故风险因素的相关度矩阵; h_{ij} 为第 i 个和第 j 个因素的相关度。

1.2 飞行事故关键风险因素识别模型

参照文献[10],从度数中心度、中介中心度和接近中心度3个维度,识别导致飞行事故的关键风险因素。同时,通过计算飞行事故视频评论情感值,辨识出引发公众负面情绪最强烈的关键人员风险因素。

度数中心度反映与某一飞行事故风险因素相连的其他因素的数量,计算如下:

$$D_i = \frac{m_i}{N-1} \quad (3)$$

式中: D_i 为因素 i 的度数中心度; m_i 为现有与因素 i 相连的边的数量; $N-1$ 为因素 i 与其他因素相连的边的数量。

中介中心度为某个风险因素在任意2个其他风险因素之间的所有最短路径上的出现次数,计算如下:

$$B_i = \sum_{j < r} \frac{g_{jr}(i)}{g_{jr}} \quad (4)$$

式中: B_i 为因素 i 的中介中心度; g_{jr} 为飞行事故风险因素 j 到 r 的所有最短路径数; $g_{jr}(i)$ 为风险因素 j 到风险因素 r 经过因素 i 的最短路径数。

接近中心度为网络中某个风险因素到其他因素距离(最短路径)的倒数,计算如下:

$$C_i = \left(\sum_{j=1}^v d_{ij} \right)^{-1} \quad (5)$$

式中: C_i 为因素 i 的接近中心度; d_{ij} 为2个风险因素

间的最短路径; v 为关联网络中风险因素的总数。

为量化公众对飞行事故的情感态度,运用BERT模型将飞行事故视频评论转换为 embedding 向量,通过Transformer编码处理,并添加 softmax 线性分类器,将文本特征向量转换为(0,1)的情感评分,计算如下:

$$P = \text{softmax}(\boldsymbol{W}\boldsymbol{S} + b) \quad (6)$$

式中: P 为情感评分; $\boldsymbol{W}$ 为事故文本权重矩阵; $\boldsymbol{S}$ 为特征提取层的输出向量; b 为飞行事故文本偏置^[11]。

2 飞行事故关键风险识别与分析

2.1 飞行事故风险因素关联关系挖掘

飞行事故多源文本挖掘的目的是尽可能全面获取飞行事故风险因素。为确保提取结果的可靠性,采用主题一致性评价识别结果。潜在狄利克雷分布模型(Latent Dirichlet Allocation, LDA)和BERTopic模型识别飞行事故风险因素的主题一致性结果分别为0.375和0.415。主题一致性得分越高,表明主题语义关联越强,因此,BERTopic模型的识别结果更为可靠。同时,BERTopic模型挖掘出27个主题,依据SHEEL模型将其划分为人、机、环、管4大类。基于多源文本的飞行事故风险因素见表1。

为分析飞行事故风险因素间的潜在关联,选取与主题内涵最接近的主题词,运用Word2Vec模型计算其词向量,并依据式(1)计算不同主题间的相关度,进而通过式(2)得出相关度矩阵。飞行事故风险因素语义相似度计算结果见表2。

2.2 飞行事故关键风险因素识别

2.2.1 基于度数中心度的飞行事故关键风险因素识别

飞行事故风险因素关联网络的度数中心度如图2所示,由图2可知:未严格执行操作程序、机组成员疾病和鸟击等风险因素具有较高的度数中心度,相较于其他飞行事故风险因素更为关键。

表1 基于多源文本的飞行事故风险因素

Table 1 Flight accident risk factors based on multi-source texts

飞行事故风险	飞行事故风险因素	部分主题关键词	飞行事故风险	飞行事故风险因素	部分主题关键词
人员风险 A	未严格执行操作程序 A_1	频率、管制、夜班、管制员、扇区、航班、信号、机场、报告、值夜班、离场、程序	环境风险 C	鸟击 C_1	鸟类、鸟击、发动机、驱鸟、叶片、飞鸟、风扇、鸟群、动物、鸟害、威胁、栖息地
	机组成员疾病 A_2	机长、医生、医疗、症状、血压、袖带、驾驶舱、救助、乘务员、身体、酒精、医药		空中湍流 C_2	事故、机长、机组、报告、机场、飞行员、高度、遭遇、数据、湍流、气象、航班

续表 1

飞行事故风险	飞行事故风险因素	部分主题关键词	飞行事故风险	飞行事故风险因素	部分主题关键词
人员风险 <i>A</i>	管制员工作失误 A_3	管制员、间隔、指令、冲突、飞机、航班、管制、咨询、机组、高度、航路、航空	环境风险 <i>C</i>	地面强光干扰 C_3	灯光、荧光屏、机场、光源、红灯、目视、亮度、闪光灯、反射光、强光、标志灯
	飞行员驾驶经验不足 A_4	速度、空速、飞机、机长、驾驶仪、襟翼、高度、油门、飞行员、机组、迎角、慢车		能见度较低 C_4	目视、航空器、高度、飞机、交通、区域、管制、流量、冲突、视线、信标台、意识
	操作失误 A_5	直升机、作业、旋翼、事发、高压线、尾桨、飞行员、调查组、航空器、事故、坠地	管理风险 <i>D</i>	航图标识不准确 D_1	定位点、航图、航路、航线、错误、页面、名称、路线、偏差、指令、航向、标识
	忽视警告 A_6	警告、高度、飞机、空速、地形、程序、机组、机长、机场、仪表、目视、飞行员		空中管理不足 D_2	空域、管制、航空器、冲突、指令、进场、距离、高度、离场、间隔、报告、航路
	机组成员安全意识淡薄 A_7	烟雾、气味、驾驶舱、空调、面罩、氧气、机组人员、报告、发动机、检查、异味		飞机配载不当 D_3	数据、错误、重量、检查、速度、飞机、参数、标识符、检测、燃油、高度限制
	没有遵照空中交通管制指令 A_8	管制员、进场、空域、终端、速度、高度、空速、程序、指令、定位点、管制、报告		航行计划执行不当 D_4	航路、航线、标准、路线、航向、间隔、滑行道、计划、区域、机务、航迹、报告
	故障排除能力不足 A_9	发动机、叶片、燃油、涡轮、检查、风扇、整流罩、高压、裂纹、故障、碎片、滑油		安全管理不到位 D_5	事故、猜测、失联、报告、失控、调查结果、黑匣子、信息管理、媒体报道、情报
	管制员情景意识不足 A_{10}	扇区、高度、航路、进场、飞行高度、飞机、飞行员、空域、余度、作业、障碍物		监督检查力度不够 D_6	人员、培训、危险品、经营、规章、机务、危险源、报告、局方、负责人、航空安全
	设备维护不足 A_{11}	发电机、故障、总线、电缆、检查单、升降舵、报告、面板、机组、步骤、线缆		培训不足 D_7	学员、教员、训练、事发、坠机、油门、调查组、燃油、动作、培训、降落伞、坠机
	引擎故障 B_1	引擎、客机、直升机、航空、空域、机上、飞机、事故、故障、航班、失联、坠机		旅客不当行为 D_8	电池、锂电池、手机、短路、失控、电子设备、锂离子、电流、电压、火灾、危险品
	控制系统故障 B_2	压力、氧气、活门、面罩、客舱、高度、外流、空调、故障、控制器、舱压、警告	机械设备风险 <i>B</i>	通信设备故障 B_4	信号、监护、频率、指挥棒、偏差、障碍物、传输、频段、终端、导航台、对讲机
	机械磨损故障 B_3	液压、起落架、线缆、软管、故障、油量、齿轮箱、液压油、扰流板、组件、壳体			

表 2 飞行事故风险因素语义相似度部分计算结果
Table 2 Partial calculation results of semantic similarity of flight accident risk factors

飞行事故风险因素	空中湍流	设备维护不足	引擎故障
空中湍流	1	-0.059 424 3	0.461 055 6
设备维护不足	-0.059 424 3	1	-0.213 631 6
引擎故障	0.461 055 6	-0.213 631 68	1

基于度数中心度的风险因素的关联网路如图 3 所示,由图 3 可知:机组成员疾病等因素与更多因素相连,对飞行事故的发生具有较大影响。

2.2.2 基于中介中心度的飞行事故关键风险因素识别

飞行事故风险因素关联网路的中介中心度如图 4 所示,由图 4 可知:旅客不当行为、鸟击等风险因素具有较高的中介中心度,在关联网路中占据了关键的中介位置,起到更为关键的信息传递作用。

基于中介中心度的风险因素关联网路如图 5 所示,由图 5 可知:旅客不当行为处于多个风险因素间

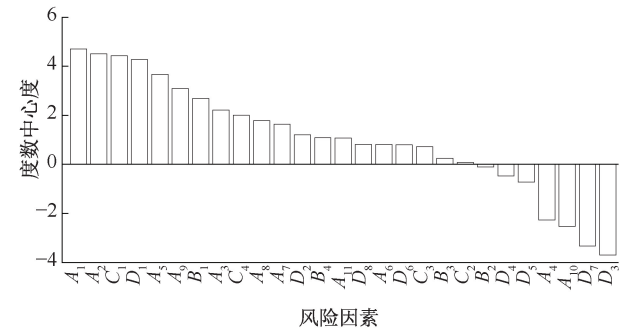


图2 飞行事故风险因素关联网络度数中心度
Fig.2 Flight accident risk factor correlation network degree centrality

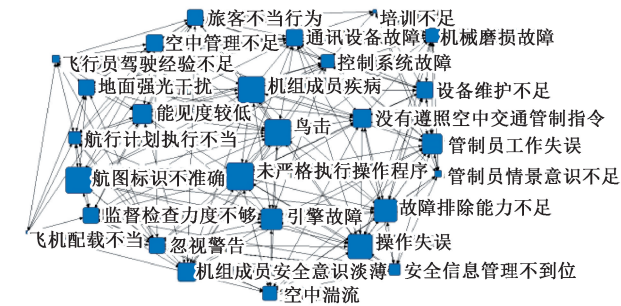


图3 基于度数中心度的风险因素关联网络
Fig.3 Correlation network of risk factors based on degree centrality

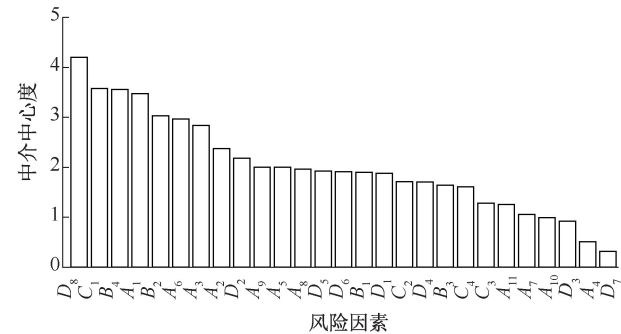


图4 飞行事故风险因素关联网络中介中心度
Fig.4 Flight accident risk factor correlation network intermediary centrality

的最短路径上,将加快风险信息的传递速率。

2.2.3 基于接近中心度的飞行事故关键风险因素识别

飞行事故风险因素关联网络的接近中心度如图6所示,由图6可知:未严格执行操作程序、鸟击、机组成员疾病和故障排除能力不足等风险因素具有较高的接近中心度,能够迅速影响网络中的其他因素。

基于接近中心度的风险因素关联网络如图7所示,由图7可知:鸟击、未严格执行操作程序等风险因素在网络中与其他风险因素的平均距离较短,这

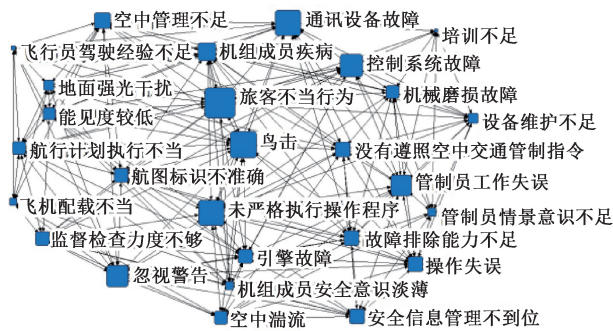


图5 基于中介中心度的风险因素关联网络
Fig.5 Correlation network of risk factors based on intermediate centrality

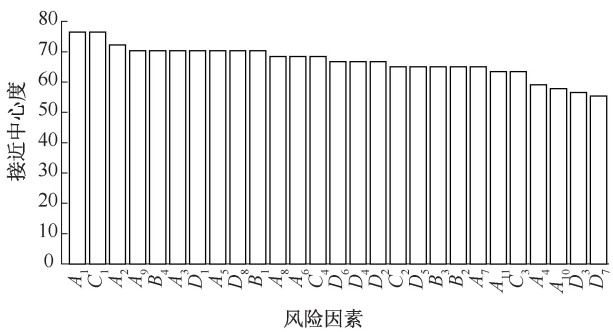


图6 飞行事故风险因素关联网络接近中心度
Fig.6 Flight accident risk factor correlation network closeness centrality

使得它们能够快速传播和接收风险信息,也容易受到其他风险因素的影响。

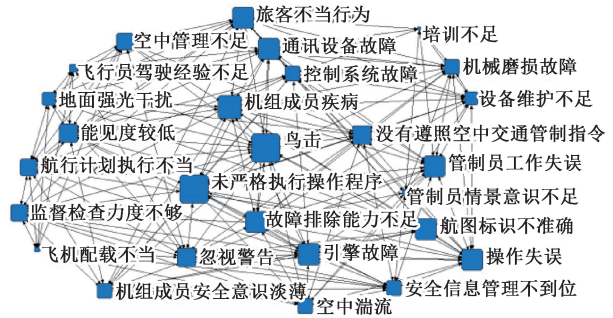


图7 基于接近中心度的风险因素关联网络
Fig.7 Correlation network of risk factors based on closeness centrality

综上所述,未严格执行操作程序这一风险因素,在度数中心度和接近中心度这2个指标上的影响尤为突出,是引发飞行事故的重要隐患。

2.3 飞行事故关键人员风险因素识别

2.3.1 基于事故统计的飞行事故关键人员风险因素识别

在众多飞行事故风险因素中,人员因素极为关键,其中未严格执行操作程序的行为常是引发飞行

事故的直接原因。依据民航局 2024 年新颁布的《航空器的运行文件》^[12],将此类人员分为飞行机组人员、空中交通管制员、机务维修人员和客舱乘务人员。筛选民航局公布的 1 159 份飞行事故报告,共得到 141 份与未严格执行操作程序相关的报告。4 类人员未严格执行操作程序事故统计如图 8 所示,由图 8 可知:在 141 份相关事故报告中,飞行机组人员的占比最高,共 64 份,占比约为 45.5%。这表明飞行机组人员未严格执行操作程序是较为突出的风险因素,对飞行安全影响重大,需重点关注。

2.3.2 基于情感分析的飞行事故关键人员风险因素识别

通过网络获取 4 种类型人员未严格执行操作程序引发飞行事故的抖音视频评论,得到 832 条飞行机组人员、1 155 条空中交通管制员、696 条机务维

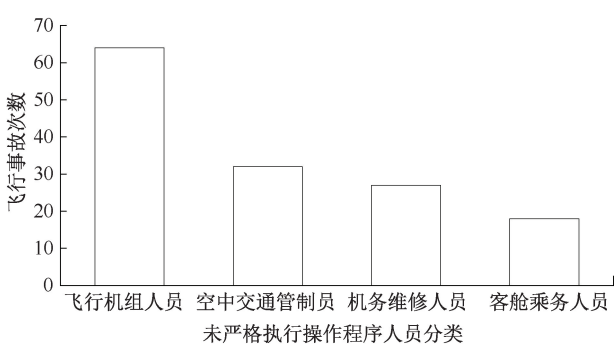


图 8 4 类人员未严格执行操作程序事故统计

Fig.8 Four types of personnel do not strictly implement operation procedures flight accident statistics

修人员和 1 231 条客舱乘务人员的有效评论。利用 BERT 模型计算 4 类人员未严格执行操作程序视频评论的情感值,部分结果见表 3。

表 3 4 类人员未严格执行操作程序视频评论情感值部分结果

Table 3 Partial results of sentiment values of video comments on four types of personnel who did not strictly follow operation procedures

飞行机组人员致因事故评论	情感值	空中交通管制员致因事故评论	情感值	机务维修人员致因事故评论	情感值	客舱乘务人员致因事故评论	情感值
就是他们啥都不干,自动驾驶也能把飞机开回去	0.261	什么都不做,都可能没事	0.459	完全是人为导致的,人命陪不了开这样的玩笑	0.549	空姐也就算了,乘客不是应该系安全带吗	0.582

由于 4 类人员视频评论的数量存在差异,通过计算视频评论情感得分的均值,使结果具有可比性,均值结果如图 9 所示。

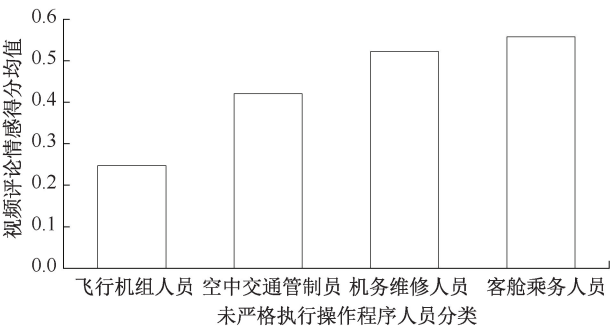


图 9 4 类人员未严格执行操作程序视频评论情感得分均值

Fig.9 Four categories of people do not strictly implement operating procedures video comments emotion score mean

由图 9 可知:飞行机组人员的视频评论情感得分最低,仅为 0.247,反映出公众对其未能严格执行操作程序的负面情绪最为强烈,容易引发更严重的社会舆论。

2.4 低空空域飞行事故关键风险因素识别

从 1 075 份飞行事故报告和 544 份航空安全自愿报告中,筛选出与低空空域相关的 59 份飞行事故报告和 19 份航空安全自愿报告。通过词频统计得出低空空域飞行事故风险因素高频词,如图 10 所示。

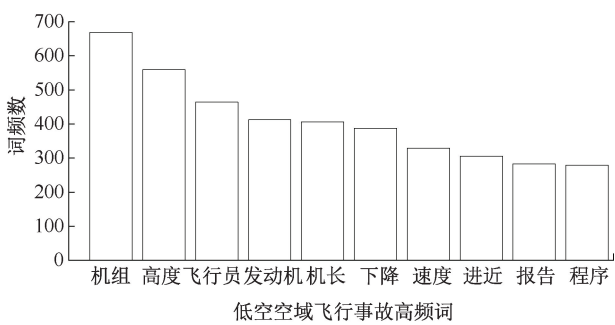


图 10 低空空域飞行事故风险因素高频词

Fig.10 Flight accident risk factors of at low altitude airspace high-frequency words

采用词云图可视化统计低空空域飞行事故风险因素,结果如图 11 所示。

47(8): 95–106.

GAO Fan, XU Sijia. Topic identification and content analysis in the field of digital government governance based on BERTopic model[J]. Information Studies: Theory & Application, 2024, 47 (8): 95–106.

[9] 刘全龙, 张晓霖, 张悦倩, 等. 基于事故致因的标准化煤矿建筑规范库构建[J]. 中国安全科学学报, 2023, 33(5): 128–133.

LIU Quanlong, ZHANG Xiaolin, ZHANG Yueqian, et al. Construction of a database of standardized coal mine building regulation based on accident causation[J]. China Safety Science Journal, 2023, 33(5): 128–133.

[10] SUNG Y Y, KIM S B. Topical keyphrase extraction with hierarchical semantic networks[J]. Decision Support Systems, 2020, 128: DOI: 10.1016/j.dss.2019.113163.

[11] YANG N, JO J, JEON M, et al. Semantic and explainable research related recommendation system based on semi-supervised methodology using BERT and LDA models[J]. Expert Systems with Applications, 2022, 190: DOI: 10.1016/j.eswa.2021.116209.

[12] 中国民用航空局. 航空器的运行文件[Z]. 2024–12–06.



作者简介: 田泽 (1995—), 男, 湖南湘西人, 博士研究生, 主要研究方向为安全风险管
理。E-mail: tianze2025@126.com。



中国科学引文数据库(CSCD)期刊引证指标公布

2025 年 7 月, 中国科学引文数据库(CSCD)期刊引证指标公布, 根据中国科技期刊引证指标 (CSCD-JCR) 数据库数据, 《中国安全科学学报》影响因子为 1.219, 他引影响因子为 0.635, 总被引频次为 2 460, 入选 Q2 区。

中国科技期刊引证指标(CSCD-JCR)数据库是根据 CSCD 年度期刊指标统计数据创建的。该统计数据以 CSCD 数据库为基础, 对刊名等信息进行了大量的规范工作, 所有指标统计遵循文献计量学的相关定律及统计方法, 这些指标如实反映期刊在学术界的影响力。

CSCD 期刊所属学科划分由期刊发表论文的分类特征和期刊的施引文献的分类特征, 以及期刊实际的发文内容综合确定。期刊的分区为期刊所属学科的范围, 依据期刊影响因子的取值对其相对位置进行四等分。

