栗婧,张志珍,杜璇,等. 基于文本分类技术的煤矿违章行为统计方法研究[J]. 矿业科学学报,2022,7(3):344-353. DOI:10.19606/j.cnki.jmst.2022.03.009

Li Jing,Zhang Zhizhen,Du Xuan,et al. Statistical method of coal mine violations based on text classification technology[J]. Journal of Mining Science and Technology,2022,7(3):344-353. DOI:10.19606/j.cnki.jmst.2022.03.009

# 基于文本分类技术的煤矿违章行为统计方法研究

栗婧[1],张志珍[1],杜璇[2],王真[1],刘紫薇[1],辛艳丽[1]

1. 中国矿业大学(北京) 应急管理与安全工程学院,北京 100083;
2. 武警特种警察学院情报侦察系,北京 100100

**摘 要**:煤矿作为高危行业,企业违章行为记录繁杂。为高效、准确、智能地检索和管理企业违章记录信息,减少违章行为发生,本文以某矿近3年的13 935条违章行为数据库为样本,将违章行为分为3大类23小类,基于计算机文本分类技术,通过Jieba分词器文本预处理、向量空间模型构建、TF-IDF模型特征值选取、相似度计算等流程搭建了违章文本数据分类器,在Python环境下构建了可视化展示平台并进行分类统计。结果表明:违章操作在总违章行为中占比最高,达到64%,其次为违章行动和违章指挥。同时对各违章子类进行了高、中、低频类别划分,为预防事故发生提供重要数据支撑。

**关键词**:文本分类技术;违章行为;安全生产;煤矿企业

**中图分类号**:TP 391,TD 79 **文献标志码**:A **文章编号**:2096-2193(2022)03-0344-10

# Statistical method of coal mine violations based on text classification technology

Li Jing[1],Zhang Zhizhen[1],Du Xuan[2],Wang Zhen[1],Liu Ziwei[1],Xin Yanli[1]

1. School of Emergency Management and Safety Engineering,China University of Mining and Technology-Beijing,Beijing 100083,China;
2. Department of Intelligence and Reconnaissance,Special Police College of CAPF,Beijing 100100,China

**Abstract**:As a high-risk industry,coal mining enterprises have a complex record of violations. In order to efficiently,accurately and intelligently retrieve and manage an enterprise's illegal record and reduce the occurrence of illegal behaviors. A database of 13,935 violations in a mine in recent three years is taken as a sample. The illegal actions are divided into 3 categories and 23 subcategories. And based on the computer text classification technology,the illegal text data classifier is built. Its process includes text preprocessing of Jieba word segmentation,vector space model construction,feature value selection of TF-IDF model, and similarity calculation process. Finally, a visual classification statistics and presentation system was constructed in Python environment,and the classified statistics were carried out. The results showed that the proportion of illegal operation is 64%,which is the highest among all illegal behavior,followed by illegal action,and illegal command accounted for the smallest proportion. At the same time,the key subcategories of high frequency,medium frequency and low frequency were analyzed to provide quantitative support for accident prevention.

**Key words**:text classification technology;violations behaviors;production safety;coal mining enterprises

**收稿日期**:2021-08-06 **修回日期**:2021-11-01

**基金项目**:中央高校基本科研业务费专项资金(2021YJSAQ12)

**作者简介**:栗婧(1980— ),女,山西长治人,博士,副教授,主要从事矿山安全、安全管理、安全应急等方面的研究工作。Tel:13811351654,E-mail:rainbow_lijing@163.com

煤矿作为高危行业,安全工作不容忽视。近年来,国家对煤矿安全生产的重视程度提高,事故发生总量呈下降的趋势(图 1),但煤矿一般事故频发,重特大事故时有发生,给我国经济和社会带来了巨大的损失[1],煤矿安全形势依然严峻[2]。因此,指导煤矿安全生产、预防事故发生,依然是安全研究亟待解决的问题。

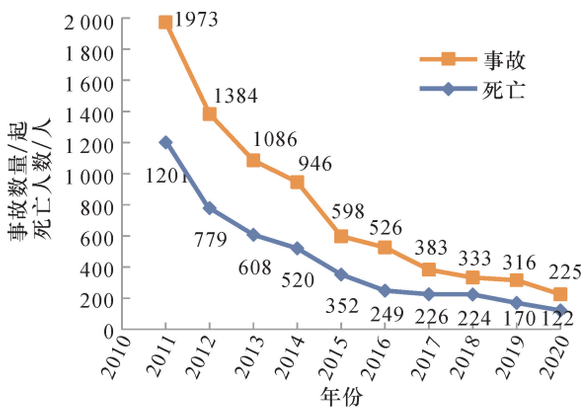


**图 1　煤矿安全生产发展趋势**

**Fig. 1　The development trend of coal mine safety production**

已有研究表明,超过 80% 的煤矿事故是由人的不安全行为造成的,主要表现为违章行为[3]。目前针对煤矿安全事故的研究,大都是基于全国重特大事故案例,而针对特定煤矿违章行为进行统计分析的较少。同时,我国各大煤矿企业累积了大量的违章信息,此类信息具有数据总量大、记录不规范、特征不明显、类别不具体、人工简单统计、数据应用率不高的特点。海因里希法则指出,当一定量的违章和隐患出现时,事故必然发生。因此,如何科学地统计日常违章行为、有效地分析违章数据,并为安全生产进行及时、高效的指导,是研究的重点问题。

近年来,文本分类技术逐渐成为研究热点。在人工智能的背景下,基于自然语言处理(Natural Language Processing,NPL)的文本分类技术对现有的杂乱数据进行结构化处理、分类型展示、动态性分析,能够提升信息处理的效率与准确率,其较强的应用性也得到了验证。崔敏[4]构造了一种融合深层注意力机制的双层双向长短时记忆网络的深度学习模型,加强了文本分析在电力系统故障方面的应用。秦欢等[5]基于改进的隐式马尔科夫算法制定了非结构化数据的结构化表达规则,形成了一份电网企业国网系统运检专业领域的数据词典库。黄亚春[6]基于自然语言处理的相关技术构建了分类效果最优的卷积神经网络(Convolutional Neural Networks,CNN)模型,加速建筑行业的文本自动分类技术的发展。鲁博仁[7]提出了一种面向铁路安全监督文本的字符级特征提取方法(Character Level-Word2Vec,CLW2V)和基于 CLW2V 的铁路安全监督文本分类方法,实现铁路安全监督文本的多类别分类应用。田继存[8]提出了一种基于局部文档频率的文本分类方法(Text Categorization Based on Local Document Frequency,TCBLDF),并将其应用于民航安全自愿报告(Aviation Safety Reporting System,ASRS)数据。可见,现有的研究主要从优化文本分类模型的角度在电力、铁路、建筑、民航等领域展开,但在煤矿安全管理领域研究较少。

本文基于 NPL 文本分类技术搭建适用于煤矿领域违章行为的文本数据分类器,结合事故致因“2-4”模型和某矿违章信息,创建分类规则,建立数据文本分类可视化平台,实现违章数据的导入、文本分类、信息统计及多因素分析等功能。通过对煤矿违章信息分类统计,煤矿企业可实现快速实时统计、挖掘数据深层含义、预测安全事故发生,为开展安全培训、落实安全教育、指导煤矿安全生产提供准确的切入点和有力的数据支撑。

## 1　违章数据分类规则

对复杂问题进行分类时,运用 MECE(Mutually Exclusive Collectively Exhaustive)原则分类可以做到不重不漏,MECE 原则来自于麦肯锡咨询公司,中文含义是“相互独立、完全穷尽”。本文通过 4 个步骤来运用 MECE 原则。

(1) 确定范围。根据某矿违章记录文本,尽可能穷尽收集该矿的违章情况。基于事故致因“2-4”模型理论中对不安全动作的定义及分类方法[9],借鉴其他学者[10-12]利用该理论进行不安全动作分类经验,结合煤矿事故特征,排除掉不违章的不安全动作,对违章的不安全动作(违章行为)进行了划分,将违章行为划分为违章操作、违章行动、违章指挥 3 类。违章操作是指违反煤矿生产相关法律、法规、规章和操作规程,具有操作主体、操作对象、操作过程的行为。违章行动是指没有操作对象的行为,或有操作对象但不以工作为目的的行为。违章指挥是指操作主体为管理员或同级工作者,以命令或安排其他人进行违章操作的行为。

(2) 寻找符合的切入点。以《企业职工伤亡事故分类标准》(GB 6441—1986)规定的 13 种“不

安全行为”为基础划分出10种违章子类(不安全姿势及位置;不按规定对机器维修检查;不正确警戒、预警或使用信号;使用不安全物品代替专业工具工作;不按规定使用安全防护装置使其失效;未对易燃、易爆及其他物品妥善保护;作业前未排查设备环境隐患;违反劳动纪律分散注意力;冒险进入危险场所;未佩戴/错误佩戴安全装备),并将其一一对应到3个违章大类中。

(3) 考虑是否需要细分和补充。通过对该矿违章情况分析发现,违章文本的记录均是以《某矿安全生产管理守则》(以下简称“守则”)为依据,且现有的10种违章子类与实际情况存在差异,需要进一步细化和补充。首先,根据安全管理“三违”(违规作业、违章指挥、违反劳动纪律)中对违章指挥的规定,增加了管理人员的违规组织作业、不合理人员安排以及未有效管控等违章指挥行为。其次,根据守则及具体违章信息记录补充了无证上岗、手指口述不合格、违反标准程序作业、危险气体监测设备使用不当、无人看护作业、安全培训不到位、不安全移动、违反休息规定、违反生产秩序、记录填写不当等违章行为。

(4) 确认是否有遗漏或重复。对上述所有违章子类具体内容、定义适度调整,确保每个子类均包含了1种或几种具有共同特征的具体违章行为,覆盖范围全面,且彼此之间独立。最终将其划分为23种违章子类(表1)。

**表1 违章分类**

**Table 1 Classification of violations**

| 违章大类 | 违章子类 | 具 体 内 容 |
|---|---|---|
| 违章操作 | A1 不安全姿势及位置 | 将自身置于危险的位置,使用错误姿势作业,设备未按照规定停放在指定位置 |
| | A2 不按规定维修检查 | 违反了检查维修相关规定,致使检查维修存在危险或检修不到位 |
| | A3 不正确警戒、预警或使用信号 | 忽视警告标志、警告信号,未能依据规定进行预警信息联络 |
| | A4 使用不安全物品 | 使用其他工具或方式代替作业,未按要求使用工具器具或以手代工具操作 |
| | A5 手指口述不合格 | 井下作业未按规定执行手指口述或手指口述不合格 |
| | A6 违反标准程序作业 | 违反了设备使用规定,错做、少做操作动作及顺序与规定不一致 |
| | A7 不按规定使用安全防护装置 | 不进行安全防护或错误进行安全防护致使安全装置失效的行为 |
| | A8 未对不安全物品妥善保护 | 未对重物、易燃易爆、高温高压等危险物品妥善存放或未加保护措施 |
| | A9 未使用/错误使用危险源检测设备 | 主要有害气体这一潜在的重大安全隐患或其他气体危险隐患,未在作业前使用或错误使用检测设备对危险源进行检测的行为 |
| | A10 未填写或伪造记录 | 每班未准确及时的填写相应记录 |
| | A11 无人看护作业 | 需要看护的作业过程中,无人看护擅自作业 |
| | A12 作业前未排查隐患 | 开始作业前未对设备环境进行安全隐患的排查,或明知设备环境存在隐患,仍强行作业的行为 |
| 违章行动 | B1 不安全移动 | 作业期间走动、跑动等违反井下安全作业条例行为 |
| | B2 破坏生产管理秩序 | 员工不服从安全管理、不配合各级检查人员检查等行为 |
| | B3 违反劳动纪律 | 员工之间发生矛盾产生争吵或进一步升级起哄谩骂、打架斗殴或不认真工作开玩笑、乱摆弄、脱岗、酒后作业 |
| | B4 违反休息规定 | 在非休息时间或非休息场所瞌睡,或在规定休息时间为方便休息而违反休息规定 |
| | B5 违规进入危险场所 | 未经组织部门许可,违规进入危险警戒区域 |
| | B6 未佩戴/错误佩戴安全装备 | 未按规定穿戴佩戴帽带、自救器等人身安全防护装备,包括特定作业时未按规定穿戴佩戴相应安全防护装备 |
| | B7 无证上岗/证件不符合规定 | 无证上岗、证件过期、证件损坏等 |
| 违章指挥 | C1 违规组织作业 | 生产作业没有获得批准的情况下擅自组织作业;或现场作业环境发生变化,未及时汇报,强行组织作业 |
| | C2 不合理人员安排 | 安排与作业要求不相吻合的资质人员执行作业 |
| | C3 未有效对井下作业秩序进行管控 | 指挥人员未随时掌握井下作业动态,对井下违章行为实施把控,维持现场作业秩序,纠正作业过程中存在的不安全问题 |
| | C4 安全培训不到位 | 未按规定组织培训或组织培训时间不够,以致在岗作业人员对其岗位作业标准、公司制度规范不清楚,或掌握不到位 |

## 2　违章文本分类器

文本分类是基于特定文本信息及规则对信息与某一或某些主题对应关系的划分[13]。它是 NLP 领域中一项具有挑战性的任务,在情感分类、自动问答、舆情分析等领域具有广泛的应用。违章文本分类器处理非结构化数据文本,其文本分类流程主要包含文本预处理、构建空间向量、特征值选取、相似度计算等步骤,如图 2 所示。

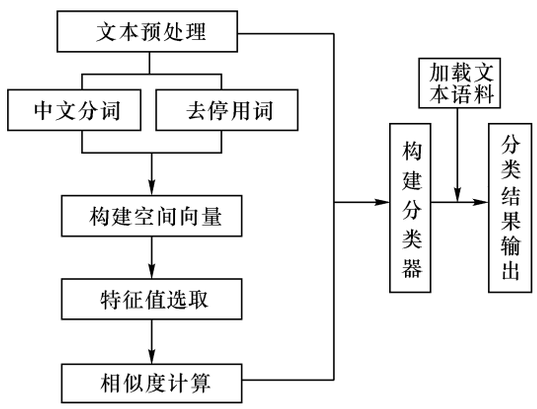


**图 2　文本分类流程**

**Fig. 2　Text classification process**

### 2.1　文本预处理

文本预处理的目的是剔除文本中所有与分类任务无关的内容,并根据需要实现同义词替换功能,将文本转化为由其包含的基本语义单位组成的表列。其首要工作是分词,分词质量的好坏直接影响分类结果。根据煤矿企业的特点,初步选择了两种分词工具:ICTCLAS 分析系统和基于 Python 开发的 Jieba 分词器。进一步比较发现,Jieba 分词虽然效果上不如 ICTCLAS,但在 Python 编写上模型简单、代码清晰、扩展性好,可自定义词库的设计和添加功能,且能够根据需要进行优化和改进,有改进的想法就能够编写代码进行修改,从而提高专业领域文本处理结果的准确性与有效性[14-15],因此分词工具最终选用 Jieba 分词器。使用 Jieba 分词工具自带的自定义词库功能对分词标准进行优化,建立“违章词典.txt”及去停用词文件“stopword.txt”。

### 2.2　构建向量空间模型

为了便于计算机识别,需将文本预处理结果转换为向量空间模型(Vector Space Model,VSM)[16],通过空间向量对文本信息进行描述,以便于后期对违章记录与所定义违章子类之间进行相似度计算及特征向量计算。具体步骤如下:

(1) 与违章记录有关的统计结果记为文档集合 C。该文档集合是 TF-IDF(Term Frequency-Inverse Document Frequency)算法实现的基础语料库。

(2) 以统计记录与违章形式文本为对象完成中文分词处理,对其中的语气词、连接词、标点符号等元素进行去除和简化,获得词项结果,并用词项集的形式对上述文本的特征内涵进行表述。

### 2.3　特征值选取

特征值选择是指从特征全集中选取一部分对于分类有贡献的特征子集。本研究采用 TF-IDF 算法对文本特征值进行筛选。TF-IDF 能够对词语在文档数据集内的重要性进行分析和评价,具体构成要素为 TF 与 IDF 两部分,用 $t_i$ 表示词项集合中第 $i$ 个词项,则其对应的词频 TF 用 $tf_{i,j}$ 表示,第 $i$ 个词项 $t_i$ 的 *IDF* 值用 $idf_i$ 表示,TF-IDF 用 $tfidf_{i,j}$ 表示,其计算公式如下:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k m_{k,j}} \tag{1}$$

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{2}$$

$$tfidf_{i,j}=tf_{i,j}\times idf_{i,j} \tag{3}$$

式中,$tf_{i,j}$ 为词项 $t_i$ 在文档 $j$ 中的词频;$n_{i,j}$ 为词项 $t_i$ 在文档 $j$ 中出现的次数;$m_{k,j}$ 为文档 $j$ 中第 $k$ 个词项;$\sum_k m_{k,j}$ 为文档 $j$ 中所有词项出现的次数总和;$idf_i$ 为词项 $t_i$ 逆文本频率;$|D|$ 为语料库中的文档总数;$|\{j:t_i\in d_j\}|$ 为包含词语 $t_i$ 的文档数量;$tfidf_{i,j}$ 为所求的词项 $t_i$ 在文档 $j$ 中的权重。

借助 TF-IDF 模型计算词项对应的权重值,以此为依据可完成词项排序处理。根据排序结果与选取标准确定关键词,所有的关键词用关键词词典 $D$ 进行表示,为违章记录关键词和违章类型的并集;而关键词的数量将作为对应空间向量模型的维度数,这个特定维度的向量模型就将作为违章统计记录文本、违章形式文本的抽象空间表达。

进一步分析发现,当违章记录文本和违章类型文本中都包含“指挥”这一关键词时,根据 TF-IDF 计算公式,“指挥”的权重就会变得很小。尤其是当违章文本与违章类型文本长度不一样时,在截取若干关键词的过程中,很可能把权重小的“指挥”

一词剔除掉。而“指挥”一词又是违章类型判定的关键动作,没有“指挥”一词,违章的描述就变得不完整了,缺少了“指挥”这个动作描述,最终会降低文本相似度的匹配效果。因此,本文对上述过程中的词典 $D$ 进行了重新设计,新的词典 $D$ 由违章形式文本中的关键词组成。利用中文分词对违章形式文本进行分词操作,分词得到的 $M'$ 个词项全部作为违章形式文本关键词进行保留。在向量空间中,$M'$ 个关键词组成 $M'$ 维向量,并以此建立新的词典 $D'$,并构建空间向量(表2)。

**表2 词典 $D'$ 构建空间向量模型**

**Table 2 Dictionary $D'$ to build a spatial vector model**

| 步骤 | 分词 | TF-IDF 计算权重 | 取若干关键词 | 关键词并集组成词典 $D'$,以词典 $D'$ 构建空间向量 |
|---|---|---|---|---|
| 违章记录文本 | 中文分词 | 计算权重 | $[(t_1,w_1),(t_2,w_2),\cdots,(t_i,w_i)]$ | $[(t_1,w_1),(t_2,w_2),\cdots,(t_j,w_j)]$ |
| 违章形式文本 | 中文分词 | 计算权重 | $[(t_1,w_1),(t_2,w_2),\cdots,(t_j,w_j)]$ | |

### 2.4 相似度计算

通过相似度分析技术判断相似度水平,从而确定违章行为的具体分布情况与分布特征。余弦相似度是通过向量之间的夹角来衡量向量相似性。基于 TF-IDF 模型可确定文本特征值对应的空间向量,以该向量为基础对其与预设分类向量之间的夹角进行分析。任意两个文档 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 之间的相似性系数 $\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)$ 指的是两个文档内容的相关程度。如图3可知,在向量夹角很小时,可认为其表现出较为显著的相似性特征;当夹角为0时,两个向量将完全相同。设文档 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 表示 VSM 中的两个向量:

$$\boldsymbol{D}_1=D_1(\omega_{11},\omega_{12},\cdots,\omega_{1n})$$

$$\boldsymbol{D}_2=D_2(\omega_{21},\omega_{22},\cdots,\omega_{2n})$$

根据夹角的取值结果,确定违章文本和预设分类之间的相似度水平,完成对违章类别的划分。其计算公式为

$$\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)=\cos\theta=\frac{\sum_{k=1}^{n}\omega_{1k}\times\omega_{2k}}{\sqrt{\sum_{k=1}^{n}\omega_{1k}^2\sum_{k=1}^{n}\omega_{2k}^2}} \tag{4}$$

式中,$\cos\theta$ 代表 $x_i$,$y_i$ 文本的相似度。

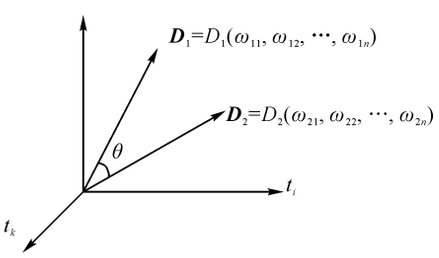


**图3 文本空间向量模型**

**Fig. 3 Text space vector model**

以煤矿登高作业为例,设违章形式1和2所对应的关键词词典分别为

$M_1$=(工作面,登高,作业,无人,扶梯)

$M_2$=(工作面,登高,作业,未系,安全带)

如果以 $X$=“登高作业无人扶梯”为具体的违章记录结果,则该违章文本所对应的两个违章词典空间向量可分别表示为

$$\boldsymbol{M}_1=[(\text{工作面},0)(\text{登高},1)(\text{作业},1)(\text{无人},1),(\text{扶梯},1)] \tag{5}$$

$$\boldsymbol{M}_2=[(\text{工作面},0)(\text{登高},1)(\text{作业},1)(\text{未系},0)(\text{安全带},0)] \tag{6}$$

其相似度结果为 $\cos\ \theta_1=\frac{2}{\sqrt{5}}$;$\cos\ \theta_2=\frac{\sqrt{2}}{\sqrt{5}}$,易知,$\cos\ \theta_1>\cos\ \theta_2$,所以 $\theta_1<\theta_2$,即 $X$ 与 $\boldsymbol{M}_1$ 表现了较为显著的相似特征。根据上述两种相似度结果的对比关系,以相似度较高的结果为依据进行划分,从而确定具体违章记录对应的违章形式。

### 2.5 分类结果检验

随机抽取278条(2%)违章信息同时进行机器分类与人工分类,利用 SPSS 软件相关性分析机器分类与人工分类相关关系的密切程度。分类数据见表3。计算 Pearson 相关系数,选择【分析】—【相关】—【双变量】,得到双变量对话框,将“机器分类”与“人工统计”选入“变量(V)”框中得到相关性表格。易知,两变量的线性关联显著性(双侧)值为0.000,小于0.10,且在显著性水平0.01下的皮尔森相关性系数为0.986,说明机器分类的结果可信度较强。

表 3　机器分类与人工分类结果

Table 3　Machine classification and artificial classification results

| 违章子类 | 机器分类/起 | 人工分类/起 |
|---|---|---|
| A1 | 11 | 10 |
| A2 | 3 | 4 |
| A3 | 10 | 9 |
| A4 | 8 | 7 |
| A5 | 19 | 23 |
| A6 | 53 | 50 |
| A7 | 26 | 27 |
| A8 | 0 | 1 |
| A9 | 14 | 16 |
| A10 | 12 | 11 |
| A11 | 16 | 14 |
| A12 | 9 | 10 |
| B1 | 8 | 5 |
| B2 | 3 | 1 |
| B3 | 13 | 16 |
| B4 | 32 | 35 |
| B5 | 1 | 2 |
| B6 | 27 | 23 |
| B7 | 1 | 3 |
| C1 | 1 | 3 |
| C2 | 4 | 3 |
| C3 | 2 | 1 |
| C4 | 5 | 4 |

## 3　违章文本分类可视化平台搭建

以第 2 章搭建的文本分类器为技术基础,后端使用 Python 开发工具完成对违章文本的分类处理,并为前端提供数据调用接口。前端使用 Vue 框架,在 Window 上搭建后端文本处理结果及统计分析的可视化展示平台。文本原始数据是由某矿企业安全检查部门给出的每日总计违章统计表格,近 3 年共 1 095 张,以某年 4 月 5 日为例(图 4),利用 pandas 对原始数据进行抽取[17],并将抽取后的结果统一归类至设定好的新表格中。该平台实现了数据导入、文本分类、信息统计、多因素分类统计等多项功能。本文主要针对文本分类实验结果进行展示及分析,将某矿近 3 年的 13 935 条违章记录加载至分类平台,完成自动分类后部分结果如图 5 所示。

### 3.1　违章分类结果展示及统计分析

对 3 种违章大类、23 种违章小类发生的频次及各违章大类的违章次数占总违章次数的比率分析可知,违章操作类违章共有 8 898 起,在 3 类违章中频次最高,占总违章次数的 64%;违章行动类共 4 453 起,在总违章次数中占比 32%;违章指挥类共 584 起。按照各违章大类中违章子类发生的

| 级别 | 班次 | 单　位 | 地　点 | 姓　名 | 工　种 | 年龄 | 工龄 | 三 违 情 况 | 性质 | 检查人单　位 | 检查人姓　名 | 处罚情况 | 诚信币扣除 | 诚信币结余 | 备　注 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 0点班 | 综掘5.3队 | 22采区轨顺 | [redacted] | 掘进工 | 40 | 3 | 不放心人员未戴黄帽 | 严重 | 安监大队 | [redacted] | 处罚500元 | 3分 | 3分 | 二次违章 |
| 17 | 8点班 | 综掘5.1队2组 | 2207皮顺 | [redacted] | 掘进工 | 39 | 3 | 登高作业联网无人扶梯 | 一般 | 安监大队 | [redacted] | 处罚100元 | 3分 | 9分 | |
| 18 | 8点班 | 公司机掘队 | 2207轨顺 | [redacted] | 队干 | 38 | 5 | 随身携带便携式瓦检仪未开启 | 一般 | 安监大队 | [redacted] | 加倍处罚200元 | 3分 | 3分 | 三次违章 |
| 19 | 4点班 | 炮掘一组 | 470南翼辅运巷 | [redacted] | 队干 | 40 | 2 | 作业地点两端无阻车器 | 一般 | [redacted]安监站 | [redacted] | 加倍处罚500元 | 3分 | 6分 | 二次违章 |
| 20 | 4点班 | 综掘5.2队 | 2207轨反 | [redacted] | 掘进工 | 40 | 3 | 打顶钻作业时钻机无油 | 一般 | 安监大队 | [redacted] | 加倍处罚500元 | 3分 | 6分 | 二次违章 |

4月5日上报违章行为20人次，其中1人次不算违章，一般违章行为发18人次，严重违章行为发生1人次；矿领导、科室人员查纠8人次，队组自查6人次，安全员查纠5人次；

**典型违章行为分析：**综掘3.1队[redacted]工作面打帮锚杆时未用助力器，经落实情况属实，矿领导及科室人员进入工作面看发现打帮锚杆时，责任人[redacted]使用锚杆钻机助力，一旦别脱容易上机周围作业人员，规程明确要求：大帮锚杆助推时必须使用专用助力器，经核实现场无助力器，属于源头安全管理不到位违章行为，根据《员工行为考核细则》，给予责任人[redacted]处罚200元，队组源头安全管理不到位，给予队长[redacted]处罚500元。下来各队组要进一步强化源头管理，工作面必须按规定存放必备工器具，必须保持完好，有效杜绝被迫违章行为现象发生。

审核人：　　制表人：[redacted]

图 4　某年 4 月 5 日违章记录表

Fig. 4　Record of violations on April 5


图 5　煤矿违章行为文本分类可视化平台

Fig. 5　Visual platform for text classification of coal mine violations

次数由高到低进行排序(图 6),违章操作中排名较为靠前的分别是 A6“违反标准程序作业”、A7“不按规定使用安全防护装置”、A5“手指口述不合格”3 类;违章行动中发生次数较多的是 B4“违反休息规定”、B6“未佩戴/错误佩戴安全装备”;违章指挥中主要的违章子类是 C1“安全培训不到位”、C2“不合理的人员安排”。

其主要原因分析如下:

违章操作主要是认知和应对环节出现问题[18]。煤矿井下生产工艺复杂,设备种类繁多,涉及大量繁杂的标准化生产操作过程。员工安全操作规程掌握不足、对危险的预判能力不足、不进行或错误进行安全防护以及企业未落实煤矿标准化生产、未健全标准化体系、安全培训缺失等,均会造成违章操作的产生。

违章行动是一种个人行为的失当[19]。违章行动主要受个人态度、主观规范、直觉行为控制[20],也受周围员工潜移默化的影响,具有强烈的主观能动性。员工消极、低迷的工作态度以及企业排班作业制度存在问题,均会增加违章行动的发生。

违章指挥是指管理决策环节出现的差错,主要受管理层行为人个体素质所影响。行为的主体责任大,行为的影响意义重要,危害性完全不亚于另外两种违章类别,但是由于管理层的人员基数远远小于普通员工,因此违章数量是最小的。企业现场指挥人员个人能力不足、同级作业人员盲目指挥、企业培训不合理等均会导致违章指挥的出现。

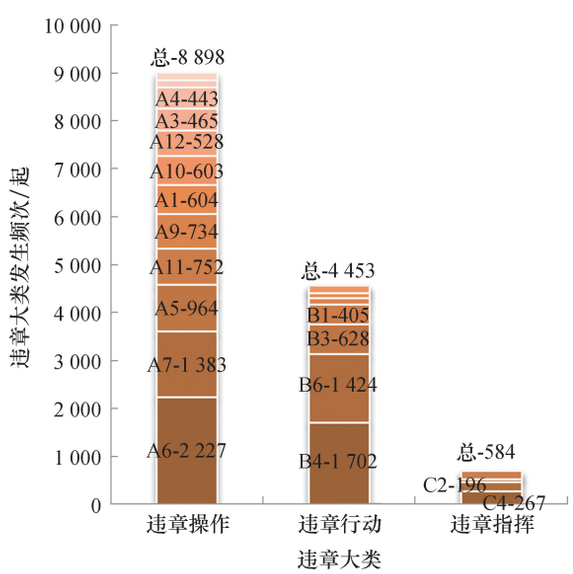


**图 6 违章类别频次**

**Fig. 6 Frequency chart of violation categories**

### 3.2 违章子类频次分类分析

为了对违章行为进行精准预防,需要对违章子类发生频次及其在所属的违章大类中的占比(以下简称“占比”)进行分析。基于 SPSS 软件的 K-均值聚类(K-means)分析违章子类发生频次的聚类[21],将各违章子类划分为高频、中频、低频违章子类。

以违章操作的 12 种子类为例,将违章操作子类代码及对应的频次导入 SPSS 数据编辑器,选择【分析】-【分类】-【K-平均值聚类分析】,“频次”设为变量,“违章操作子类”设为标注个案,聚类数选择 3,方法选择“迭代与分类”,“保存”中勾选“聚类成员”与“与聚类中心的距离”,“选项”中勾选“每个个案的聚类信息”,点“确定”,即可得到聚类之后的 3 类数据信息(表 4)。

**表 4 违章操作子类 K-平均值聚类分析结果**

**Table 4 K-mean clustering analysis results of illegal operation**

| 个案编号 | 违章操作子类 | 丛集 | 距离 |
|---|---|---|---|
| 1 | A6 | 1 | 0.000 |
| 2 | A7 | 2 | 424.750 |
| 3 | A5 | 2 | 5.750 |
| 4 | A11 | 2 | 206.250 |
| 5 | A9 | 2 | 224.250 |
| 6 | A1 | 3 | 198.571 |
| 7 | A10 | 3 | 197.571 |
| 8 | A12 | 3 | 122.571 |
| 9 | A3 | 3 | 59.571 |
| 10 | A4 | 3 | 37.571 |
| 11 | A2 | 3 | 256.429 |
| 12 | A8 | 3 | 359.429 |

同理,将违章行动和违章指挥进行聚类之后,将丛集 1 认定为高频违章子类、丛集 2 认定为中频违章子类、丛集 3 认定为低频违章子类,如图 7 与表 5 所示。

分析统计结果可得,“违反标准程序作业”2 227 起,占比 25.03%,为高频违章操作;“不按规定使用安全防护装置”1 383 起,“手指口述不合格”964 起,“无人看护作业”752 起,“未使用/错误使用危险源检测设备”734 起,总占比 43.07%,为中频违章操作;其余违章操作发生的频数(起)均小于 605,总占比 31.9%,记为低频违章操作。“违反休息规定”1 702 起,占比 38.22%,记为高频违章行动;“未佩戴/错误佩戴安全装备”1 424 起,占比 31.98%,记为中频违章行动;“违反劳动纪律、不安全移动、无证上岗/证件不符合规定、违规进入

危险场所、破坏生产管理秩序”分别为 628 起、405 起、143 起、105 起、46 起,总占比为 29.79% ,记为低频违章行动。“安全培训不到位”267 起,占比 45.72% 记为高频违章指挥;“不合理的人员安排”196 起,占比为 33.56% ,记为中频违章指挥;“违规组织作业、未有效对井下作业秩序进行管控”为 60 起左右,总占比为 20.72% ,记为低频违章指挥。总的来说,违章操作呈现总体频次基数大、违章子类多的特点;违章行动的高、中频违章子类种类少、占比高,低频违章子类种类较多;违章指挥由于基数少,违章子类频次均较低,但高、中频违章子类占比极高。

企业违章行为管控措施要根据不同违章类别的不同特点来制定。对违章子类占比数据进行深入分析可知,高频违章类别需要企业高度重视,立即采取措施予以整改,如通过优化安全生产管理信息系统[22]、完善作业程序标准化规范员工操作行为,通过制定实施针对性的整改措施减少高频违章行为的发生,通过定期检查整改执行进度严控整改措施的效果,通过严格开展安全检查人员的绩效考核规范违章行为监督工作等。中、低频违章类别是发展成为高频违章类别的潜在因素,不容忽视,企业要定期组织技能培训,对不安全行为组织干预[23],帮助员工自觉养成按章操作的习惯。此外,要增加安全融入企业管理的程度,强化企业安全文化建设,从而达到从根本控制企业生产风险的效果。

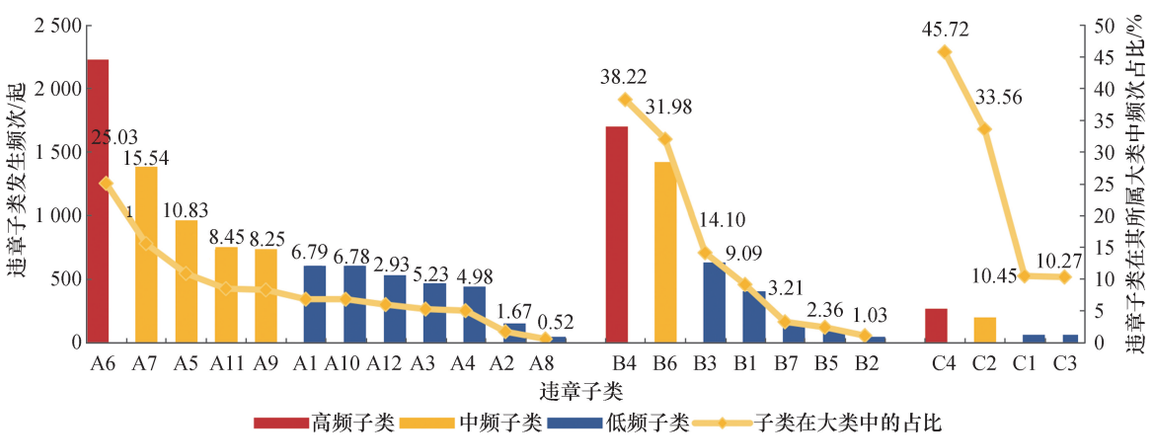


**图 7　违章子类占比**

**Fig. 7　Chart of violation of the subclass**

**表 5　违章子类频次分类分析**

**Table 5　Frequency classification and analysis of illegal subcategories**

| 违章类别 | 频次分类 | 总占比/% | 违章子类 |
|---|---|---|---|
| 违章操作 | 高频 | 25.03 | A6 违反标准程序作业 |
| | 中频 | 43.07 | A7 不按规定使用安全防护装置、A5 手指口述不合格、A11 无人看护作业、A9 未使用/错误使用危险源检测设备 |
| | 低频 | 31.9 | A1 不安全姿势及位置、A10 未填写或伪造记录、A12 作业前未排查隐患、A3 不正确警戒、预警或使用信号、A4 使用不安全物品、A2 不按规定维修检查、A8 未对不安全物品妥善保护 |
| 违章行动 | 高频 | 38.22 | B4 违反休息规定 |
| | 中频 | 31.98 | B6 未佩戴/错误佩戴安全装备 |
| | 低频 | 29.79 | B3 违反劳动纪律、B1 不安全移动、B7 无证上岗/证件不符合规定、B5 违规进入危险场所、B2 破坏生产管理秩序 |
| 违章指挥 | 高频 | 45.72 | C4 安全培训不到位 |
| | 中频 | 33.56 | C2 不合理的人员安排 |
| | 低频 | 20.72 | C1 违规组织作业、C3 未有效对井下作业秩序进行管控 |

## 4　结　论

本文采用自然语言文本分类技术搭建了适用于煤矿的违章文本分类器与煤矿违章行为分类可视化平台,选取某矿违章数据进行了智能分类和统计,得到如下结论:

(1) 本文基于“2-4”模型将煤矿违章划分为3大类23小类。利用文本分类技术,结合违章数据特点构建完成了煤矿违章行为文本数据自动化分类器,其流程为:Jieba分词器文本预处理→向量空间模型构建→TF-IDF筛选文本特征值→相似度水平计算。

(2) 在文本分类器的基础上,利用Vue框架开发了煤矿违章分类可视化软件平台,利用Python环境实现了对违章数据的自动化分类及数据调用,使平台具有违章数据导入、违章信息文本分类、违章多因素分类统计等多项功能,极大地提升数据分析的速度和准确度。

(3) 违章数据分析结果显示,3类违章中违章操作类频次最高,占违章总数的64%,其次为违章行动与违章指挥,分别占32%和4%。进一步对违章子类分析可知,其中违章操作含1种高频、4种中频、7种低频子类,高频操作占比25.03%,为“违反标准程序作业”;违章行动含1种高频、1种中频、5种低频子类,高频违章行动占比38.22%,为“违反休息规定”;违章指挥含1种高频、1种中频子类、2种低频子类,高频违章指挥占比45.72%,为“安全培训不到位”。

本文立足于某煤矿近3年违章数据建立煤矿违章行为分类平台,其违章分类方法及依据违章词典的向量空间模型较适用于该煤矿,但在其他煤矿的适用性缺乏验证,且该分类器准确率有待进一步提升,智能化水平也有待提高。同时,应开展违章行为统计方法及可视化平台在其他煤矿的适用性验证,分析全国煤矿的违章数据,以期建立一个普遍适用于煤矿专业领域的自然语言文本分类模型。

## 参考文献


[1] 佟瑞鹏,赵辉,张娜,等. 矿工不安全行为涌现性建模研究[J]. 矿业科学学报,2020,5(3):311-319.
Tong Ruipeng,Zhao Hui,Zhang Na,et al. Research on emergency modeling of unsafe behavior of coal miners[J]. Journal of Mining Science and Technology,2020,5(3):311-319.

[2] 丁百川. 2020年全国煤矿事故特点及原因分析[N]. 中国能源报,2021-02-22(15).

[3] 时砚. 群体动力学在安全管理中违章行为矫正的应用[D]. 北京:北京交通大学,2008.

[4] 崔敏. 基于文本识别技术的电气设备监测数据处理[D]. 北京:华北电力大学,2019.

[5] 秦欢,门业堃,于钊,等. 基于隐马尔科夫和主成分分析的电网数据词典构建[J]. 电力大数据,2019,22(1):16-21.
Qin Huan,Men Yekun,Yu Zhao,et al. The construction of grid data dictionary based on HMM and PCA[J]. Power Systems and Big Data,2019,22(1):16-21.

[6] 黄亚春. 基于自然语言处理的建筑工程安全事故报告风险研究[D]. 武汉:华中科技大学,2019.

[7] 鲁博仁. 面向铁路安全监督文本的分类技术研究[D]. 郑州:郑州大学,2020.

[8] 田继存. 文本分类及其在民航安全自愿报告分析中的应用研究[D]. 天津:中国民航大学,2010.

[9] 傅贵. “2-4”模型视角下的行为安全[J]. 现代职业安全,2019(12):17-19.
Fu Gui. Behavior safety from the perspective of “2-4” model[J]. Modern Occupational Safety,2019(12):17-19.

[10] 傅贵. 安全科学学及其应用探讨[J]. 安全,2019,40(2):1-10.
Fu Gui. The science of safety science and its practical implications[J]. Safety & Security,2019,40(2):1-10.

[11] 贺莹鸽,连民杰,江松,等. 矿工习惯性违章行为风险态势评估[J]. 中国安全科学学报,2020,30(12):62-69.
He Yingge,Lian Minjie,Jiang Song,et al. Risk state assessment of coal miners' habitual violation behavior[J]. China Safety Science Journal,2020,30(12):62-69.

[12] 傅贵,王秀明,李亚. 事故致因“2-4”模型及其事故原因因素编码研究[J]. 安全与环境学报,2017,17(3):1003-1008.
Fu Gui,Wang Xiuming,Li Ya. On the 2-4 model and the application of its causative codes to the analysis of the related accidents[J]. Journal of Safety and Environment,2017,17(3):1003-1008.

[13] 于游,付钰,吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报,2019,5(5):1-8.
Yu You,Fu Yu,Wu Xiaoping. Summary of text classification methods[J]. Chinese Journal of Network and Information Security,2019,5(5):1-8.

[14] 祝永志,荆静. 基于Python语言的中文分词技术的研究[J]. 通信技术,2019,52(7):1612-1619.
Zhu Yongzhi, Jing Jing. Chinese word segmentation technology based on python language[J]. Communications Technology,2019,52(7):1612-1619.

[15] 尤众喜,华薇娜,潘雪莲. 中文分词器对图书评论和情感词典匹配程度的影响[J]. 数据分析与知识发现,2019,3(7):23-33.
You Zhongxi,Hua Weina,Pan Xuelian. Matching book reviews and essential sentiment lexicons with Chinese word segmenters[J]. Data Analysis and Knowledge Discovery,2019,3(7):23-33.

[16] 马艳荣,温煜坤. 基于向量空间模型的对外汉语应用文写作词汇分类系统研究[J]. 现代电子技术,2021,44(8):137-140.
Ma Yanrong, Wen Yukun. Study on VSM-based vocabulary classification system of TCFL practical writing[J]. Modern Electronics Technique, 2021, 44(8):137-140.

[17] 丁宇,李瑞祥. 利用 pandas 的数据清洗功能来提取宽带用户的相关信息[J]. 网络安全和信息化,2021(9):94-96.
Ding Yu, Li Ruixiang. Pandas uses its data cleaning function to extract information about broadband users. [J]. Cybersecurity & Informatization, 2021(9):94-96.

[18] 殷文韬,傅贵,公建祥. 煤矿工人违章操作的"认知-行为"失效机理与管理措施[J]. 中国安全科学学报,2015,25(10):153-159.
Yin Wentao, Fu Gui, Gong Jianxiang. Research on coal miners' operating against safety regulation:"cognition-behavior" failure mechanism and control measures[J]. China Safety Science Journal, 2015, 25(10):153-159.

[19] 曹家琳,傅贵. 煤与瓦斯突出事故不安全动作分类研究[J]. 煤矿安全,2016,47(9):240-242,246.
Cao Jialin, Fu Gui. Classified study on unsafe Acts in coal and gas outburst accidents[J]. Safety in Coal Mines, 2016, 47(9):240-242, 246.

[20] 李乃文,马跃,牛莉霞. 基于计划行为理论的矿工故意违章行为意向研究[J]. 中国安全科学学报,2011,21(10):3-9.
Li Naiwen, Ma Yue, Niu Lixia. Research on miners' deliberate violation behavior intentions based on theory of planned behavior[J]. China Safety Science Journal, 2011, 21(10):3-9.

[21] 宋仁旺,苏小杰,石慧. 基于空间分布优选初始聚类中心的改进 K-均值聚类算法[J]. 科学技术与工程,2021,21(19):8094-8100.
Song Renwang, Su Xiaojie, Shi Hui. An improved K-mean clustering algorithm based on spatial distribution to optimize the initial clustering center[J]. Science Technology and Engineering, 2021, 21(19):8094-8100.

[22] 宋曦,丁文梅,宁云才,等. 煤矿安全生产管理体系优化研究——以陕西某煤矿为例[J]. 矿业科学学报,2019,4(2):187-94.
Song Xi, Ding Wenmei, Ning Yincai, et al. Research on the optimization of coal mine safety Production management system——take a coal mine in Shaanxi province as an example[J]. Journal of Mining Science and Technology, 2019, 4(2), 187-194.

[23] 佟瑞鹏,陈策,刘大鹏. 矿工不安全行为组织干预时效性研究[J]. 矿业科学学报,2016,1(2):155-61.
Tong Ruipeng, Chen Ce, Liu Dapeng. Timeliness analysis on organization Intervention of miners' unsafe behavior[J]. Journal of Mining Science and Technology, 2016, 1(2):155-161.

(责任编辑:王晓玲)