文章编号：1007-2780(2023)11-1531-11

基于有监督对比学习的航天信息获取与图像生成

齐翌辰[1]，赵伟超[2*]
(1. 东北大学 计算机科学与工程学院，辽宁 沈阳 110167；
2. 中国科学院 长春光学精密机械与物理研究所 网络与信息化技术中心，吉林 长春 130033)

摘要：为了提高获取开源航天信息的效率并解决开源航天信息内容较长、数量较为有限、应用常用文本分类模型鲁棒性较差以及文本信息不够直观等问题，本文提出一种基于有监督对比学习的航天信息分类方法。该方法基于带有注意力机制(Attention)的双向长短期记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)，融合对比学习技术，对开源的信息进行处理并分析，进而高效地筛选出航天类的信息，利用unCLIP(un-Contrastive Language-Image Pre-Training)模型生成信息对应的图像。实验结果表明，对比CNN(Convolutional Neural Networks)、BiLSTM、Transformer和BiLSTM-Attention等常用的文本分类方法，该方法在准确率、召回率和F1-Score上均表现良好，其中F1-Score达到0.97，同时以图像的形式呈现信息，使信息更加清晰直观。本文方法可以充分使用网络公开的数据资源，有效地提取开源航天信息并生成对应图像，对航天信息的分析和研究具有重要价值。

关　键　词：有监督文本分类；对比学习；文本生成图像；航天信息
中图分类号：TP391.1；TP751.1　　**文献标识码：**A　　**doi：**10.37188/CJLCD.2023-0056

Aerospace information acquisition and image generation based on supervised contrastive learning

QI Yi-chen[1], ZHAO Wei-chao[2*]

(1. *School of Computer Science & Engineering, Northeastern University, Shenyang* 110167, *China*;
2. *Network and Information Technology Center, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun* 130033, *China*)

Abstract: In order to improve the efficiency of obtaining open source aerospace information, and solve the problems of long open source aerospace information content, relatively limited quantity, poor robustness of commonly used text classification models, and unintuitive text information, this paper proposes a method for aerospace information text classification based on supervised contrastive learning. The method is based on the bidirectional long short-term memory (BiLSTM) network with the attention mechanism, integrates comparative learning technology, processes and analyzes open source information, efficiently screenes out aerospace information, and uses the unCLIP (un-Contrastive Language-Image Pre-Training) model to generate an image corresponding to the information. The experimental results show that compared with commonly used text classification methods such as CNN (Convolutional Neural Networks), BiLSTM, Transformer and BiLSTM-Attention, this method performes well in accuracy, recall and F1-Score, among

收稿日期：2023-02-15；**修订日期：**2023-03-01.

*通信联系人，E-mail：zhaoweichao@ciomp.ac.cn

them, F1-Score reaches 0.97. At the same time, information is presented in the form of images to make information clearer and more intuitive. It can make full use of open data resources on the network, effectively extract open-source space information and generate corresponding images, which is of great value to the analysis and research of aerospace information.

Key words: supervised text classification; contrastive learning; text-to-image synthesis; aerospace information

1 引　　言

近年来,随着科学技术和工业制造水平的提高,我国航天事业实现了快速且长足的发展,取得举世瞩目的成就,而这些成绩的获得离不开航天信息工作的强有力支持。由于航天信息存在数据量较稀疏等特点,使得信息采集过程中存在数据不易获取,人工搜集效率较低等问题[1]。在互联网时代的浪潮之下,开源的航天信息资源呈现快速增长的趋势,涉及的内容较为广泛,来源更加丰富[2]。因此,如何自动化且高效地从海量数据中获取航天信息,是当前国内外航天信息科技研究的热点问题。郭颂[3]等人提出了基于支持向量机(Support Vector Machines, SVM)的主题爬虫采集方法,针对航天领域的信息,通过SVM方法增强指定领域的特征权重,利用训练好的分类模型对网页信息进行筛选。张亚超[4]等人基于深度学习的文本分类方法,采用基于注意力机制的TextRCNN-A文本分类算法,面向航天信息领域实现了自动分类。刘秀磊[5]等人提出一种基于Bert与XGBoost融合模型的航天科技开源信息分类算法,使用Bert提取文本的特征后,采用XGBoost对特征进行分类,提升了航天信息分类的准确率。张玉峰[6]等人基于Web数据挖掘的方法,研究了基于Web文本挖掘的企业竞争信息的获取步骤。黄胜[7]等人面向军事领域,基于爬虫技术从Web开源信息中搜集军事信息,通过层次聚类的算法生成信息主题。王明乾[8]等人通过基于机器学习的文本分类方法从互联网的开源信息中筛选出军事信息,并且分析了多种常用的词向量模型和文本分类模型对于获取开源军事信息的效果。

通过分析当前包含航天信息的文本数据集发现,航天信息文本的内容普遍较长,并且样本数量相对较少[9],这些特点导致现有的提取航天信息的模型准确率较低。为进一步提高航天信息研究工作的效率,将信息较长的文本内容可视化,本文提出了基于有监督对比学习的航天信息文本分类方法。该方法利用BiLSTM捕获长文本的特征信息,使用注意力机制感知全局的文本特征,同时通过对比学习方法增强模型的鲁棒性和泛化能力。即使样本数量少也能够提高模型预测的准确率,从而面向互联网的开源信息筛选出航天信息,同时根据信息生成清晰直观的图像,达到高效获取并可视化开源航天信息的目的,对于航天信息科技领域的研究具有重要意义。

2 相关研究工作

2.1 有监督文本分类

近年来,在文本分类领域,机器学习方法受到了人们的高度关注。深度学习作为实现机器学习的技术之一,也广泛应用在文本分类中。深度学习处理文本分类任务现已慢慢成熟,相对于人工方法,提高了准确率和效率,并在实际工作中效果良好。

Kim[10]将卷积神经网络(CNN)应用在文本分类,即在通过word2vec获取词向量后,利用卷积学习文本特征并完成分类。循环神经网络(Recurrent Neural Networks, RNN)[11]将文本看作词语的序列,目的在于捕获词语之间的依赖关系和文本结构。Peng Zhou[12]等人提出基于RNN的变体,即双向长短期记忆网络,同时加入注意力机制来捕捉句子中最重要的语义信息,可以自动聚焦对分类有关键作用的词语。CNN和RNN在捕获句子中词语之间关系的计算成本会随着句子长度的增加而增大,Transformer[13]解决了这个问题,通过注意力机制为文本中的每个词语计算并记录注意力分数,来模拟每个词语对另一个词语的影响。

2.2 对比学习

在使用深度学习进行有监督的分类任务时,

会出现由于交叉熵损失或者正则化带来的鲁棒性差和泛化性能低等问题[14]。然而对比学习作为一种无监督学习方法，可以使模型中心学习编码器的特征，尽可能所缩小相似样本的距离，同时增大正负样本之间的距离，因此对比学习的研究为有监督学习带来了重大进展，可以有效地解决这些问题。

Khosla[15]提出了一种基于对比性自监督文献的有监督学习损失函数，即利用标签信息，也就是来自同一类的标准化嵌入比来自不同类的嵌入更接近，同时增加不同类之间的距离。Gao[16]在无监督学习的任务中，利用dropout为样本加入噪声，实现数据扩增；在有监督学习的任务中，利用自然语言推理数据集，将蕴含对作为正样本，矛盾对作为负样本，用于后续的对比学习训练。Liang[17]针对在训练深度神经网络时dropout技术的随机性导致的训练和预测不一致问题，提出了一致性训练策略R-drop来正则化dropout，强制dropout生成的不同子模型的输出保持一致。

2.3　文本生成图像

根据文本的语义信息生成图像是计算机视觉和自然语言处理两个领域的综合性任务，也是近几年的研究热点，即输入是一段文本描述，输出则是包含该文本语义信息的图像，便于人们通过视觉更直接得感受文本表达的信息。

Reed[18]等人首次将生成式对抗网络（Generative adversarial Network，GAN）用于解决文本生成图像的问题，提出了GAN-INT-CLS模型，在GAN模型的基础上，同时将输入数据的文本描述训练为句嵌入向量，构建文本和图像匹配的对抗损失，对生成器和鉴别器进一步约束，最后生成分辨率为64×64的图像。Radford[19]等人提出了对比图文预训练（Contrastive Language-Image Pretraining，CLIP）模型。该模型利用两个编码器分别处理文本和图像，然后计算文本特征和图像特征匹配的相似度。利用对比学习思想，最大化正样本对的相似度，最小化负样本对的相似度。该模型可以转换到图像生成、图像检索[20]和视频动作识别等多种下游任务。Ramesh[21]等人在CLIP的基础上提出了unCLIP模型，针对文本生成图像任务，首先利用一个先验模型将文本编码生成图像的嵌入，然后通过图像解码器根据图像嵌入生成给定文本对应的图像。该图像更加真实多样。

3　开源航天信息的获取与图像生成

本文提出了有监督对比学习的文本分类模型，可以在爬取到互联网的开源信息之后进行分析和判断，进而获取航天类别的信息。该模型使用基于注意力机制的BiLSTM作为深度神经网络架构，即BiLSTM-Attention，同时针对神经网络中由于随机进行dropout带来的训练与预测不一致的问题，引入基于R-Drop的对比学习方法训练模型，即在原本的交叉熵损失函数中加入两次前向传播所形成概率分布的KL散度损失，共同进行反向传播，完成参数更新。最后通过unCLIP模型将获取到的航天信息生成对应的图像。

3.1　BiLSTM-Attention模型

基于深度学习的有监督文本分类模型[22]通常由文本预处理、文本的词向量表示、神经网络层和分类层组成。

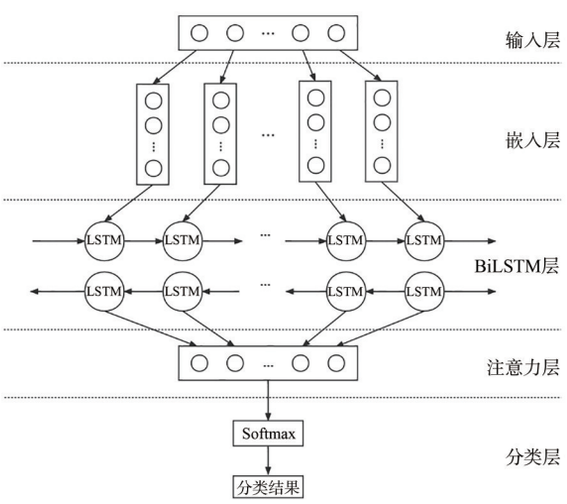

图1　BiLSTM-Attention模型

Fig. 1　BiLSTM-Attention model

第一步将文本中的词语分隔开，然后去除停用词，可以降低对下一步工作的干扰。第二步将文本转化成一种便于计算机识别、数学表达的形式来表达词语的特征，即词向量。第三步将词向量输入到神经网络的模型进行运算，得到变换过的文本序列。最后通过sigmod函数或者softmax函数

计算文本属于某个类别的概率，概率最大的就是模型预测的文本类别。本文使用BiLSTM-Attention作为神经网络模型，该模型的示意图见图1。

3.1.1 输入层

将输入模型的文本进行预处理。首先对文本进行分词处理，本文使用Jieba工具实现分词[23]，将句子中的所有词语快速地全部切分出来，同时利用正则表达式re将文本中的空白字符删除。分词后去除停用词[24]，然后统计词频，选择出现频数较高的词汇生成词汇表。

3.1.2 嵌入层

文本的词向量在嵌入层获取。本文使用word2vec方法[25]训练词向量，便于神经网络层之间的计算，训练时联系上下文，使词向量的语义信息更加丰富。

在Word2vec模型中，主要有两个关键的模型，第一个是CBOW模型，根据某词的上下文信息，确定该词的词向量；第二个是skip-gram模型，根据某词的词向量，确定上下文的词向量。

本文使用gensim框架，选用skip-gram模型训练得到词向量，将词向量的维度限定为200维。

3.1.3 BiLSTM层

LSTM[26]是一种RNN的变体，通过引入“门”结构，有效地解决了RNN的训练过程中激活函数的导数累积而导致的“梯度消失”和“梯度爆炸”的问题。LSTM由遗忘门、输入门和输出门组成，用于控制记忆单元的状态是否存储和更新。

LSTM在t时刻的计算过程可以通过公式(1)～(7)表示，其中f_t、i_t和o_t分别表示遗忘门、输入门和输出门，$\boldsymbol{W}$和$\boldsymbol{b}$表示权重矩阵和偏置向量，x_t和h_t分别表示隐藏层在时刻t的输入和输出，σ和tanh分别表示sigmod和tanh激活函数。

遗忘门根据上一个单元的输出以及当前时刻t的输入共同决定记忆单元中信息的丢弃与否，计算公式为：

$$f_t=\sigma\left(\boldsymbol{W}_f\left[h_{t-1},\ x_t\right]+\boldsymbol{b}_f\right). \tag{1}$$

输入门用于控制网络当前时刻输入的信息有多少保存到记忆单元中，同时计算前一时刻记忆单元的状态更新值$\tilde{C}$，计算公式为：

$$i_t=\sigma(\boldsymbol{W}_i\cdot\left[h_{t-1},\ x_t\right]+\boldsymbol{b}_i), \tag{2}$$

$$\tilde{C}=\tanh(\boldsymbol{W}_c\cdot\left[h_{t-1},\ x_t\right]+\boldsymbol{b}_c). \tag{3}$$

根据遗忘门、输入门以及上一个时刻记忆单元的状态，共同计算当前记忆单元C_t的状态，公式为：

$$C_t=f_tC_{t-1}+i_t\tilde{C}. \tag{4}$$

输出门用于控制记忆单元对于当前时刻t输出值的影响，也就是记忆单元中的哪些信息会输出，公式为：

$$o_t=\sigma(\boldsymbol{W}_o\cdot\left[h_{t-1},\ x_t\right]+\boldsymbol{b}_o). \tag{5}$$

隐藏层节点在时刻t的输出h_t的计算公式为：

$$h_t=o_t\times\tanh C_t. \tag{6}$$

LSTM虽然可以获取长距离的特征信息，但是只完成了神经网络前向的计算，也就是只利用当前时刻之前的信息而忽略了当前时刻之后的信息产生的影响。BiLSTM[27]在LSTM的基础上同时计算前向和后向的信息，可以更加有效地提取文本的特征。本文采用BiLSTM神经网络结构，联合前向和后向的输出，计算后传递给下一个隐藏层节点。

第i个词语的隐藏层节点的输出h_i通过前向传递和后向传递的信息共同决定，计算公式为：

$$h_i=[\overrightarrow{h_i}\oplus\overleftarrow{h_i}]. \tag{7}$$

3.1.4 注意力层

注意力机制[28]最早出现在人类的视觉领域，是关于人类大脑的注意力分配机制，也就是对于重要的信息分配更多的注意力。应用在人工智能领域，注意力机制旨在从较为复杂的信息中有效地提取出关键特征，在调整注意力的权重之后，筛选出关键的信息。在自然语言处理的任务中，注意力机制可以更好地捕捉全局的信息，更加充分地提取文本的特征。

首先确定注意力机制的矩阵$\boldsymbol{H}_t$，也就是BiLSTM层输出的文本特征向量矩阵，计算对应的权重$\boldsymbol{u}_t$，公式为：

$$u_t=\tanh(W_w\boldsymbol{H}_t+\boldsymbol{b}_w). \tag{8}$$

然后对得到的权重矩阵进行归一化处理，计算输入的第t个词语对于判定文本类别的影响程度α_t，公式为：

$$\alpha_t=\frac{\exp\left(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_w\right)}{\sum_t\exp\left(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{u}_w\right)}. \tag{9}$$

最后将向量矩阵进行加权求和，计算得到输出到分类层的句子向量表示$\boldsymbol{h}^*$，公式为：

$$r=\sum_{i=1}\alpha_i \boldsymbol{H}_i\,, \tag{10}$$

$$\boldsymbol{h}^*=\tanh r\,. \tag{11}$$

3.1.5　*分类层*

分类层的输入来自注意力层，经过softmax函数计算后，得到模型预测文本类别的概率，公式为：

$$s_i=\frac{v_i}{\sum_j v_j}\,. \tag{12}$$

3.2　R-Drop技术

在神经网络进行训练的过程中，通过调整隐藏层节点的权重参数，可以学习到输入模型的向量与预测标签之间的关系，但在模型结构较为复杂特别是训练集较少的情况下，会导致过拟合的问题，即模型在训练集的表现很好，但在测试集则表现很差。Hinton[29]等人针对过拟合的问题提出了dropout技术，也就是在神经网络前向传播的过程中，使某个神经元的激活值以一定的概率p停止工作，提高了模型的泛化能力。

虽然dropout技术的效果很好，但是由于dropout的随机性，导致了在训练时随机抽样的子模型和预测时完整模型之间的不一致问题。对比学习的方法可以有效地解决该问题，即通过对同一个样本进行不同的增强得到正样本对，保证相同样本的输出一致性，进一步提高了模型的鲁棒性和泛化能力。Liang等人[17]针对dropout带来的不一致问题，基于对比学习的思想提出了R-Drop技术，用于正则化dropout，可以保证由于dropout产生的不同子模型的输出一致。R-Drop方法的示意图见图2。

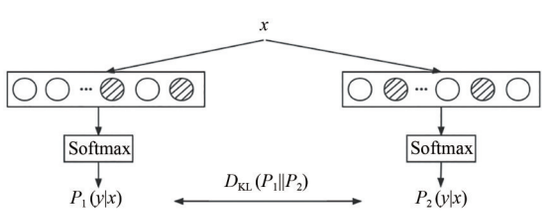

图2　R-Drop示意图（阴影表示随机dropout的神经元）

Fig. 2　R-Drop diagram (Shading represents random dropout neurons)

在训练时的每个批次中，对于同一个样本x_i，前向传播两次之后，由于随机dropout一些隐藏层节点，所以会得到两个不同但差别很小的概率分布，分别表示为$P_1^w(y_i|x_i)$和$P_2^w(y_i|x_i)$，其中y_i表示对应的标签。通过使两个分布之间的KL散度最小化，可以将两个子模型输出的数据样本保持一致，达到缓解训练和预测之间不一致的目的，计算公式为：

$$L_{KL}^i=\frac{1}{2}\left(D_{KL}\left(P_1^w(y_i|x_i)\|P_2^w(y_i|x_i)\right)+D_{KL}\left(P_2^w(y_i|x_i)\|P_1^w(y_i|x_i)\right)\right), \tag{13}$$

式中，$D_{KL}(P_1\|P_2)$表示P_1和P_2两个分布之间的KL散度。

在两次前向传播的过程中，本文使用交叉熵损失函数作为模型学习的目标函数的一部分，公式为：

$$L_{CE}^i=-\log P_1^w(y_i|x_i)-\log P_2^w(y_i|x_i)\,. \tag{14}$$

对于输入数据x_i和对应的标签y_i，最终训练需要最小化的目标函数为：

$$L^i=L_{CE}^i+L_{KL}^i=-\log P_1^w(y_i|x_i)-\log P_2^w(y_i|x_i)+\frac{\alpha}{2}\left(D_{KL}\left(P_1^w(y_i|x_i)\|P_2^w(y_i|x_i)\right)+D_{KL}\left(P_2^w(y_i|x_i)\|P_1^w(y_i|x_i)\right)\right), \tag{15}$$

式中，α表示控制KL散度的权重系数。通过这种方式可以约束模型，将模型的鲁棒性和泛化能力进一步提高。

3.3　unCLIP模型

CLIP作为一种基于对比学习的文本-图像预训练模型，根据数据集中的文本描述和其对应的图像，利用文本编码器和图像编码器分别提取文本和图像的特征，然后计算文本特征和图像特征的余弦相似度，最后进行对比学习，即模型的训练目标为最大化正样本对的相似度，最小化负样本对的相似度。

为了利用CLIP学习到带有语义信息的图像特征来生成图像，unCLIP模型的训练集由文本-图像对(x,y)组成，即图像x和对应的文本描述y。对于给定的图像x，令z_i表示CLIP的图像嵌入，首先利用先验模型$P(z_i|y)$根据文本描述生成对应的CLIP图像嵌入，然后解码器根据CLIP图像嵌入z_i反向生成图像x，这两个阶段对应的图像生成模型$P(x|y)$如式(16)所示：

$$P(x|y)=P(x,z_i|y)=P(x|z_i,y)P(z_i|y)\,. \tag{16}$$

3.4　开源航天信息的获取与图像生成框架

基于上述相关技术，本文提出一种基于开源

文本信息的航天信息与图像生成获取框架，主要包括开源信息采集模块、文本处理模块、模型训练和模型预测模块、文本生成图像模块。

开源信息采集模块通过爬虫技术从互联网上的公开信息中搜集相关信息，信息主要来自科技文献题录数据和航天主题网站的公开数据。

文本处理模块首先对搜集到的数据进行清洗，然后进行分词和去停用词，最后通过 Word2Vec 方法获取文本的词向量。

模型训练模块以词向量作为输入，通过 R-Drop 方法训练 BiLSTM-Attention 模型，对模型进行多轮训练，保存表现最佳的模型。

模型训练之后，将待分类的数据输入模型，根据模型预测的分类结果，实现对开源信息数据进行航天类的信息高效筛选，从而获取航天信息。

最后利用 unCLIP 模型根据航天信息的文本内容生成对应的图像。

基于开源信息的航天信息获取与图像生成框架流程见图 3。

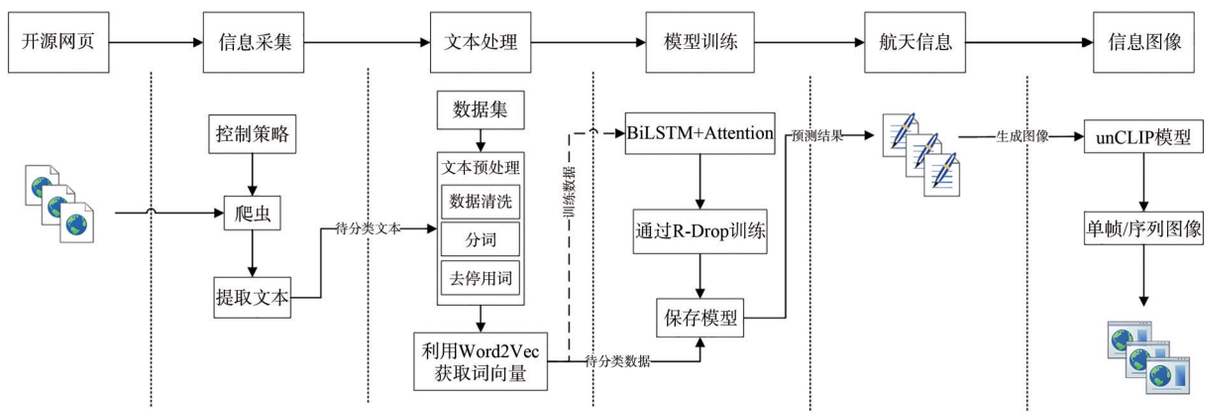

图 3 开源航天信息的获取与图像生成框架流程图

Fig. 3 Flowchart of open source aerospace information acquisition and image generation framework

4 实验与结果分析

4.1 实验环境与数据集

4.1.1 实验环境

本文所涉及到的实验使用的环境配置见表 1。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

实验工具	配置
CPU	Intel(R) Xeon(R) Silver 4210R
GPU	NVIDIA GeForce RTX 3090
内存	128G
操作系统	Ubuntu 20.04.3 LTS
开发语言	Python 3.6

4.1.2 实验数据集

本文的实验采用复旦大学发布的文本分类数据集，该数据集共含有 19 637 篇文档，其中训练集和测试集基本按照 1∶1 的比例划分，涵盖了 20 个类别，主要包括航天、艺术、教育等主题。样本的类别分布不均衡，航天类别的文本数量较少。在训练过程中，由于该数据集的文本较长，统一处理成长度为 600 的样本，同时将原训练集按照 8∶2 的比例随机划分为训练集和验证集。

4.2 评价指标

在文本分类任务中，如何判断分类的效果，即模型的性能如何，是提高分类能力、改善模型的关键。为判断基于有监督对比学习的模型分类效果，本文使用精确率 P、召回率 R 和 F1-Score 作为评价标准[30]。

在测试分类效果的过程中，判断模型预测的标签与正确的标签是否一致，会出现 4 种情况，即二维混淆矩阵，4 种情况分别定义为：

(1)真正例(True Positive, TP)：将正例预测为正例的个数；

(2)假正例(False Positive, FP)：将负例预测为正例的个数；

（3）假反例（False Negative，FN）：将正例预测为负例的个数；

（4）真反例（True Negative，TN）：将负例预测为负例的个数。

精确率（Precision，P）表示预测正确的正例样本数与预测为正例的样本总数的比例，计算公式为：

$$P=\frac{\text{预测正确的正例样本数}}{\text{预测正例的样本总数}}=\frac{\text{TP}}{\text{TP}+\text{FP}}. \tag{16}$$

召回率（Recall，R）表示预测正确的正例样本数与实际为正例的样本总数的比例，计算公式为：

$$P=\frac{\text{预测正确的正例样本数}}{\text{实际正例的样本总数}}=\frac{\text{TP}}{\text{TP}+\text{FN}}. \tag{17}$$

为了综合评价模型的分类效果，结合精确率和召回率可以计算出 F1-Score，对模型的表现做出整体评价，计算公式为：

$$F=\frac{2PR}{P+R}. \tag{18}$$

在多分类的任务中，还有宏平均、微平均和加权平均这 3 个指标对分类效果进行评价。宏平均是先计算各个类别的相关指标后再对它们求算术平均；微平均是不区分样本的类别，将所有样本整体进行计算；加权平均考虑了每个类别的样本数量的权重，对相关指标进行加权计算。

4.3　实验结果与分析

本文采用基于有监督对比学习的文本分类模型，即基于注意力机制的 BiLSTM 模型，同时融合基于 R-Drop 技术的对比学习方法进行训练，该模型的算法流程如图 4 所示。

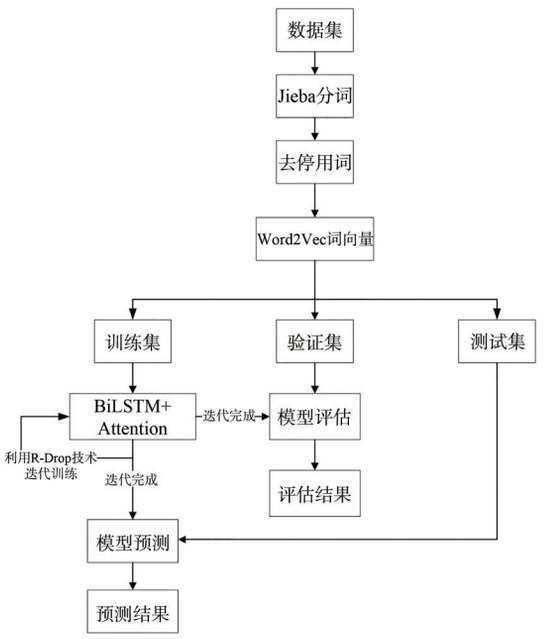

图 4　算法流程图

Fig. 4　Algorithm flowchart

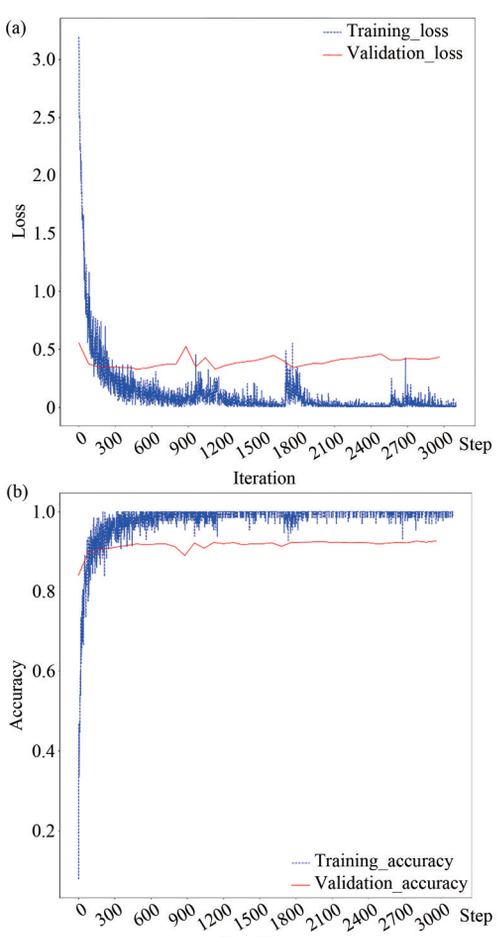

图 5　训练过程的损失（a）和准确率（b）变化

Fig. 5　Loss (a) and accuracy (b) changes during training

在训练过程中，模型在训练集和验证集上的损失和准确率变化曲线见图 5。从图 5 可以看出，随着迭代次数的增加，模型在训练数据集上的损失值一直在变小，而在验证集上的损失值先减小然后增大，说明模型在后续出现了过拟合现象。模型的准确率逐渐增加后趋于稳定，并在训练过程中保存了表现最佳的权重。

本文同时选取 CNN、BiLSTM、Transformer 和 BiLSTM-Attention 模型进行对比实验，关于航天类别文本的分类效果见表 2。

实验结果表明，对于航天类别的文本，本文提出的 BiLSTM-Attention＋R-drop 的分类模型在精确率、召回率和 F1-Score 这 3 个指标都是最高，说明该模型对于获取航天信息具有很好的效果，

表 2 航天类别文本的分类效果

Tab. 2 Classification effect of aerospace category text

模型	Precision	Recall	F1-score
CNN	0.89	0.94	0.91
BiLSTM	0.72	0.80	0.76
Transformer	0.86	0.88	0.87
BiLSTM-Attention	0.94	0.98	0.96
BiLSTM-Attention+R-Drop	0.97	0.98	0.97

有助于提高获取开源航天信息的质量和效率。

在获取到航天类别的信息后，将信息输入到unCLIP模型中，对应生成的图像样例如图6所示，可以更加直观地表达信息，如信息主体、形态和环境等，便于研究人员开展后续工作。

图 6 信息图像

Fig. 6 Information image

为了进一步说明R-Drop方法对于有监督文本分类模型的有效性，本文基于常见的文本分类模型CNN、BiLSTM、Transformer和BiLSTM-Attention做了4组对比实验。由于数据集的样本分布不均，所以采用加权平均作为评价整体分类效果的指标。同时给出当前实验环境下测试集的预测时间，以便对不同模型的处理时效进行对比分析，实验结果见表3。

表 3 各模型的整体分类效果

Tab. 3 Overall classification performance of each model

模型	Precision	Recall	F1-score	预测时间/s
CNN	0.88	0.90	0.88	11.96
CNN+R-Drop	0.89	0.90	0.88	11.20
BiLSTM	0.77	0.79	0.78	35.78
BiLSTM+R-Drop	0.79	0.80	0.80	37.52
Transformer	0.86	0.86	0.84	137.93
Transformer+R-Drop	0.87	0.88	0.87	138.78
BiLSTM-Attention	0.92	0.92	0.92	42.63
BiLSTM-Attention+R-Drop	0.93	0.93	0.93	41.16

实验结果表明，即使在样本分布不均衡的情况下，R-Drop也可以有效地提升多数模型的分类效果，但是对于CNN这种结构过于简单的模型，R-Drop对它的效果影响不大。从各个模型的预测时间可以看出，处理信息的时间与模型的结构复杂程度或者隐藏层神经元的数量有关。R-Drop作为一种训练策略，不会影响模型预测结果的时间。因此可以从实验结果得出，在多数情况下，R-Drop的有效性与模型和样本的分布无关，不会影响模型的信息处理时间，能够优化基于深度学习的分类模型，提升模型的泛化能力和鲁棒性。

5 结 论

本文针对开源航天信息的获取和分析过程中，存在样本的内容过长且相关样本数量较少的问题，提出了基于有监督对比学习的文本分类模型，从互联网爬取到开源的信息之后，通过该模型进行分类并筛选出航天类别的信息。实验结果表明，针对航天类别的文本，融合R-Drop技术的BiLSTM-Attention模型具有较高的精确率、召回率和F1-Score，F1-Score可以达到0.97，比原模型提高了1%，能够做到高效地获取开源航

天信息,并且将信息以图像的形式呈现,更加直观地展示文本内容。本文研究可以提升信息研究人员的工作效率,对于航天科技领域的信息研究工作具有重要意义。

参 考 文 献:

[1] 佟艳春. 基于项目知识管理的航天科技情报协同工作系统研究[D]. 哈尔滨:哈尔滨工业大学,2015.
TONG Y C. Research on the cooperative work system of aerospace science and technology intelligence based on project knowledge management [D]. Harbin: Harbin Institute of Technology, 2015. (in Chinese)

[2] 孔凡芃,刘旭红,刘秀磊,等. 基于BERT模型的航天科技开源情报分类[J]. 北京信息科技大学学报(自然科学版),2021,36(1):28-33.
KONG F P, LIU X H, LIU X L, *et al*. Classification of open source intelligence of aerospace science and technology based on BERT model [J]. *Journal of Beijing Information Science & Technology University*, 2021, 36(1): 28-33. (in Chinese)

[3] 郭颂,边伟,刘洋,等. 基于SVM主题爬虫的航天情报采集应用研究[J]. 电子设计工程,2016,24(17):28-30,34.
GUO S, BIAN W, LIU Y, *et al*. Research on the application of SVM-based focused crawler for space intelligence collection [J]. *Electronic Design Engineering*, 2016, 24(17): 28-30, 34. (in Chinese)

[4] 张亚超. 面向航天情报领域的文本分类算法研究与实现[D]. 西安:西安电子科技大学,2018.
ZHANG Y. The research and implementation on text classification algorithm applied for aerospace intelligence [D]. Xi'an: Xidian University, 2018. (in Chinese)

[5] 刘秀磊,孔凡芃,谌彤童,等. 基于BERT与XGBoost的航天科技开源情报分类[J]. 郑州大学学报(理学版),2021,53(3):15-22.
LIU X L, KONG F P, CHEN T T, *et al*. Research on classification of aerospace science and technology open source information based on BERT and XGBoost [J]. *Journal of Zhengzhou University* (*Natural Science Edition*), 2021, 53(3): 15-22. (in Chinese)

[6] 张玉峰,朱莹. 基于Web文本挖掘的企业竞争情报获取方法研究[J]. 情报理论与实践,2006,29(5):563-566.
ZHANG Y F, ZHU Y. Enterprise competitive intelligence acquisition method based on Web text mining [J]. *Information Studies: Theory & Application*, 2006, 29(5): 563-566. (in Chinese)

[7] 黄胜,郭继光,陆泽健,等. 面向军事领域的Web开源情报主题挖掘研究[J]. 中国电子科学研究院学报,2017,12(4):400-405.
HUANG S, GUO J G, LU Z J, *et al*. Study of web open source intelligence topic mining in military domain [J]. *Journal of China Academy of Electronics and Information Technology*, 2017, 12(4): 400-405. (in Chinese)

[8] 王明乾,倪林,张斌. 基于文本分类的开源军事情报获取方法[J]. 情报探索,2021(7):17-23.
WANG M Q, NI L, ZHANG B. An open source military intelligence acquisition method based on text classification [J]. *Information Research*, 2021(7): 17-23. (in Chinese)

[9] 刘舆,曾德贤,胡远方,等. 基于知识图谱的卫星情报分析方法研究[J]. 情报探索,2021(11):1-7.
LIU Y, ZENG D X, HU Y F, *et al*. Research on satellite intelligence analysis method based on knowledge graph [J]. *Information Research*, 2021(11): 1-7. (in Chinese)

[10] ZHANG Y, WALLACE B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [C]//*Proceedings of the Eighth International Joint Conference on Natural Language Processing. Taipei*, China: ACL, 2017: 253-263.

[11] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning [C]//*Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. New York: IJCAI, 2016: 2873-2879.

[12] ZHOU P, SHI W, TIAN J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification [C]//*Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*). Berlin: ACL, 2016: 207-212.

[13] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need [C]//*Proceedings of the* 31*st Interna-*

tional Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000-6010.

[14] 郭东恩,夏英,罗小波,等. 基于有监督对比学习的遥感图像场景分类[J]. 光子学报,2021,50(7):0710002.
GUO D E, XIA Y, LUO X B, *et al*. Remote sensing image scene classification based on supervised contrastive learning [J]. *Acta Photonica Sinica*, 2021, 50(7): 0710002. (in Chinese)

[15] KHOSLA P, TETERWAK P, WANG C, *et al*. Supervised contrastive learning [C]//*Proceedings of the* 34*th International Conference on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2020: 1567.

[16] GAO T, YAO X, CHEN D. SimCSE: simple contrastive learning of sentence embeddings [C]//*Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*. Punta Cana: ACL, 2021: 6894-6910.

[17] LIANG X B, WU L J, LI J T, *et al*. R-drop: regularized dropout for neural networks [C]//*Proceedings of the* 34*th International Conference on Neural Information Processing Systems*. Virtual: NeurIPS, 2021: 10890-10905.

[18] REED S, AKATA Z, YAN X C, *et al*. Generative adversarial text to image synthesis [C]//P*roceedings of the* 33*rd International Conference on International Conference on Machine Learning*. New York: JMLR. org, 2016: 1060-1069.

[19] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision [C]//*Proceedings of the* 38*th International Conference on Machine Learning*. Virtual: PMLR, 2021: 8748-8763.

[20] 周林鹏,姚剑敏,严群,等. 融合多尺度特征及注意力机制的医学图像检索[J]. 液晶与显示,2021,36(8):1174-1185.
ZHOU L P, YAO J M, YAN Q, *et al*. Medicalimage retrieval with multiscale features and attention mechanisms [J]. *Chinese Journal of Liquid Crystals and Displays*, 2021, 36(8): 1174-1185. (in Chinese)

[21] RAMESH A, DHARIWAL P, NICHOL A, *et al*. Hierarchical text-conditional image generation with CLIP latents [J/OL]. *arXiv*, 2022: 2204.06125.

[22] 刘婷婷,朱文东,刘广一. 基于深度学习的文本分类研究进展[J]. 电力信息与通信技术,2018,16(3):1-7.
LIU T T, ZHU W D, LIU G Y. Advances in deep learning based text classification [J]. *Electric Power Information and Communication Technology*, 2018, 16(3): 1-7. (in Chinese)

[23] 韦人予. 中文分词技术研究[J]. 信息与电脑,2020,32(10):26-29.
WEI R Y. Research on Chinese word segmentation technology [J]. *China Computer & Communication*, 2020, 32(10): 26-29. (in Chinese)

[24] 周钦强,孙炳达,王义. 文本自动分类系统文本预处理方法的研究 [J]. 计算机应用研究,2005,22(2):85-86.
ZHOU Q Q, SUN B D, WANG Y. Study on new pretreatment method for Chinese text classification system [J]. *Application Research of Computers*, 2005, 22(2): 85-86. (in Chinese)

[25] MIKOLOV T, CHEN K, CORRADO G, *et al*. Efficient estimation of word representations in vector space [C]. 1st International Conference on Learning Representations. Scottsdale: ICLR, 2013.

[26] SHI X J, CHEN Z R, WANG H, *et al*. Convolutional LSTM network: a machine learning approach for precipitation nowcasting [C]//*Proceedings of the* 28*th International Conference on Neural Information Processing Systems*. Montreal: MIT Press, 2015: 802-810.

[27] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [J/OL]. *arXiv*, 2015: 1508.01991.

[28] ALLPORT A. *Visual attention*[M]//POSNER M I. Foundations of Cognitive Science. Cambridge: The MIT Press, 1989: 631-682.

[29] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, *et al*. Improving neural networks by preventing co-adaptation of feature detectors [J/OL]. *arXiv*, 2012: 1207.0580.

[30] 奉国和. 文本分类性能评价研究[J]. 情报杂志,2011,30(8):66-70.
FENG G H. Review of performance evaluation of text classification [J]. *Journal of Intelligence*, 2011, 30(8): 66-70. (in Chinese)

作者简介：

齐翌辰（1997—），女，吉林长春人，硕士研究生，2020年于延边大学获得学士学位，主要从事模式识别和多模态方面的研究。E-mail：qiyichen_619@163.com

赵伟超（1992—），男，吉林辽源人，硕士，助理研究员，2016年于西北工业大学获得硕士学位，主要从事普适计算、边缘计算、自然语言处理方面的研究。E-mail：zhaoweichao@ciomp.ac.cn