面向开源科技情报分析的智能文本分类方法研究

彭鹏　徐红姣
（中国科学技术信息研究所，北京 100038）

摘要：随着网络信息的爆发式增长，从海量的网络文本信息中识别有价值的科技情报并对其进行智能分类成为开源科技情报分析的关键。针对开源科技情报文本的特点，构建了面向开源科技情报分析的文本智能去噪与分类一体化模型。结合大语言模型与提示工程的自动标注方法进行噪声数据标注及文本分类数据标注；基于预训练语言模型进行噪声识别与过滤，过滤非科技情报文本；利用多语言预训练模型及蒸馏技术，改进损失函数设计，解决类别分布不均和数据不足的问题，实现在一定程度上提升多标签科技情报文本分类的精度和稳定性的目标。实验结果表明，与TextCNN与BERT方法相比，所提出的方法具有较高的分类性能、更好的鲁棒性和适应性。

关键词：开源科技情报；文本分类；信息过滤；预训练语言模型
中图分类号：TP391.1　DOI：10.3772/j.issn.1673-2286.2025.02.007
引文格式：彭鹏，徐红姣. 面向开源科技情报分析的智能文本分类方法研究[J]. 数字图书馆论坛，2025，21（2）：65-72.

随着科技领域的快速发展，在各种网络平台上发布的开源科技信息规模呈指数级增长，这些信息通常以非结构化或半结构化的文本形式存在，蕴含了大量有价值的科技情报。近年来，科技战略情报分析、企业竞争情报分析、政府舆情情报分析等相关领域的研究人员不断挖掘互联网开源科技情报，开展各类研究。然而，开源科技情报文本的复杂性、多样性和噪声干扰为开源科技情报的有效提取和利用带来了巨大挑战：①发布科技信息的平台众多，涉及政府部门、科研机构、行业企业、科技媒体、社交媒体等，不同来源的数据格式和结构不一致，处理难度较大；②这些平台除了发布科技类的开源情报，还会涉及其他各种非科技类的主题信息；③这些平台发布的科技信息中通常包含一定量的专业术语和领域特定知识，具有领域专业性，通用模型难以直接适用；④针对这类科技文本的高质量标注数据获取成本高，尤其是在新兴科技领域。因此，高效地过滤噪声并实现精准分类成为开源科技情报分析领域的重要研究方向。

近年来，文本信息处理技术取得长足的发展，研究人员针对文本数据的信息过滤与分类问题开展了广泛研究，模型在噪声环境中的鲁棒性和对复杂场景的适应性都得到一定程度提升。为此，本文针对开源科技情报特点，利用文本处理领域高效的智能文本过滤与分类技术，构建面向开源科技文本的去噪与分类一体化模型，提高开源科技情报获取效率和智能分析能力。

1　相关研究

文本分类是自然语言处理的基础任务之一，研究者提出了多种创新性方法以提升分类模型的性能。

收稿日期：2025-01-23

文本数据质量是影响文本分类模型性能的关键因素之一。文本数据中的噪声不仅会干扰分类模型的训练，还会导致模型在预测任务中的准确性下降。因此，研究人员开发了多种信息过滤方法，以应对这些挑战。针对数据本身的噪声问题，Sun等[1]提出基于掩码语言模型的上下文文本过滤方法，通过设计掩码预测机制，使模型能够学习上下文关系并填补语义空缺，从而有效地修复噪声文本。这种方法显著提升了文本表示的质量，尤其适用于非正式文本数据的场景。高雄等[2]针对大规模文本分类任务提出了一种训练语料去噪方法，通过构建“类目-类目关键词”知识库，并结合词嵌入技术检测和清除类别外噪声，显著提高了训练语料的质量，从而有效优化了下游文本分类任务的性能。在弱监督场景中，Ren等[3]提出了结合多源弱监督信号的去噪框架，通过聚合不同来源的噪声信号，提升了模型在无标注数据分类任务中的表现。这一研究为在缺乏标注数据的情况下实现有效分类提供了重要技术支持。除了数据本身的噪声，标签噪声对模型的干扰同样值得关注。Xu等[4]提出标签特定去噪框架，通过建模标签与数据之间的依赖关系，有效减少了错误标签对多标签分类任务的影响，从而显著提升了模型的准确性。但是总的来说，这些方法多依赖特定的噪声类型或场景，缺乏通用性。

标注数据不足是基于弱监督学习的文本分类方法需要解决的核心问题。Xiao[5]提出的极弱监督分类方法结合了词集挖掘和同步过滤技术，在标注数据极少的情况下实现了高效的文本分类。Kargupta等[6]提出MEGClass方法，通过多粒度增强策略，结合不同层次的文本特征，进一步提升了弱监督分类模型的鲁棒性。此外，Park等[7]提出的LIME方法完全摆脱了对种子标签的依赖，通过弱监督策略实现高效分类，为无标注数据的情感分析和主题分类提供了技术支持。虽然弱监督学习方法在标注数据极少的情况下表现良好，但在完全无监督的场景下，模型的稳定性和准确性仍有待提升。

标签语义的利用和融合是多标签文本分类任务中重要的步骤。任彦凝等[8]提出了一种融合标签信息的分类方法，通过结合标签语义与标签关系，显著提升了模型处理复杂样本的能力，增强了多标签分类任务的准确性。在开源情报分析领域，陆佳丽[9]将规则和深度学习方法结合，根据情报源特点构建正则匹配规则进行分类，并针对标题和正文分别建立Bert_TextCNN多标签分类模型，全面提升模型的分类性能。多标签分类虽然具有较高的实用性和灵活性，但面临着标签不平衡、标签语义少、标签之间存在复杂关联等问题，需要根据数据和应用场景进行优化以提升性能。

总的来说，开源科技情报分析通常以新闻报道、社交媒体等为数据源，此类数据源中通常蕴含大量噪声数据，文本内容长短不一且含有一定量的领域专有名词，现有文本过滤和分类方法难以有效适用。因此，本文借助大语言模型进行噪声数据标注及文本分类数据标注，利用标注数据分别训练噪声判别模型和多标签分类模型，以达到高效地从开源平台发布的文本中过滤非科技情报信息，并对有价值的科技情报按主题类别进行多标签分类的目标。

2 模型设计

随着开源科技情报的爆发式增长，文本信息过滤与智能分类成为高效管理和利用开源科技情报的关键。本文构建了一个面向开源科技情报文本的去噪与分类一体化模型（见图1），主要包括如下模块。①数据标注模块：采用结合大语言模型与提示工程的自动标注方法，针对科技文本专业性强、标注成本高的特点，设计特定的提示模板，引导模型生成准确的标签，降低人工标注的成本，获取高质量的噪声标注数据和文本分类标注数据，分别用于后续非科技情报过滤模型及科技情报多标签分类模型的训练。②非科技情报过滤模块：采用预训练语言模型构建噪声判别器，通过学习科技文本的上下文语义信息，识别并过滤数据中的非科技情报数据。该模块通过掩码语言模型和对比学习技术，进一步提升噪声识别的准确性和鲁棒性。③科技情报多标签分类模块：针对科技情报文本分类任务的特点，设计基于预训练语言模型的多标签分类器，利用损失函数缓解标签分布不平衡带来的模型优化问题，提升对少数类标签的识别能力。

2.1 数据预处理

数据预处理包括数据收集、清洗和格式转换。利用爬虫工具从多个来源（新闻网站、学术期刊网站、社交媒体等）自动收集原始数据，从中提取标题、摘要、正文文本数据（学术期刊网站有摘要数据，其他来源没

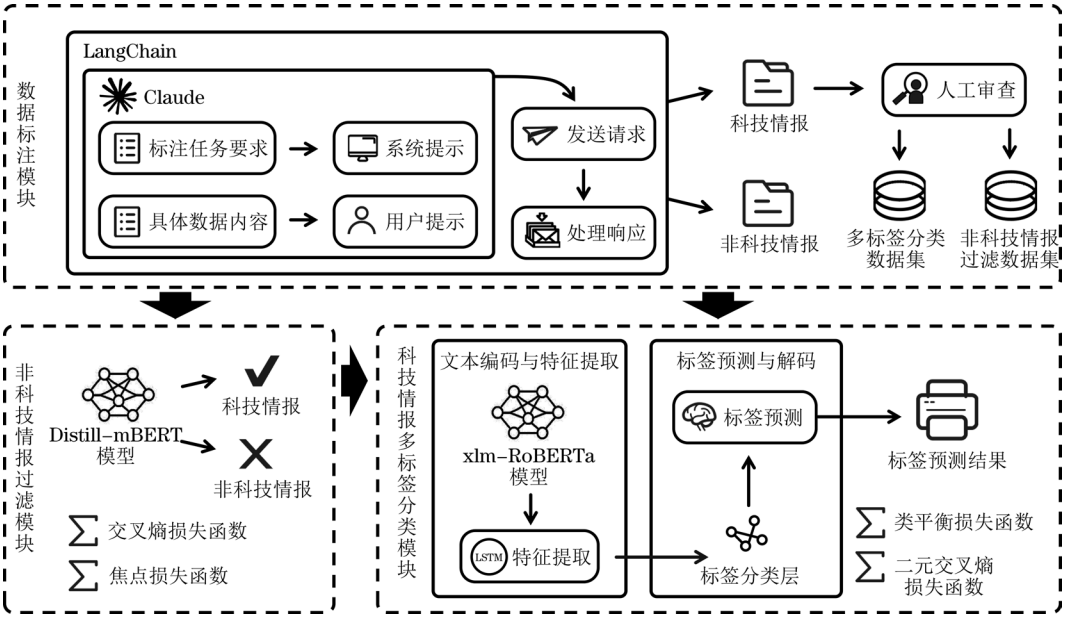


图1　面向开源科技情报文本的去噪与分类一体化模型

有；社交媒体只有正文数据），去除重复和无关内容，得到待标注数据。

2.2　数据标注

针对大规模、高噪声、多领域数据集的科技文本去噪与分类任务，本文提出了一种结合大语言模型与提示工程的自动标注方法。该方法利用大语言模型的文本理解能力实现自动化标注，并设计人工规则，确保标注结果的准确性与一致性。为提高标注效率，采用了函数调用、提示配置和回退机制等技术，构建了灵活的数据标注系统，能够有效处理非结构化文本数据并转化为结构化信息，从而提升数据标注的准确性、稳定性和可靠性。

Claude模型通过Anthropic API提供访问权限，并允许分别设置系统提示和用户提示。对于数据标注任务，将整体提示作为系统提示、具体数据内容作为用户提示，使Claude模型可以更容易、准确地标注数据。引入LangChain库定义函数、发送请求和处理响应。

（1）非科技情报过滤任务。以“标题：[标题]摘要/正文：[摘要/正文]”的形式将标题和摘要/正文组合成模型输入格式，使用大模型对数据进行自动标注，标注结果的分类为“科技情报”和“非科技情报”。为了提高标注数据的质量，人工对大模型标注为“科技情报”的数据进行检查，确保噪声数据判断更加准确并且与人类判断标准相符。采用的提示模板1如下所示。

你是一个噪声识别过滤器。给定数据标题和摘要/正文，你将判断该数据是否为科技情报。

输入将以JSON格式给出，你的输出是一个JSON对象，其属性分别是“reason：判断理由”“sci-tech：布尔值”，为true时表示该数据是科技情报，为false时表示该数据是噪声。“confidence：置信度”，0~1的数表示对判断的自信程度。

（2）科技情报多标签分类任务。在该任务中采用的提示模板2如下所示。

你是一个专业领域分类专家，请根据给定数据标题和摘要/正文，将其归入以下N个领域之一：[C_1, C_2, ⋯, C_N]

任务流程：

①分析标题和摘要/正文：对标题和摘要/正文进行详细分析，结合其内容描述，与上述领域进行匹配。

②列出候选领域：根据分析结果，列出可能相关的所有领域。以<candidates>标签呈现每个备选领域，格式如下：

<candidates>

<candidate id=“C_i”/>

⋯

</candidates>

③选择最恰当的领域：从候选领域中选择最符合内容的领域。描述选择依据，通过自然语言详细阐释你的思考过程。逐个分析不选择其他候选项的原因。

④给出最终分类

以<choice>标签呈现最终归类结果，格式如下：

<choice id="最终领域" />

2.3 非科技情报过滤方法

2.3.1 多语言BERT预训练模型

BERT是一种双向预训练语言模型，能够同时考虑文本的前后文信息，从而提升语义理解能力。BERT在大规模单语言语料上进行预训练，并可通过微调应用于下游任务。mBERT在BERT的基础上扩展至多语言环境，在多种语言的大规模语料上进行预训练，从而能够在不同语言间共享表示空间，适用于多语言文本处理任务。mBERT在跨语言文本分类、翻译和信息检索等任务中表现优越，能够有效处理科技情报分类任务涉及的多语言文本。

在非科技情报过滤任务中，面对多语言文本的复杂性与噪声数据的干扰，采用Distill-mBERT模型进行高效的自动化分类与过滤。Distill-mBERT是通过知识蒸馏技术[10]优化的mBERT模型[11]，在保持多语言理解能力的同时，大幅减少了模型参数量，提高了推理效率，适用于大规模文本分类任务。知识蒸馏过程如式（1）所示。

$$L_{\text{Distill}} = \alpha L_{\text{CE}} + (1-\alpha) L_{\text{KL}} \quad (1)$$

式中：L_{CE}为标准交叉熵损失，衡量模型预测与真实标签的差异；L_{KL}为Kullback-Leibler散度，衡量学生模型与教师模型输出分布的接近程度；α为权衡系数。

2.3.2 损失函数

在非科技情报过滤任务中，将数据划分为科技情报（正类）与无关数据（噪声，负类）。首先，使用Distill-mBERT对由标题和摘要/正文构成的文本进行初步分类。为了进一步提升数据质量，对标注为“科技情报”的样本进行人工复核，以减少噪声干扰，使最终标注结果更符合人类标准。

分类任务采用交叉熵损失函数进行优化，其定义如式（2）所示。

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \text{lb} p_i + (1-y_i)\text{lb}(1-p_i)] \quad (2)$$

式中：N为样本数量；$y_i \in \{0, 1\}$为第i个样本的真实标签（1表示科技情报，0表示噪声）；p_i为模型预测样本为科技情报的概率。

此外，在类别不均衡情况下，采用焦点损失进行优化，如式（3）所示。

$$L_{\text{FL}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_i(1-p_i)^{\gamma}[y_i \text{lb} p_i + (1-y_i)\text{lb}(1-p_i)] \quad (3)$$

式中：α_i为类别权重；γ为调节参数，使模型为难分类样本赋予更高的权重，从而提高分类的鲁棒性。

2.4 科技情报多标签分类方法

本文特别针对不完全标注的情境提出一种新的多标签文本分类方法，目标是开发一种能够从真实标签y_i中学习，并尽可能削弱未知标签y^j对分类器的影响的策略。

面向开源科技情报研究一种多标签文本分类模型，该模型能够处理自动化标注的数据集$\boldsymbol{D}=\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$，每个数据点包含一段文本$\boldsymbol{x}_i$及其对应的多标签向量$\boldsymbol{y}_i$。这个设置中，标签向量$\boldsymbol{y}_i=(y_i^1, y_i^2, \cdots, y_i^n)$，其中$y_i^k$指示第$k$个标签的存在（用1表示）或缺失（用0表示）。由于实际标注过程中可能遗漏一些标签，进一步引入了完美标签向量$\tilde{\boldsymbol{y}}_i=(\tilde{y}_i^1, \tilde{y}_i^2, \cdots, \tilde{y}_i^n)$，其中正例标记为1，负例标记为0，以表示理想状态下的全面标注。实际标签向量$\boldsymbol{y}_i$与理想标签向量$\tilde{\boldsymbol{y}}_i$的差异通过缺失标签向量$\hat{\boldsymbol{y}}_i=(\hat{y}_i^1, \hat{y}_i^2, \cdots, \hat{y}_i^n)$来表达，其中缺失的标签记为1，已标注的为0。

在多标签文本分类的常规任务中，目标是构建一个能从文本$\boldsymbol{x}_i$映射到标签向量$\boldsymbol{y}_i$的分类器。然而，实际应用场景中常见的不完全标注问题意味着实际的标签向量$\boldsymbol{y}_i$可能与理想的标签向量$\tilde{\boldsymbol{y}}_i$不完全一致，需要引入缺失标签向量$\hat{\boldsymbol{y}}_i$。因此，面对不完全标注的挑战，本文旨在开发一种多标签文本分类方法，该方法基于实际可观察的标注$\boldsymbol{y}_i$进行学习，并力图减少未知标签$\tilde{\boldsymbol{y}}_i$对分类器性能的负面影响。

（1）文本编码与特征提取。编码采用xlm-RoBERTa预训练语言模型[12]，将输入文本$\boldsymbol{x}_i$转换为编码矩阵$\boldsymbol{H}$，如式（4）所示。

$$\boldsymbol{H}=\text{xlm-RoBERTa}(\boldsymbol{x}_i) \quad (4)$$

文本编码矩阵$\boldsymbol{H}$的维度为$n \times d$，每行代表文本中每个字符的d维向量表示。

进一步，采用LSTM模块，从编码矩阵$\boldsymbol{H}$中提取文本的高级语义特征向量$\boldsymbol{v}$，用于后续的标签预测，如式（5）所示。

$$\boldsymbol{v}=\mathrm{LSTM}(\boldsymbol{x}_i) \tag{5}$$

（2）标签预测与解码。在解码阶段，文本特征向量$\boldsymbol{v}$通过一个标签分类层转化为最终的标签概率向量$\boldsymbol{p}_1$计算公式如式（6）所示。

$$\boldsymbol{p}_i=\mathrm{Sigmoid}(\boldsymbol{v}\boldsymbol{W}_i+\boldsymbol{b}_i) \tag{6}$$

式中：$\boldsymbol{W}_i$为可学习的权重矩阵；$\boldsymbol{b}_i$为偏置项；Sigmoid函数用于将线性输出映射到（0，1）区间，表示每个标签与文本相关的概率。标签的判定依据概率值与预设阈值的比较结果。

（3）训练与损失函数。在模型训练阶段，针对多标签分类任务的特点，采用二元交叉熵作为损失函数，处理每个标签独立的二分类问题，如式（7）所示。

$$L_{\mathrm{BCE}}=\begin{cases}-\mathrm{lb}(p_i^k), y_i^k=1\\ -\mathrm{lb}(1-p_i^k), \text{其他}\end{cases} \tag{7}$$

式中：p_i^k为模型预测第k个标签属于正类的概率，而y_i^k为实际标签值。尽管二元交叉熵损失函数在多标签分类中表现良好，但当遇到标签分布不平衡的情况时，其优化效率会降低，因为其为所有标签赋予相同的权重，未能考虑到不同标签出现频率的差异。

为了缓解标签分布不平衡的问题，进一步引入类平衡损失函数。类平衡损失函数通过引入一个标签频率相关的调整因子r_{CB}来平衡不同标签的权重，具体计算公式如式（8）~式（9）所示。

$$r_{\mathrm{CB}}=\frac{1-\varepsilon}{1-\varepsilon^f} \tag{8}$$

$$L_{\mathrm{CB}}=\begin{cases}-r_{\mathrm{CB}}\left(1-p_i^k\right)^r \mathrm{lb}\left(p_i^k\right), y_i^k=1\\ -r_{\mathrm{CB}}\left(p_i^k\right)^r \mathrm{lb}\left(1-p_i^k\right), \text{其他}\end{cases} \tag{9}$$

式中：ε是一个预设的超参数，其值范围为[0，1)，用于调整每个标签的权重，以便在一定程度上减少高频标签对损失函数的贡献，从而增加低频标签的影响力；f是训练集中每个标签的出现频率；r是一个非负的调节参数，用于进一步聚焦难以分类的样本。类平衡损失函数能够有效地缓解标签分布不平衡带来的模型优化问题，使模型在处理多标签分类任务时，对于少数类标签具有更好的识别能力。

3　实验与结果

3.1　实验数据与设置

（1）实验数据。从多个来源（新闻网站科技模块、学术期刊网站、社交媒体等）采集2024年10月份发布的信息，借助大语言模型和人工校对相结合的方法，对爬取到的数据进行噪声过滤，识别原始数据中的非科技情报数据，得到科技情报数据。过滤的噪声数据包括：新闻网站科技模块中的科技相关展览、活动、宣传等；学术期刊网站中的学术活动、会议信息、培训信息、征文启事等；社交媒体中的大量非科技文本。具体数据来源分布和数据量如表1所示。

根据样本内容和实际需求，共设置8个类别标签，涉及新材料、人工智能等多个领域，以此数据为基础进行实验，验证所提方法的有效性。每个类别下的样本数量如表2所示。在科技情报多标签分类任务中，采用8∶2的比例划分训练集与测试集。

（2）实验设置。在模型训练流程中使用了Hugging Face的Trainer类来进行管理和控制，这个类提供了一套方便的API，能够处理训练、验证和测试等任务。具体的训练配置如下：使用AdamW优化器，通过引入权重衰减来防止模型过拟合；设置的学习率为2×10^{-5}，Warm-Up比例为总训练步数的5%，权重衰减率为0.1；使用余弦学习率调整策略，在训练过程中动态调整学习率，在训练初期使用较高的学习率，随着训练的进行逐渐降低学习率；焦点损失函数中类别权重α_i，以及类平衡损失函数中超参数ε的值均初始化为0.5，在模型训练过程中不断学习优化，提高准确率。

表1　数据来源分布

类别	数据来源	数据量/条	噪声数据量/条
新闻网站科技模块	CNN（https://edition.cnn.com/science）、《纽约时报》（https://www.nytimes.com/science；https://www.nytimes.com/section/technology）	2 103	823
学术期刊网站	《先进材料》《太阳能学报》《计算机学报》《中国软科学》《航空学报》《生物过程》等26种期刊的官方网站	6 593	1 002
社交媒体	Twitter、Facebook	3 021	1 238

表2　数据类别分布

类别	数据量/条
新材料	1 156
先进能源	1 221
人工智能	1 392
新一代通信网络	1 062
生物科技	1 109
集成电路	1 088
航空航天	943
其他	683
合计	8 654

（3）评价标准。模型性能的评估主要采用了精确率、召回率以及F1分数。精确率衡量模型识别为正类的实例中真正属于正类的比例，召回率则评估模型找出所有正类实例的能力，而F1分数是精确率与召回率的调和平均。精确率、召回率以及F1分数的计算公式如式（10）~式（12）所示。

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}} \tag{10}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}} \tag{11}$$

$$F_1=2\times\frac{P\times R}{P+R} \tag{12}$$

式中：N_{TP}表示正确预测为正类的样本数；N_{FP}表示错误预测为正类的样本数；N_{FN}表示实际为正类但被模型预测为负类的样本数。平均精确率均值（Mean Average Precision，mAP）是所有类别上平均精确率（Average Precision，AP）的平均值。单个类别中平均精确率的计算方法如式（13）所示。

$$P_{A,q}=\frac{\sum_{k=1}^{n}P_q(k)\cdot \mathrm{rel}_q(k)}{N_q} \tag{13}$$

式中：N_q表示结果为q的样本数；$P_q(k)$表示前k个结果中的精确率；$\mathrm{rel}_q(k)$是一个指示函数，表示第k个结果是否相关，如式（14）所示。

$$\mathrm{rel}_q(k)=\begin{cases}1,\text{第}k\text{个样本属于}q\\0,\text{第}k\text{个样本不属于}q\end{cases} \tag{14}$$

设Q是类别总数，则mAP如式（15）所示。

$$P_{mA}=\frac{1}{Q}\sum_{q=1}^{Q}P_{A,q} \tag{15}$$

3.2　非科技情报过滤实验结果

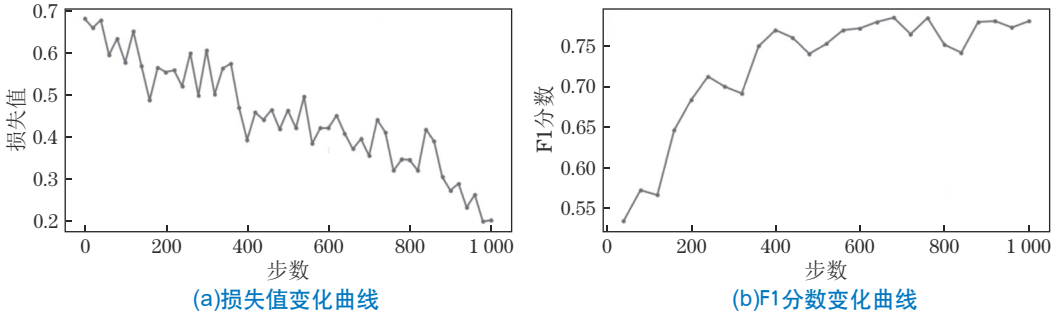
由图2（a）可见，模型在开始阶段没有任何的前置知识，这时它还不会进行分类。随后，损失值迅速下降，在实验过程中逐步降低并稳定。这表明模型在初期经历了一段学习过程，在调整和优化后逐渐适应了过滤任务，展现出更好的泛化能力。

由图2（b）可见，模型的F1分数在开始时已处于较高水平，随后出现波动，但总体上显示出上升趋势。F1分数的上升可能是由于模型在学习过程中逐渐捕捉到了更为精细的特征或模型参数得到调整优化。

模型整体性能表现较为均衡。平均精确率为0.79，表明模型对噪声的过滤具有较高的准确性，能很好地从大量带有噪声的开源科技情报文本中过滤掉无用的非科技情报，精简出有价值的情报文本。平均召回率为0.77，反映出其对噪声的识别较全面。平均F1分数为0.78，进一步验证了模型的稳定性。

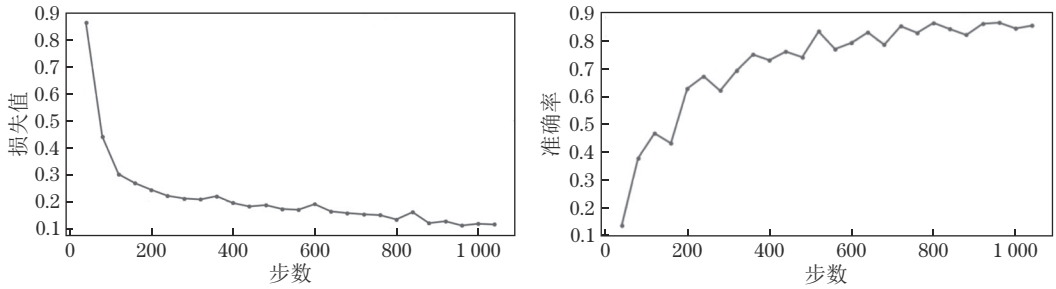
3.3　科技情报多标签分类实验结果

由图3（a）可见，在实验初期，损失值较高，表明模型在开始阶段没有任何的前置知识，不具备学习能力。随后，损失值迅速下降，在实验过程中逐步降低并稳定。这可能表明模型在初期经历了一段学习过程，在调整和优化后逐渐适应了多标签分类任务，展现出更好



(a)损失值变化曲线　　(b)F1分数变化曲线

图2　非科技情报过滤损失值与F1分数变化曲线



(a)损失值变化曲线　(b)准确率变化曲线

图3　科技情报分类损失值和准确率变化曲线

的泛化能力。

由图3（b）可见，在实验初期，模型的准确率较低，随着训练的进行显著提高，随后出现波动，但总体上显示出上升趋势，表现了该方法的可行性。准确率上升是由于模型在学习过程中逐渐捕捉到了更为精细的特征，以及模型参数得到调整优化。

选择文本分类领域的两个经典基线模型TextCNN和BERT进行对比实验，模型性能对比结果如表3所示。尽管从召回率来看，所提模型结果稍低于BERT，但其精确率比其他两种方法有一定程度提升，F1分数比TextCNN提升3.7%、比BERT提升1.2%。综合对比来看，提出的模型的表现优于基线模型，能在一定程度上提高文本分类性能。

表3　模型分类性能比较

类别	精确率	召回率	F1分数
TextCNN	0.78	0.80	0.79
BERT	0.80	0.82	0.81
所提方法	0.84	0.80	0.82

为进一步分析所提模型性能，在科技情报多标签分类任务中评估了模型在各个类别上的性能，包括先进能源、新一代通信网络、人工智能等8个预定义分类，结果如表4所示。整体上，模型的mAP达到0.84，反映了其综合表现良好。在样本量较少的数据上或者特征不太明显的类别，模型性能不佳，需要更多数据或是针对性的模型调整来改进对这些类别的预测效果。精确率和召回率在某些类别间有明显差异，这表明模型在特定领域可能存在过拟合现象或者对某些类别的特征学习不足，导致无法泛化到所有类别。比如，新一代通信网络类别的精确率、召回率较低，这可能是由于其涉及太赫兹频段通信、空天地一体化网络，尤其是人工智能和机器学习等关键技术，与模型的人工智能标签存在一定程度重叠，导致分类效果不佳。相反，先进能源、生物科技和新材料3个类别的主题定义相对清晰，因此分类性能表现优于其他类别。

表4　模型在不同类别上的评估结果

类别	精确率	召回率	F1分数
先进能源	0.90	0.78	0.84
新一代通信网络	0.71	0.78	0.74
生物科技	0.82	0.85	0.84
集成电路	0.88	0.83	0.85
人工智能	0.79	0.78	0.78
航空航天	0.88	0.75	0.81
新材料	0.88	0.83	0.86
其他	0.79	0.75	0.77
平均值	0.84（mAP）	0.80	0.82

综上所述，所提模型具有较高的稳定性，在大多数类别上表现出了较高的精确率和召回率，说明模型能够有效将开源科技情报文本按研究人员的重点关注主题类别进行分类，从而帮助其方便、迅速地从海量开源科技情报文本中按需筛选自己关注的类别进行浏览阅读，提高信息的利用效率。

4　结论

本文针对开源科技情报的智能分类方法开展了研究，通过引入基于预训练语言模型的噪声判别算法和多标签文本分类模型，在节省人工标注成本的同时一定程度上提升了模型的分类性能。实验结果表明，该方法在多个复杂场景中具有较高的鲁棒性与适应性。在非科技情报过滤方面，提出的过滤方法有效减少了噪声干扰，提升了文本质量，为后续的分类任务提供了可靠的数据支持。在科技情报多标签文本分类方面，构建了高效的分类模型，充分利用多语言预训练模型及其蒸馏版本进行优化，使得分类模型在计算效率和预测性能之

间达成平衡。通过改进损失函数设计，有效应对了类别分布不均和标注不足的挑战。本文采用大模型进行数据标注，有效降低了人工标注数据的成本。未来的研究可以进一步探索增强模型在低资源环境下的表现，并针对特定领域数据进行深度定制与优化。

参考文献

[1] SUN Y F，JIANG H M. Contextual text denoising with masked language model[C]//Proceedings of the 5th Workshop on Noisy User-generated Text（W-NUT 2019）. Stroudsburg：Association for Computational Linguistics，2019：286-290.

[2] 高雄，韩红旗，王力，等. 大规模文本分类的训练语料去噪方法研究[J]. 情报工程，2021，7（4）：117-126.

[3] REN W D，LI Y H，SU H T，et al. Denoising multi-source weak supervision for neural text classification[C]//Findings of the Association for Computational Linguistics：EMNLP 2020. Stroudsburg：Association for Computational Linguistics，2020：3739-3754.

[4] XU P Y，JING L P，YU J. Enhancing multi-label text classification under label-dependent noise：a label-specific denoising framework[C]//Findings of the Association for Computational Linguistics：EMNLP 2024. Stroudsburg：Association for Computational Linguistics，2024：5674-5688.

[5] XIAO L. Extremely weakly-supervised text classification with wordsets mining and sync-denoising[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies（Volume 1：Long Papers）. Stroudsburg：Association for Computational Linguistics，2024：7167-7179.

[6] KARGUPTA P，KOMARLU T，YOON S，et al. MEGClass：extremely weakly supervised text classification via mutually-enhancing text granularities[C]//Findings of the Association for Computational Linguistics：EMNLP 2023. Stroudsburg：Association for Computational Linguistics，2023：10543-10558.

[7] PARK S，LEE J. LIME：weakly-supervised text classification without seeds[C]//Proceedings of the 29th International Conference on Computational Linguistics. Stroudsburg：Association for Computational Linguistics，2022：1083-1088.

[8] 任彦凝，陈俊霖，刘群. 一种融合标签信息的多标签文本分类方法[J]. 小型微型计算机系统，2024，45（3）：584-590.

[9] 陆佳丽. 基于Bert-TextCNN的开源威胁情报文本的多标签分类方法[J]. 信息安全研究，2024，10（8）：760-768.

[10] 黄震华，杨顺志，林威，等. 知识蒸馏研究综述[J]. 计算机学报，2022，45（3）：624-653.

[11] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2024-12-12]. https://arxiv.org/abs/1810.04805v2.

[12] PRYTULA M. Fine-tuning BERT，DistilBERT，XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews[J]. Artificial Intelligence，2024，29（2）：85-97.

作者简介

彭鹏，女，副研究员，研究方向：企业管理、信息资源管理。
徐红姣，女，硕士，高级工程师，通信作者，研究方向：多语言信息分析，E-mail：xuhj@istic.ac.cn。

Intelligent Text Classification Method for Open-Source Technology Intelligence Analysis

PENG Peng XU HongJiao
(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract: With the explosive growth of network information, identifying valuable technology intelligence from massive network text information and classifying it intelligently have become the key to open-source technology intelligence analysis. Based on the characteristics of open-source technology intelligence texts, this paper constructs an integrated model of text denoising and classification for open-source technology intelligence analysis. It combines large language model with automatic annotation method of prompt engineering to annotate noise data and text classification data. A pre-trained language model is constructed for noise recognition and filtering, filtering non-technology intelligence texts. Multilanguage pre-trained models and distillation techniques are used to improve the loss function design, solve the problems of uneven class distribution and insufficient data, and achieve the goal of improving the accuracy and stability of multi-label technology intelligence text classification to a certain extent. The experimental results show that compared with TextCNN and BERT methods, the method proposed in this paper has higher classification ability, robustness, and adaptability.

Keywords: Open-Source Technology Intelligence; Text Classification; Information Filtering; Pre-Trained Language Model

（责任编辑：王玮）