

[Classifying Free Texts Into Predefined Sections Using AI in Regulatory Documents: A Case Study with Drug Labeling Documents | Chemical Research in Toxicology \(acs.org\)](#).[↗]

标题

《Classifying Free Texts Into Predefined Sections Using AI in Regulatory Documents: A Case Study with Drug Labeling Documents》

使用人工智能（AI）技术，将监管文档中的自由文本自动分类到标准化的章节中：

一个根据药品标签文档进行的研究

AI总结

论文《Classifying Free Texts Into Predefined Sections Using AI in Regulatory Documents: A Case Study with Drug Labeling Documents》介绍了使用人工智能（AI）技术，特别是基于双向编码器表示（BERT）的自然语言处理（NLP）方法，将药品标签等监管文档中的自由文本自动分类到标准化的章节中，以改进药品安全和有效性评审流程。以下是论文的翻译与分析总结。

翻译概述：

- **引言：**美国食品和药物管理局（FDA）的监管文档通常非常庞大，涉及广泛的信息。由于这些信息的组织结构不一，FDA提出了标签结构（如“Physician Label Rule” PLR）来规范药品标签信息。然而，许多文档仍然格式不当，影响了信息的有效获取。为了改善这一情况，作者们探讨了利用BERT模型将非结构化的文本分类为标准章节的可行性。

- **方法**：研究使用FDA的药品标签文档作为测试对象，将这些文档分为结构良好的PLR格式和其他非PLR格式（包括SmPC）。BERT模型被训练来处理这些文本，并在二分类和多分类任务中表现出较高的准确率。使用的数据集涵盖了超过5.5万份文档和1900多万句文本。
- **模型及其改进**：除了BERT外，作者还尝试了改进版的BERT模型，如ALBERT、DistilBERT和RoBERTa，以提高效率。每个模型经过10个训练周期，使用PLR、非PLR和SmPC数据进行测试。研究表明，BERT及其变种在分类任务中表现出色，特别是在PLR数据集上的表现最佳。
- **结果**：对于二分类任务（如“使用说明”和“警告和注意事项”），BERT模型的准确率接近95%；而在多分类任务中，准确率稍有下降，但仍能达到82%。在面对非结构化的文档（如非PLR或SmPC）时，模型的表现也具有一定的鲁棒性。
- **讨论**：作者指出，未来研究可以进一步扩展分类任务的类别，提高模型的精确度。此外，研究发现使用段落而非句子作为输入，可以进一步提高模型的准确性，但在时间和资源上有更高的要求。
- **结论**：本研究证明了基于BERT的深度学习模型能够有效地将非结构化的药品标签信息自动分类到标准化章节中，展示了AI在监管科学领域的应用潜力。

分析总结：

1. **研究背景与意义**：本文的研究旨在解决药品监管文档格式不一致的问题，提出通过AI技术自动分类自由文本，从而简化FDA的评审流程。这一方法有助于提高药品安全性评估的效率，减少文档信息不当组织对药品评审带来的负面影响。
2. **方法与模型选择**：本文使用了BERT模型及其变种，这些模型在自然语言处理领域表现出色，尤其适用于文本分类任务。BERT通过双向理解文本中的上下文信息，能够在多种任务中实现高效的分类效果。
3. **结果与应用前景**：模型的高准确率证明了AI技术在自由文本分类中的应用潜力，尤其是在FDA监管流程中的应用，可以帮助自动处理和分析药

品标签文档，提升审查效率。未来研究可以进一步优化模型，使其适应更多的文本格式和语言环境。

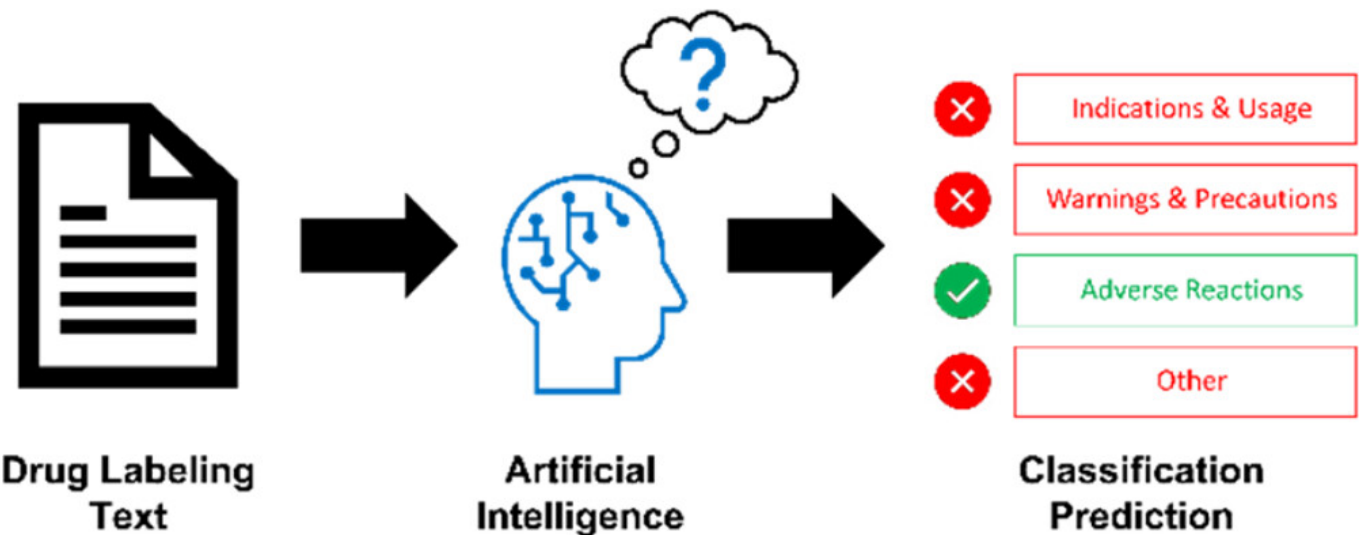
4. **局限性与未来研究方向：** 尽管研究结果令人鼓舞，但当前模型仍存在一些局限性。尤其是在处理非PLR和SmPC格式的文本时，模型的分类准确性相对较低。未来研究可以进一步扩展训练数据集，并尝试其他深度学习架构，如基于GPT的模型，以应对更多的应用场景。

总结：

本文的研究展示了AI在分类非结构化文本中的潜力，尤其是在药品标签等监管文档中的应用。通过优化BERT模型，研究人员成功地实现了自动分类，未来可以进一步扩展其应用，推动监管科学的智能化进程。

原文全译

摘要



美国食品和药物管理局（FDA）的监管流程通常涉及多个审核者，他们专注于与各自审核领域相关的信息。因此，向监管机构提交文件的制造商被要求使用某种结构组织内容，以便信息能够被轻松分配、检索和审查。然而，这一做法并不总是被正确遵循，导致部分文件没有良好结构，相似的信息散布在不同部分，妨碍了整体数据的有效访问和审核。为改善这种常见现象，我们评估了一种基于人工智能（AI）的自然语言处理（NLP）方法——双向编码

器表示（BERT），该方法能够自动将自由文本信息分类为标准化章节，从而支持对药品安全性和有效性的全面审查。特别是，我们使用FDA的标签文件作为概念验证，采用了由“Physician Label Rule”定义的标签章节结构来开发模型。该模型随后在结构良好的标签文件（基于PLR的标签）以及结构较差或不同结构的文档（非PLR和SmPC标签）上进行了评估。在训练过程中，模型在二分类任务和多分类任务中分别实现了96%和88%的准确率。对于二分类模型，PLR、非PLR和SmPC测试数据集的测试准确率分别为95%、88%和88%，而对于多分类模型，测试准确率分别为82%、73%和68%。我们的研究表明，利用AI语言模型自动将自由文本分类为标准化章节，可以成为支持审核流程的一种先进监管科学方法，有效处理未格式化的文档。

引言

监管文档通常非常庞大，涵盖广泛的信息。个别审查员通常会根据其审查任务专注于特定的信息集，例如安全性或有效性。因此，监管文档需要使用一种结构化的方式来进行组织，使得信息能够被轻松分配、检索和审查。不幸的是，尽管结构化文档被认为对改善监管审查流程至关重要，但这种结构并没有始终如一地得以应用。例如，FDA的药品标签文档在过去40年中发生了变化。

2005年，FDA发布了《行业指南：以电子格式提供监管提交文件——标签内容》（Guidance for Industry: Providing Regulatory Submissions in Electronic Format – Content of Labeling），为结构化产品标签（Structured Product Labeling, SPL）格式的监管提交文件提供了指导。在SPL格式下，文本被预先标注在特定的标签章节中，使得FDA的研究人员和审查员可以更轻松地检索和分析文档中的文本信息。

然而，许多标签文件仍然格式不佳，妨碍了对其所含信息的访问和使用。此外，未来标签文档的结构可能会再次修订，以提高信息的清晰度和组织性。这种情况引发了一个问题：如何将早期版本的标签信息与未来的格式进行标准化，以便在审查同一治疗领域或药物类别的药物时，能够提供关于药物安全性和有效性的整体视图。此外，不同国家有各自的结构化信息格式；为了扩展FDA的知识库，其他国家的监管文件也必须转换为FDA使用的格式。

文本分类是自然语言处理（NLP）中的主要任务之一，随着基于Transformer的语言模型的发展，NLP取得了显著进步。尽管文本分类模型通常专注于情感分析，但多个研究已表明NLP在将自由（非结构化）文本分类到预定义类别中具有应用潜力。例如，Dernoncourt和Lee研究了约200,000篇医学摘要的句子，将其分类为背景、目标、方法、结果或结论部分，准确率超过90%。此外，在区分是否包含宣传手段的句子方面，基于BERT的分类模型在18种不同的宣传技术和类别上实现了55%-80%+的分类准确率。

一些研究还尝试根据药品标签文本中的模式对药品或药品标签文档进行分类或分组。例如，2011年，Bisgin等人使用无监督文本挖掘方法（主题模型）对794种FDA批准药物的标签文档进行分类，发现具有相似安全问题、治疗用途或两者兼有的药物。2019年，Wu等人通过层次聚类分析对367种单成分药物的标签文档进行分组，以发现药品警示框部分中MedDRA（医疗监管活动词典）首选术语和不良反应的相似模式。这些研究表明，药品标签文档的文本中隐藏了某些模式，能够通过计算机辅助或机器学习技术对这些文档及其章节进行分组。

本研究基于上述概念，开发了一种语言模型，能够自动将FDA药品标签文档中的自由文本分类为定义好的标准化章节。我们选择PLR标签格式作为构建语言模型的标准。鉴于药品标签章节的多样性，我们在训练分类模型时使用了多种类别配置，以研究不同情况下模型性能的变化，例如类别数量增加时性能的变化。该模型使用PLR格式的标签文本进行训练，并随后应用于非PLR格式的标签文档，这通常出现在较旧的标签文档中。我们还将该模型应用于分类英国药品标签文档中的SmPC格式。

材料与amp;方法

美国FDA药品标签

基于PLR，FDA处方药标签文件通常有两种格式。PLR格式由FDA在2006年首次发布，是处方药标签格式的黄金标准，所有2001年6月以后提交的处方药标签必须符合该格式（2001年6月至2006年提交的文件需要追溯更新为

PLR格式)。另一方面，非处方药标签（如非处方药）和2001年之前批准的处方药标签则不需要采用PLR格式，被视为非PLR格式的文件。

值得注意的是，PLR格式和非PLR格式之间存在一些差异。FDA认为PLR格式“增强了人类处方药的安全和有效使用.....并减少了由于药物信息误解或误用而导致的副作用的数量”。此外，PLR格式使处方信息对医疗从业者、患者和研究人员更加易于获取。FDA声称，PLR格式采用了现代化的方法来传达准确的药物使用信息，使处方信息“更容易与电子处方工具和其他电子信息资源一起使用”。尽管PLR格式相较于非PLR格式有诸多优势，但一些较早的标签文件仍保留了非PLR格式；只有2001年至2006年间批准的新药申请（NDA）和生物制品许可申请（BLA）被追溯更新为符合PLR格式。

鉴于PLR格式在从药品标签文件中检索和使用信息时的优势，本研究的目标是将所有类型的药品标签组织成适当的PLR格式章节。为此，我们从DailyMed的完整发布中获取了45,626份处方药标签文件（截至2022年2月28日）。其中，29,709份（65%）为PLR格式，15,917份（35%）为非PLR格式。使用Python和自然语言工具包（NLTK）库，我们提取了总计17,453,802句子。这些句子进一步与逻辑观察标识符、名称和代码（LOINC）映射，LOINC是用于确定标签文件中的章节位置的官方代码。由于PLR和非PLR标签文件使用不同的LOINC代码，因此必须分别处理它们。

Figure 1

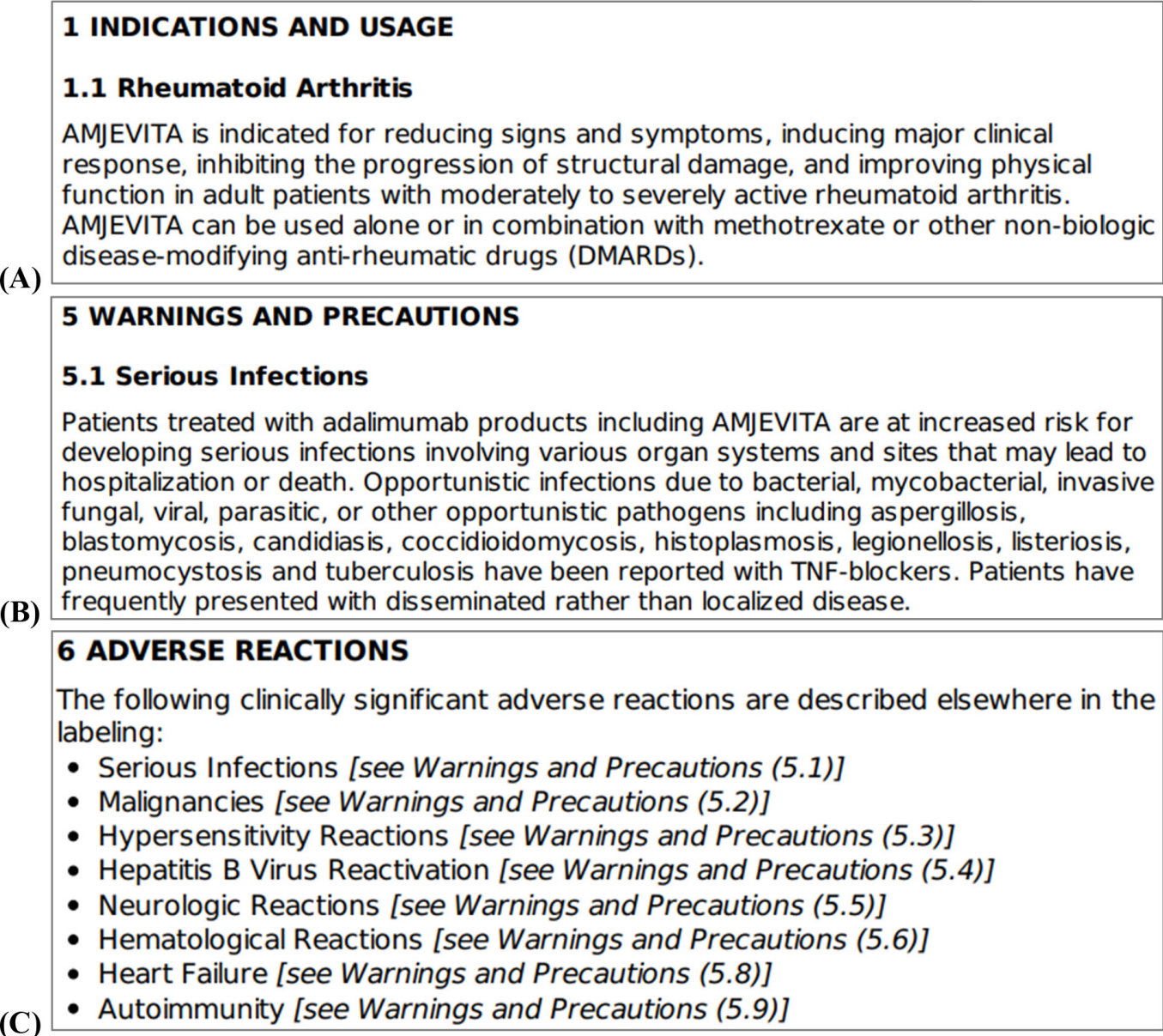


图1展示了FDA处方药标签文件中几个主要章节的内容示例。以阿达木单抗-atto注射液的标签文件为例（本研究与此无关） ，

- (A)摘录了“适应症和用途”部分
- (B)摘录了“警告和注意事项”部分
- (C)摘录了“不良反应”部分。

这些图为这些关键药品标签章节中的信息提供了大致概览。

英国药品标签

在英国，主要的药品标签文件是SmPC（药品特性总结）。这些文件为医疗专业人员提供了至关重要的信息，例如如何使用和处方药物。SmPC由制药公司根据其研究撰写并更新，由英国或欧洲的药品许可机构进行检查和批准。SmPC类似于FDA监管的处方药标签，每个文件包含类似于PLR或非PLR格式中的标签章节。例如，SmPC中的“治疗适应症”章节包含的信息与FDA标签中的“适应症和用途”类似。

为了确定本研究中产生的语言模型是否适用于外部药品标签文件，我们从英国药品数据库Electronic Medicines Compendium（电子药品汇编）中收集了9,580份SmPC（截至2022年6月26日）。使用类似的数据处理技术，我们收集了2,180,388句子。

数据集概述

Table 1. Summary of Datasets

Data set	Origin	No. Documents	No. Sentences
1. PLR	US – DailyMed	29,709	14,072,802
2. non-PLR	US – DailyMed	15,917	3,380,819
US Total		45,626	17,453,802
3. SmPC	UK – EMC	9580	2,180,388
Overall Total		55,206	19,634,190

表1总结了本研究使用的三个数据集。总体而言，我们收集了超过55,000份标签文件和超过1,900万句子，来源分别为：（1）PLR、（2）非PLR、（3）SmPC格式文件。

建模算法

在主要任务中，我们使用BERT来训练句子分类模型。BERT是一种用于多种NLP任务的最先进的语言模型，包括文本分类、问题回答和下一句预测。顾名思义，BERT是一个具有Transformer架构的多层编码器，是一种基于注意

力机制的模型。BERT预训练于BooksCorpus（8亿个词）和维基百科（25亿个词）。由于其自注意机制，训练后的模型可以通过在模型架构上训练不同的头来进一步微调，适用于多种任务。

除了基本的BERT模型（BERT-base）外，还有许多不同的BERT模型。为了探索模型对结果的影响，我们对几种替代的BERT模型进行了微调和测试，分别是ALBERT、DistilBERT和RoBERTa。这些模型之所以被选中，是因为它们具有独特且经过验证的功能。ALBERT通过参数缩减技术来降低内存消耗并加快BERT的训练速度，同时最大限度地减少语言理解能力的损失。DistilBERT在预训练阶段利用知识蒸馏来减少模型规模并加快训练速度，同时保留了大部分语言理解能力。RoBERTa通过更长时间的训练和更多的数据、动态改变训练数据应用的掩码模式，使其在性能上与BERT发布后的模型一样好，甚至更好。我们还使用了随机森林（RF）和支持向量机（SVM）模型作为基线。

表2. 所选模型概述

模型	描述
RF	随机森林（RF）分类器是一种机器学习算法，通过结合多个决策树的输出生成一个单一结果。具体而言，它适用于从数据集的子样本中拟合若干决策树分类器，并通过平均值来提高预测准确性并控制过拟合。
SVM	支持向量机（SVM）是一种监督学习算法，主要用于分类问题。在该算法中，每个数据项被绘制为一个n维空间中的点，每个特征表示一个坐标。然后，算法通过使用极点/向量（即支持向量）找到最佳决策边界（即超平面），以分离不同类别。
BERT	BERT（双向编码器表示的Transformer）是一种通用语言模型，预训练于BookCorpus（8亿词）和英文维基百科（25亿词）数据集。通过使用自注意力机制，BERT可以通过在模型架构顶部训练不同的层来完成新任务，并通过新数据进行微调，使其成为许多基于Transformer的语言模型的基础。在发布时，BERT在多个通用语言理解评估（GLUE）基准任务中实现了最先进的性能。基础模型包含大约1亿个参数。

模型	描述
ALBERT	ALBERT（轻量版BERT）通过因子化嵌入参数化和跨层参数共享等参数缩减技术，解决了BERT的内存限制和长训练时间问题。尽管基础模型只有约1200万个参数，但在语言理解上几乎没有损失，表现几乎与BERT相同。
DistilBERT	DistilBERT是BERT的精简版本，通过在BERT的预训练阶段应用知识蒸馏，解决了大规模Transformer语言模型的计算效率问题。此方法将BERT的大小减少了40%，同时保留了97%的语言理解能力，并且训练速度提高了60%。
RoBERTa	RoBERTa是一种经过大幅优化的BERT预训练方法，它通过延长训练时间、增加数据量、在更长的序列上训练并动态改变训练数据的掩码模式来对BERT进行改进。更详细地说，RoBERTa的训练数据集由BookCorpus和英文维基百科（16GB）、CC-News（76GB）、OpenWebText（38GB）和Stories（31GB）数据集组成。通过这些改进，RoBERTa在多个任务中超越了BERT，取得了最先进的结果。基础模型包含约1.1亿个参数。

模型微调

为了探讨模型预测特定药品标签句子所属章节的能力，我们开发了一系列二分类和多分类任务，重点关注几个关键的PLR章节：（1）“适应症和用途”，（2）“警告和注意事项”，（3）“不良反应”。在主要的二分类任务中，终点为“适应症和用途”和“警告和注意事项”，由于这些章节的文本易于区分，因此预计能够为模型的语言理解能力提供坚实的基础。相反，主要的多分类任务则包含这些终点外加“不良反应”和“其他/未知”（包括所有剩余的药品标签章节），从而给模型带来更具挑战性的任务，并提供其区分多个药品标签章节的能力衡量标准。

随后，准备了训练和测试数据集。对于每个分类建模任务，从每个数据集中获取每个终点的10,000个句子，确保数据集类别平衡。

对于基于BERT的模型，这些数据集使用各自的HuggingFace自动标记器进行了标记。处理和标记后的训练数据集按照80%用于训练，20%用于验证进行

划分。对于每个基于BERT的模型，使用PLR格式的训练数据集对其进行了微调。更具体地说，每个模型在10个训练周期内，使用模型的默认参数和“准确率”指标进行了微调。模型仅进行了10个周期的微调，因为在这一阶段之前或之后性能的提升趋于平稳。最后，每个模型使用PLR、非PLR和SmPC格式的测试数据集进行了评估，每个数据集包含每个终点的新/未见过的10,000个句子。

另一方面，对于RF和SVM模型，使用Python的NLTK库对数据集进行了标记和处理。更详细地说，使用“word_tokenizer”函数对句子进行了标记化，使用“WordNetLemmatizer”对每个词进行了词形还原。然后，使用scikit-learn库中的“TfidfVectorizer”工具对句子进行了向量化处理。同样，训练数据集按80%用于训练，20%用于验证进行了划分。对于RF和SVM模型，分别使用PLR格式的数据集进行了训练。最后，使用PLR、非PLR和SmPC格式的测试数据集，对每个模型进行了评估，并使用scikit-learn库中的“accuracy_score”函数计算了模型的准确率。

模型可解释性分析

在获取上述模型的结果后，计算了Shapley加法解释（SHAP）值，以确定在三种药品标签文档格式的不同章节中哪些词语具有最大影响。Shapley值的平均值提供了特定词语对模型预测给定句子章节终点的相对影响。此外，SHAP值可以通过图形化方式可视化，以展示导致某个预测的句子部分，以及那些没有起到决定性作用的部分。

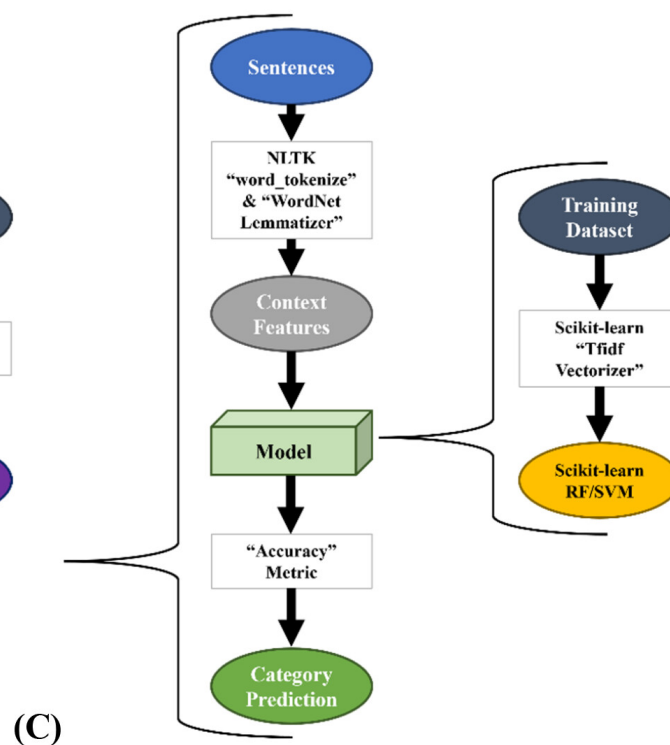
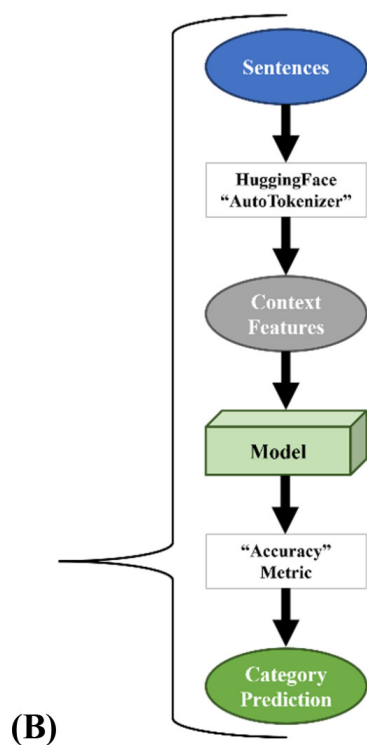
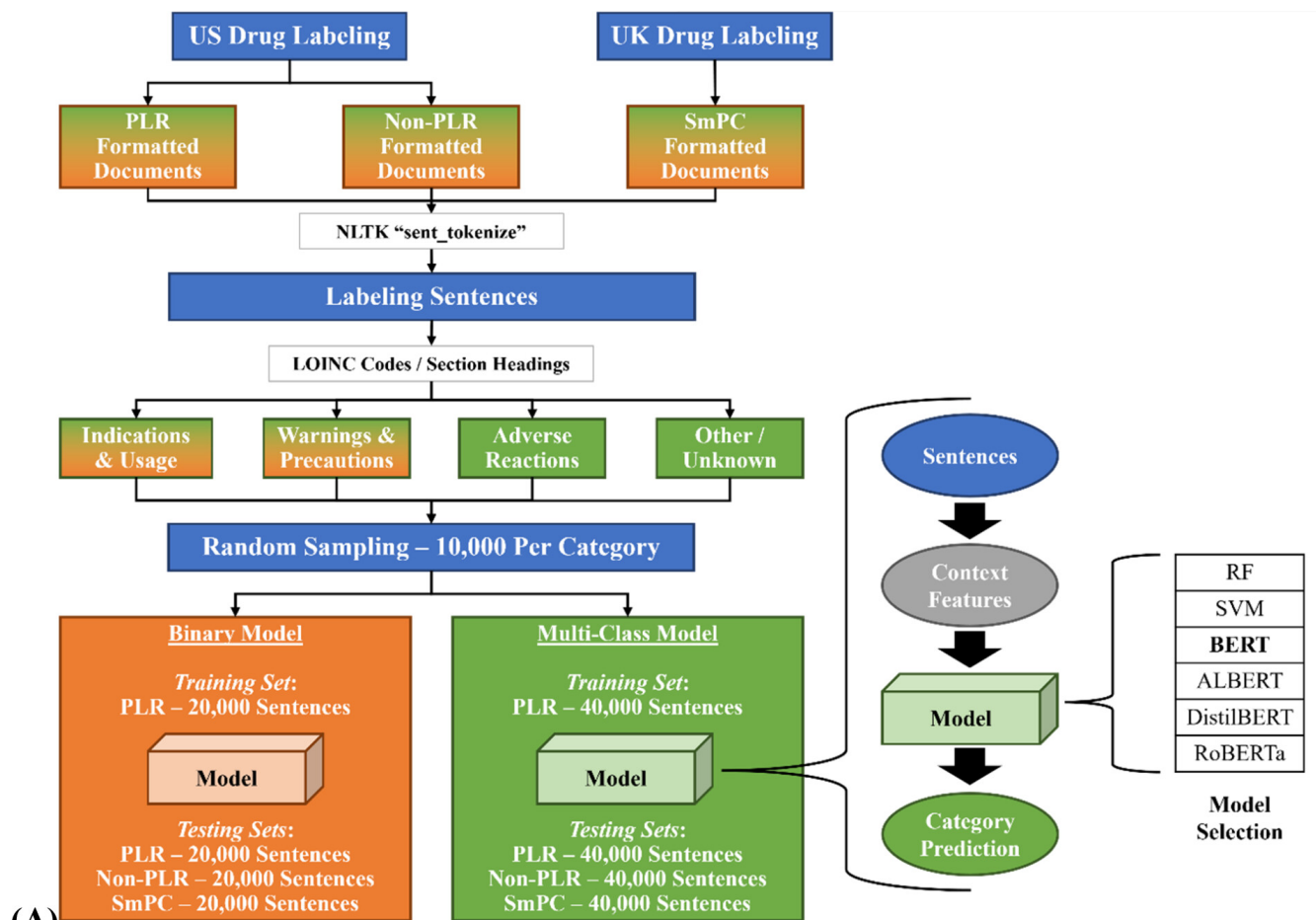
结果

模型开发流程

图2A展示了本研究的整体工作流程，图2B详细说明了基于BERT模型的建模过程，图2C则详细说明了RF和SVM模型的建模过程。标签数据从美国和英国的药品标签资源中收集，每个句子根据其标签章节进行分类。数据收集和处理完成后，开发了两种分类模型。第一种是一个二分类模型，用于区分“适应

症和用途”和“警告和注意事项”文本。另一种是多分类模型，包括四个类别：“适应症和用途”、“警告和注意事项”、“不良反应”和“其他/未知”。这两种分类模型均基于PLR格式的标签文本开发，并在PLR、非PLR和SmPC格式的标签文本上进行了测试。首先，通过标记化和编码表示法将文本转换为上下文特征。接着，六种建模算法（RF、SVM、BERT-base、ALBERT、DistilBERT和RoBERTa）被应用于模型开发。有关更多详细信息，请参见材料与方法部分。

开发流程图示



- (A) 整体工作流程
- (B) 基于BERT模型的建模流程
- (C) 基于RF和SVM模型的建模流程

模型测试结果

在训练二分类和多分类模型时，经过10个训练周期后评估准确率趋于稳定，因此我们在这一点结束训练以避免过度训练。此部分报告的结果是10个测试样本的平均准确率（每个样本包含每个终点随机选择的10,000条记录），括号内显示了这些结果的标准偏差。

表3. 二分类模型结果（预测准确率）

整体	适应症和用途	警告和注意事项
验证准确率 (PLR)	0.9635	-
测试平均值 (标准差)	PLR	0.9486 (0.0010)
非PLR	0.8756 (0.0019)	0.8564 (0.0013)
SmPC	0.8827 (0.0018)	0.8809 (0.0013)

表3显示了BERT二分类模型的结果。如预期所示，PLR测试数据集的结果最高，因为模型通过PLR格式文件进行了微调。然而，非PLR和SmPC数据集的准确率和精确度非常相似，这表明该模型在所有类型的外部测试数据集上表现良好。此外，训练验证的准确率为0.9635，而在PLR、非PLR和SmPC测试数据集上的平均测试准确率分别为0.9486、0.8756和0.8827。该模型在区分这两个类别的句子时表现出色。

表4展示了BERT多分类模型的结果。同样地，非PLR和SmPC数据集的准确率和精确度也非常相似。此外，训练验证的准确率为0.8798，而在PLR、非PLR和SmPC测试数据集上的测试准确率分别为0.8194、0.7302和0.6846。该模型能够有效区分来自这四个类别的句子。然而需要注意的是，新增的两个新类别显著降低了模型对“警告和注意事项”部分的预测精度。这可能是由于该部分与“不良反应”部分的内在相似性，或者“其他/未知”部分的多样性导致了模型预测的错误。

表4	总体	适应症和用法	警告和注意事项	不良反应	其他/未知
Val. PLR	0.8798				

表4	总体	适应症和用法	警告和注意事项	不良反应	其他/未知
测试结果 (平均值 (标准差))					
PLR	0.8194 (0.0019)	0.9040 (0.0017)	0.9044 (0.0023)	0.8166 (0.0038)	0.6525 (0.0039)
非PLR	0.7302 (0.0019)	0.8061 (0.0021)	0.5982 (0.0045)	0.7812 (0.0038)	0.7351 (0.0018)
SmPC	0.6846 (0.0012)	0.8538 (0.0015)	0.6554 (0.0036)		

比较不同的建模算法

用于文本分类任务的BERT模型有多种选择，因此我们比较了一系列模型在二分类和多分类任务中的表现，使用同一训练和测试样本。表5显示了BERT模型相比基线模型（如RF和SVM）具有更低的错误率（例如，在PLR测试的二分类任务中，RF和SVM模型的错误率为7-8%，而BERT模型的错误率为5%，即错误减少了约50%）。此外，基线模型的“黑箱”性质进一步显示了使用BERT与SHAP解释模型预测的优势。

	验证	测试				
	PLR		PLR		非PLR	
	Binary	Multiclass	Binary	Multiclass	Binary	Multiclass
RF	0.94	0.81	0.92	0.81	0.88	0.73
SVM	0.95	0.85	0.93	0.81	0.88	0.74
BERT	0.96	0.88	0.95	0.84	0.89	0.74
ALBERT	0.96	0.87	0.95	0.84	0.89	0.72
DistilBERT	0.96	0.88	0.94	0.83	0.88	0.74
RoBERTa	0.97	0.88	0.95	0.83	0.89	0.74

尽管大多数BERT模型在总体表现上相似，但不同模型在某些方面略有优势。例如，RoBERTa模型在PLR测试中的验证准确率略高，特别是在SmPC数据集上的二分类任务中表现出色。因此，如果时间不是问题，这个模型可能是更好的选择，因为它有更长的训练阶段。此外，值得注意的是，ALBERT和DistilBERT模型尽管比BERT模型小得多且速度更快，但精确度损失很小。这表明在存在时间限制时，这些模型可能是更好的选择。总的来说，这一分析提供了对所选BERT模型的优劣势的进一步洞察，有助于未来研究。

图3

tx3c00028_0003.jpeg (1000×1140) (acs.org)

图3比较了四种BERT模型在10个训练周期中的训练和评估损失。图3A比较了模型的训练损失，而图3B比较了评估损失。图中显示，模型的训练损失随着时间逐渐下降，评估损失则逐渐增加。为了防止评估损失过高，微调在10个周期后结束。

对预测影响最大的关键词

为了确定在不同药品标签章节中对预测影响最大的词语，使用微调后的BERT模型计算了每个终点的Shapley值。对于每种格式，使用了每个终点的1000个记录进行计算。

图4

tx3c00028_0004.jpeg (1952×1188) (acs.org)

图4显示了在PLR格式文档中对预测影响最大的词语。记录的平均Shapley值提供了特定词语对模型预测的相对影响的平均值。与预测正相关的值用红色显示，而负相关的值用蓝色显示。图中展示了前五个正相关和负相关值。

根据结果，词语“Indicated”（表明）和“Prevention”（预防）是“适应症和用途”部分中最具影响力的词语，而“Affected”（受影响）、“Stop”（停止）和“Consider”（考虑）是“警告和注意事项”部分的关键词。此外，“不良反应”部分主要受“tolerate”（耐受）、“alleviate”（缓解）和“occasionally”（偶尔）的影响。总体来看，这些部分中的主导词语是合理的，它们的使用与各自部分所涵盖的信息高度一致。然而，“其他/未知”类别中选定的词语似乎较为随

机，这可能是因为该类别涵盖了药品标签的多种章节。总体而言，这一分析提供了更多关于用于区分PLR格式章节的重要词语的见解。

模型权重分析

为了展示某些词语对分类的影响，我们使用SHAP库和微调的BERT模型对一些句子进行了可视化。

图5

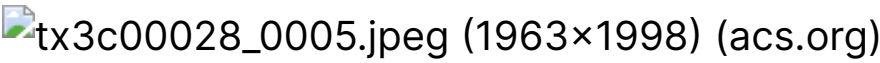
tx3c00028_0005.jpeg (1963×1998) (acs.org)

图5显示了从PLR格式文件中提取的四个句子的文本图（每个分类各一个）。蓝色高亮的词语与句子所属章节的终点负相关，红色高亮的词语则正相关。这些图中的SHAP值展示了哪些词语在句子中的影响最大，导致了它们的分类。例如，在图5A中，来自“适应症和用途”部分的句子中，词语“Indicated”和“Treatment [of]”在分类中起到了最大的作用。此外，在图5B中，来自“警告和注意事项”部分的句子中，词语“Consider”（正面影响）和“Adversely”（负面影响）对该句子的分类起到了关键作用。总体来说，此分析有助于揭示模型如何通过词语影响来进行句子的分类。

讨论

使用段落或句子作为输入

我们进一步探讨了不同输入层次如何影响结果。为此，进行了另一组实验，使用段落而非句子作为输入。表6比较了BERT二分类模型的结果。在这项任务中，段落输入模型在各项指标上均优于句子输入模型：其训练验证准确率、总体测试准确率以及各终点的预测精度均表现得更为出色。由于段落本质上为模型提供了比句子更多的上下文或信息，因此模型能够做出更准确的预测。

表6		总体	适应症和用法	警告和注意事项	
句子输入	验证	PLR			
	测试	PLR	0.95	0.94	0.96
		非PLR	0.89	0.91	0.88

表6		总体	适应症和用法	警告和注意事项	
段落输入	验证	PLR	0.98		
	测试	PLR	0.97	0.96	0.98
		非PLR	0.92	0.92	0.92

总体而言，这些发现表明，与句子输入模型相比，段落输入模型在主要的药品标签章节分类任务中表现更好，这表明这可能是未来研究中的一个有前景的方向。然而，使用段落作为输入也存在一些缺点。例如，段落输入模型的训练和测试时间比句子输入模型要长得多，这可能会对时间有限的审核员造成困扰。此外，最重要的是，对个别句子的分类可能对研究人员和审核员更为有用，因此句子级别的预测在未来项目中可能更为方便。总体来说，尽管段落输入模型的结果略好，但在任务中句子输入模型因其更为均衡且易于使用的预测而更受欢迎。

局限性与未来研究方向

在本研究中，我们只使用了有限的章节终点来训练和测试模型。具体来说，多分类模型仅分析了三个药品标签章节：“适应症和用途”、“警告和注意事项”和“不良反应”，其余章节的终点被合并为“其他/未知”类别。这些合并的章节终点也提供了重要或有价值的信息，因此，未来研究应重点开发具有不同终点配置的模型，这可能会带来新的发现。此外，未来的研究可以在模型微调中使用更多的数据点，因为研究表明更多数据可以提高模型的性能。我们注意到，从每个终点1,000条记录增加到10,000条记录时，性能有所提升。尽管如此，本文开发的模型为将非结构化的监管文件组织为结构化数据集以供高效监管使用奠定了基础。

本案例研究旨在评估基于Transformer的语言模型将自由文本分类为适当药品标签章节的能力。鉴于编码器风格的语言模型（如BERT）在情感分析和文本分类等任务中的表现优异，因此我们选择了这种架构。然而，解码器风格的语言模型（如GPT系列）也可能适用于为药品标签文本应用格式或结构，因此这些模型应在未来的研究中进行探索。尽管如此，本案例研究证明了深度

学习神经网络能够将来自不同格式的药品标签文件中的文本进行连接并归类为标准化的类别。


监管科学中的应用前景

总体来说，本研究在监管科学领域独具特色，具有广泛的应用前景。首先，基于所获得的知识开发的语言模型，自动化的监管提交结构化技术可能会应运而生，简化监管文件的提交过程。其次，这项研究可能有助于基于结构良好的文件（如PLR vs非PLR格式）生成更易于理解的、安全导向的处方药标签。未来，本研究开发的语言模型可能应用于处理其他未格式化的文件（如扫描件或照片），并将这些内容加入监管知识库中。

结论

在本研究中，为了使非结构化的文本信息更加易于监管审查员和研究人员访问，我们开发了一种能够将文本或句子分类为标准化药品标签章节的语言模型。通过使用基于BERT的模型，自动将自由文本分类为适当的药品标签章节成为可能，且取得了显著的效果。因此，这一项目为未来的监管科学工作奠定了基础。

关联内容

数据可用性声明：本项目中使用的代码和数据集已在GitHub上公开，地址为：https://github.com/magnusgray1/drug_label 

附加信息

支持信息文件可免费下载，网址为：

<https://pubs.acs.org/doi/10.1021/acs.chemrestox.3c00028> 