arXiv:2302.06474v1 [cs.CL] 7 Feb 2023

应用 BERT 和 ChatGPT 对科学文献中的莱姆病进行情感分析

Teo Susnjak[1]

[1]新西兰奥克兰梅西大学数学与计算科学学院
计算机科学与信息技术高级讲师

摘要

本章介绍了一份实用指南，说明如何运用自然语言处理（NLP）技术，对蜱传疾病文本进行情感分析。本文旨在展示如何评估围绕该疾病慢性表现的论述中是否存在偏倚。为此，本文使用从科学期刊中收集的 5643 篇有关慢性莱姆病的摘要作为数据集，通过 Python 演示如何使用预训练语言模型进行情感分析，并介绍如何运用可解释机器学习工具以及一种采用 ChatGPT 等新兴先进大语言模型的新方法，验证初步结果。本章可为有意在医学领域运用 NLP 技术进行情感分析的研究人员和从业者提供有益参考。

关键词— 情感分析、莱姆病文本分析；自然语言处理；BERT、ChatGPT、SHAP、语言模型、可解释人工智能

1 引言

自然语言处理（NLP）是一种有效的工具，可用于从文本中提取见解。自然语言处理的底层技术可用于收集

并分析文本内容，例如进行主题发现，以及识别模式和趋势。本章旨在展示如何在科学领域的蜱传疾病文本中，将自然语言处理与情感分析有效结合。本章的目标是展示确定有关该疾病慢性表现的论述中是否存在偏倚这一过程的初步步骤，而总体方法也广泛适用于其他情境。

学术界目前对慢性莱姆病是否存在的共识是，这是一个颇具争议的话题 [4, 6]。一些科学家和医疗专业人员认为，慢性莱姆病（也称为莱姆病治疗后综合征，PTLDS）确实存在，而且是一种使人衰弱的病症，可能发生在一个人感染导致莱姆病的细菌之后 [8]。他们认为，PTLDS 的症状（如疲劳、疼痛和认知障碍）在初始治疗后可能持续数月甚至数年，而这些症状是由莱姆病细菌持续感染引起的。

另一方面，另一些科学家和医疗专业人员认为，几乎没有科学证据支持慢性莱姆病的存在 [3]。他们声称，PTLDS 的症状并非由莱姆病细菌持续感染引起，而是由免疫功能障碍、慢性疲劳综合征或抑郁症等其他因素导致。他们认为，现有证据不支持对 PTLDS 进行长期或反复的抗生素治疗，而且此类治疗可能有害。

本章介绍如何使用自然语言处理技术开展情感分析，并提供实用指南，旨在展示识别医学期刊在蜱传疾病是否具有慢性这一问题上的立场所存在潜在偏倚的过程。为实现这一目标，本章使用了一个数据集，其中包含 2000 至 2021 年间从科学期刊收集的 5643 篇有关慢性莱姆病的摘要。本章旨在通过少量 Python 代码，展示如何利用预训练语言模型开展初步情感分析，并使用 SHAP 等可解释人工智能工具以及 ChatGPT 等新兴先进大语言模型对分析结果进行验证。将这些语言模型与全面的数据集结合使用，有助于深入分析医学界围绕慢性莱姆病的讨论。本章可为有意在医学领域使用自然语言处理技术开展情感分析的研究人员和从业者提供参考。

2 材料

选择自然语言处理工具时，考虑自身的技术能力，以及是否愿意投入时间学习用于此任务的新工具，会很有帮助。可供选择的工具种类繁多，且各有独特的功能和能力，因此需要仔细评估工具是否符合个人的具体需求和长期目标。

目前有一些无需编写代码、只需通过点击操作即可使用的工具。它们有时会受到限制；不过，对于刚接触自然语言处理和编程的人来说，这类工具可能是更合适的入门途径，因为通常无需掌握大量编程知识，就能更快地开始使用。

反之，对于熟悉编码的资深开发者而言，Python 或 R 等高级工具可能更合适；这些工具拥有丰富的自然语言处理库和软件包。不过，本章将说明，只需编写少量 Python 代码，也能完成大量分析；此外，结合多种技术开展混合分析，同时使用编码工具和点击式软件，也是一种实用的方法。

2.1 数据收集工具

本章假定研究人员已完成数据收集任务，因此不再介绍这一过程。不过，下面会提及几种工具，以供参考；如果读者需要执行此项任务，或对其他技术感兴趣，可以加以了解。

Scrapy 是一个开源网络爬虫框架，文档完善，具备多种功能，可简化使用 Python 构建网络爬虫的过程。ParseHub 是一款易于使用、无需编写代码的网页抓取工具，并可将提取的数据导出为多种格式。Octoparse 是一款基于云端、通过点击操作即可使用的网页数据提取解决方案，无需编程技能即可使用。

2.2 自然语言处理工具

本章使用 Python 进行演示。Anaconda 是推荐使用的 Python 编程语言发行版，包含一系列用于数据科学的库和工具，可安装在个人计算机上。不过，使用托管这项技术的云服务提供商，通常能更快地入门。注释 1 简要介绍了安装过程和其他云端解决方案。R 也是适用于自然语言处理的编程工具。下面列出了一些其他自然语言处理工具。

- KNIME 是一个开源数据分析平台，可用于自然语言处理任务。它提供了用户友好的图形界面，用户可以借此创建、执行和共享用于文本预处理、特征提取和情感分析等自然语言处理任务的工作流。KNIME 的一些优势包括灵活性、处理大型数据集的能力，以及通过拖放界面提供的多种可视化选项。不过，用户最初需要克服一定的学习门槛，而且它提供的预构建模型可能不如其他自然语言处理工具丰富。

- OpenNLP 是一个用于自然语言处理的开源 Java 程序库，包含分词、词性标注、命名实体识别等工具。OpenNLP 文档齐全，并提供种类丰富的自然语言处理工具。不过，在某些任务中，OpenNLP 的性能可能不够理想，而且可能不如其他程序库易于使用。

- MonkeyLearn 是一个云端自然语言处理平台，提供包括情感分析在内的通用文本分析工具。它无需编程，并支持使用可自定义的预构建模型。

- SAS 是一款商业软件套件，提供丰富的数据分析工具和自然语言处理功能。它提供文本挖掘、情感分析和实体识别等功能。它能够与其他 SAS 工具集成并处理大型数据集，因此非常适合自然语言处理任务。不过，使用时也必须考虑其成本和较高的学习门槛。

3 方法

本方法部分介绍了对从同行评审期刊中提取的摘要进行情感分析所需的步骤。其目标是展示如何从这些摘要所表达的态度和观点中进行洞察提取的初步工作，以便更好地了解学术界如何讨论这一感兴趣的主题。本文不会在此得出结论，而是为研究人员奠定基础，使他们能够开展后续更复杂的分析。

所述方法专为分析同行评审文献摘要的情境而设计，并考虑了此类文本的独特特征和限制。主要限制在于，学术散文通常更为客观，不太容易显露主观情感。本节概述了用于收集和分析数据的流程与技术，并附有输出示例。总体目标是提供一些指导，以便复现初步结果。

3.1 流程

1. 数据收集：在莱姆病领域开展情感分析自然语言处理工作的第一步，是收集大量文本数据集。数据来源可以包括患者和／或医师讨论论坛、患者testimonials、社交媒体平台、医学期刊或其他科学研究出版物。收集过程中，务必筛选数据，确保其内容与莱姆病相关，并根据需要聚焦于特定子主题。图 1 展示了此处所用数据集的示例，其中呈现了分析所需的最少列数以及以 CSV 格式存储的数据集结构。

	A	B	C	D
1	publication	title	year	abstract
2	journal of psychiatric practice	Panic attacks may reveal previously unsuspected chronic disseminated lyme disease.	2000	The author describes the histories of three patients with panic-like episodes that turned out to be related to underlying, previously unsuspected tick-borne diseases.
3	journal of medical entomology	Efficiency of drag sampling for estimating population sizes of Ixodes pacificus (Acari: Ixodidae) nymphs in leaf litter.	2000	Drag sampling is a commonly used method to obtain relative estimates of the density of questing nymphal Ixodes scapularis Say and I pacificus Cooley & Kohls
4	journal of medical entomology	Populations of Ixodes scapularis (Acari: Ixodidae) are modulated by drought at a Lyme disease focus in Illinois.	2000	From 1990 through 1997, Ixodes scapularis Say larvae and nymphs were sampled between May and October along a 400-m segment of a nature trail in a Lyme disease
5	journal of medical entomology	Methodology for sampling questing nymphs of Ixodes ricinus (Acari: Ixodidae), the principal vector of Lyme disease in Europe.	2000	To assess the Lyme borreliosis vector population density we set up a methodology for sampling the Ixodes ricinus L. population host questing on the vegetation. We

图 1：数据集示例，展示前五行及其关键列，包括期刊名称、论文标题、发表年份和摘要文本。

2. 数据清洗： 收集数据后，重要的是清理数据集，去除任何无关信息。如果能够高效地完成或实现自动化，纠正拼写和语法错误会有所帮助[1]；不过，对于本示例中使用这些模型进行情感分析而言，这并非绝对必要。发现重复数据时应将其删除。通过网络抓取数据时，可能会无意中收集到广告等无关数据，一旦发现就应将其删除。数据清洗通常有助于提高结果的准确性和稳健性；然而，全面清理数据所需的时间，应与这项工作在大型数据集中有时只能带来的边际收益权衡。数据集中存在一定程度的噪声是可以预料的。使用所选工具清理数据集后，即可使用 Python 将其读入变量，以进行情感分析，如图 2 所示。

[1]GPT 系列和类 BERT 的 Transformer 模型可以自动完成文本中的拼写和语法错误纠正任务。

```
# Import the pandas library used for data processing
import pandas as pd

# Read into a data frame variable the contents of the dataset
abstracts_df = pd.read_csv('abstracts.csv')
```

图 2：将清理后的数据集读入 Python 变量的示例。

3. 情感分析： 清理数据后，可以按照以下步骤进行情感分析。

I. 模型选择：有许多模型可用于执行情感分析。研究人员可以选择使用广泛可用的机器学习模型或基于词典的方法。这里演示一种名为 BERT [2] 的预训练深度学习变换器模型。BERT（来自 Transformer 的双向编码器表示）是由谷歌开发的语言模型，已广泛用于各种自然语言处理任务 [1] （注 4）。该模型可自定义，此处使用的特定模型（详情见注 2）已针对商业领域的情感分析任务进行了微调。此外，在使用该模型之前，无需执行某些额外的自然语言处理步骤（注 3）。图 3 展示了如何在 Python 脚本中调用该特定模型的代码。

```
# Import the pipeline module from the transformers library
from transformers import pipeline

# Define the name of the BERT model to be used for sentiment analysis
model = "nlptown/bert-base-multilingual-uncased-sentiment"

# Define the task as sentiment analysis
task = "sentiment-analysis"

# Defining the pipeline which represents the classification model
sentiment_pipeline = pipeline(task, model=model)
```

图 3：定义具备情感分析能力的特定 BERT 模型。

II. 文本分类：选定预训练模型并将其导入 Python 脚本后，接下来需要将模型应用于每个单独的文本数据点（摘要），并存储模型输出的得分和情感标签。该 BERT 模型会输出一个介于 0 到 1 之间的情感值，以及一个代表 1 到 5 级情感分类的标签。图 4 展示了如何

应用 BERT 模型对每篇摘要进行分类，以及如何将情感得分和标签存储到数据框变量的新列中。

```
# Iterate through each abstract in dataset and extract sentiment
for i, text in abstracts_df.iterrows():
    # Perform sentiment analysis on a given abstract
    result = sentiment_pipeline(text.abstract[:512])

    # Save the sentiment score and label
    abstracts_df.iloc[i, 4] = result[0]['score']
    abstracts_df.iloc[i, 5] = result[0]['label']

# Save the data including the extracted sentiment values
abstracts_df.to_csv('abstracts_with_sentiment.csv')
```

图4：应用 BERT 模型对每篇摘要的情感值进行分类，并提取模型输出，将其保存到单独的文件中，以便后续处理和分析。

III. 可视化：为整个文本语料库分配情感得分后，直观地分析和评估这些得分的有效性至关重要。情感得分的有效性和意义取决于情感分析模型是否适用，以及该模型所应用的具体领域。以下图示展示了此过程的一些实用可视化示例。图5展示了演示数据集中所有摘要的情感得分分布。

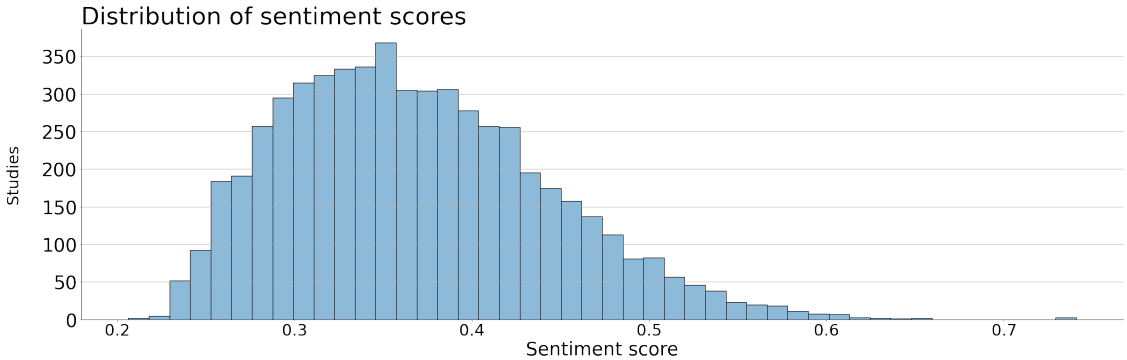


图5：数据集中所有摘要的情感得分分布。

另一项分析可以考察摘要已发表情感得分随时间的变化。如图 6 所示。

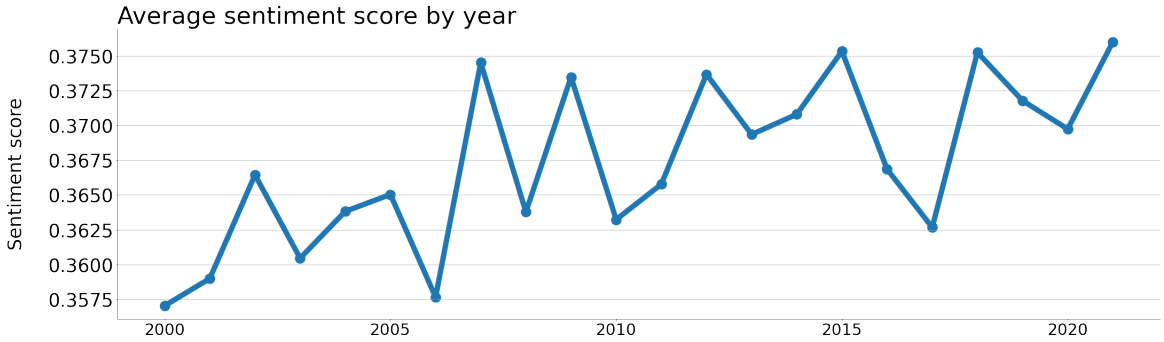


图6：2010年至2020年平均情感得分的变化趋势。

也可以按期刊对情感进行分组和汇总，以突出显示是否存在显著差异。如图 7 所示。

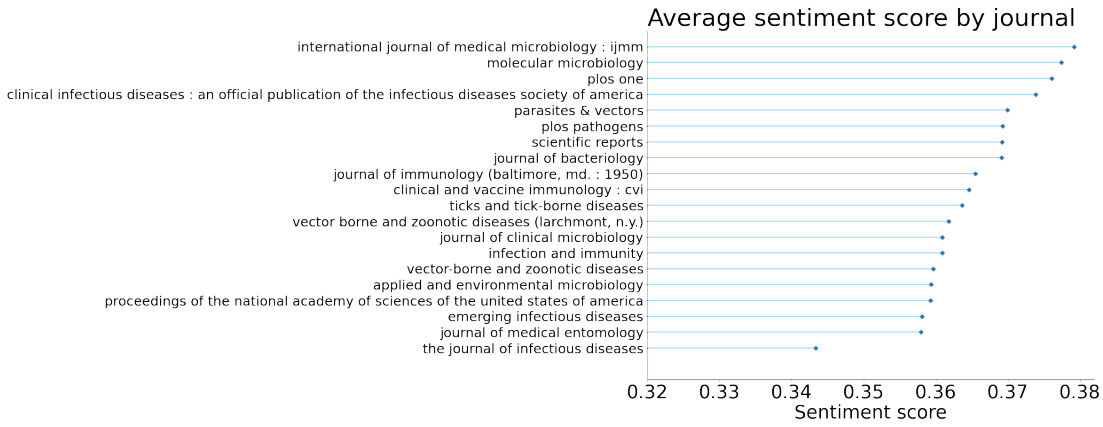


图7：数据集中按已发表摘要数量排名前20的期刊的平均情感得分分布。

IV. 文本检查与验证：研究人员在开展更高级的分析并根据情感分析结果得出结论之前，一项重要步骤是查看一组选定的待分类文本样本，并检查模型如何根据情感得分来解读不同词语。这一点至关重要，因为在某个特定领域训练的模型，面对来自其他领域的文本时，可能无法发挥最佳性能，原因在于术语和行话存在差异。在本例中，一个用于检测客户评论情感的模型被应用于医疗语境，

其中，特定领域的词语可能会对应一些意料之外的情感值。图 8 展示了一个使用 SHAP [5]工具按词检查文本分类结果的示例。图中展示了一段被分类为总体情感得分低于平均水平的文本，其中红色句子（如“导致诊断困难”）有力地印证了其较低得分，而“及时治疗可以预防残疾。”等句子则会使总体得分倾向于更高的情感类别。

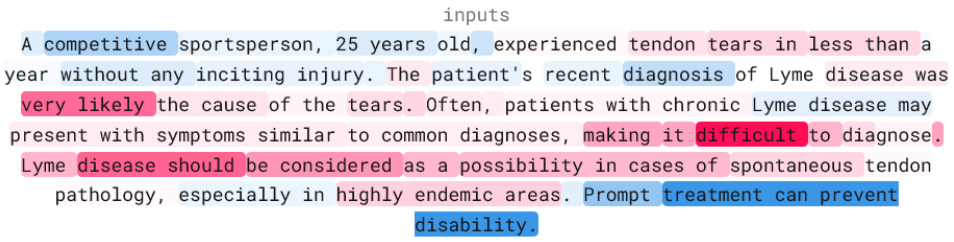


图 8：一篇探讨慢性莱姆病的论文的期刊论文摘要示例。该样本文本的总体情感得分为 0.31，情感分类在 1 到 5 的量表上为 2。图中展示了不同词语对情感最终预测的影响：红色高亮的词语会使预测更倾向于此分类，而蓝色词语则产生相反的效果；颜色越深，影响越强。

用于生成图 8 的 Python 代码见图 9。

近期开发的一种 AI 技术可用于检查和验证不同的文本样本，并借助 ChatGPT 等先进的大语言模型（LLM）（或通过应用程序接口调用 GPT-3）交叉核验对文本情感和主观性的解读。下图展示了 ChatGPT 对给定文本的解读示例，包括其主观性程度（图 10）和情感（图 11）；这些结果与 BERT 模型及先前的分析一致。

V. 替代模型与模型微调：目前大多数可用的预训练语言模型虽然能够按情感对文本进行分类，但并非以医学或健康领域为目标开发的。这种情况可能在不久的将来发生改变。注释 5 指引研究人员前往基于变换器的模型的主要资源库，以查找替代模型和新模型。

```python
# Import the shap library
import shap

# create an explainer object using the sentiment pipeline
explainer = shap.Explainer(sentiment_pipeline)

# compute the shap values for the first abstract in the dataframe
shap_values = explainer(abstracts_df.abstract.values[0])

# plot the shap values as a text plot
shap.plots.text(shap_values)
```

图 9：这段 Python 代码使用 SHAP 程序库创建一个解释器对象，用于计算输入文本（摘要）的 SHAP 值。随后，这些 SHAP 值会以文本图的形式绘制出来，展示输入文本中的每个词如何影响模型给出的整体情感预测。



图 10：要求 ChatGPT 识别主观性较强的特定词语和短语时的输出。

研究人员也可以针对医学与健康领域定制底层的基于变换器的模型，例如 BERT，以提升情感模型的性能。这是一个可选步骤，本章不作介绍。此过程需要具备一定的技术能力，并投入时间创建可用于微调底层 BERT 模型的数据集；不过，这将是一项有价值的贡献。

4 注释

1. 软件安装：可从 https://www. anaconda.com/products/distribution 免费下载 Anaconda。Deepnote（https://deepnote.com）是提供科学计算 Python 库的另一家云服务商，并拥有易于使用的编程界面。本研究通过安装图 12 所示的另外两个软件包开展。

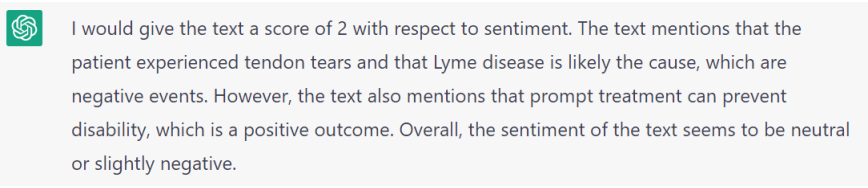


图 11：要求 ChatGPT 对摘要进行情感分类（范围为 1 到 5）时的输出。

```
!pip install shap==0.41.0
!pip install transformers==4.25.1
```

图 12：需要安装的 Python 软件包：Transformer [7]和 SHAP。

2. 本演示所用 BERT 模型的相关信息请参见：https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment。当前所用特定 BERT 模型的一项局限是，每次最多可处理 512 个词元（单词）。由于有些摘要篇幅超过此限制，为了绕过这一限制，可以将摘要拆分为较小的片段，然后计算所有片段的加权平均值，从而为整篇摘要赋予情感值。

3. 分词是一个额外步骤，通常在使用依赖标准机器学习模型或基于词典的方法的大多数情感分析模型之前执行。

4. 对莱姆病主题的科学文献进行情感分析面临两大难题。首先，科学文献力求以客观、基于事实的方式呈现研究结果，不表达任何正面或负面情感，因此模型难以识别情感上的细微差别。其次，大多数预训练模型都是基于商业语境中的文本数据集进行训练的，其中包含客户评论，这使问题更加复杂。这里的挑战在于，某些领域特定术语，例如医疗语境中的“病理学”，可能带有中性情感，而在商业语境中，这个词可能具有负面含义。预训练模型的源领域与目标领域之间的不匹配是一项重大挑战，也是当前持续研究的方向 [1]。跨领域情感分类和迁移

学习是旨在应对这一挑战的重点研究领域。

5. 此处可查看其他可选及经过微调的、基于 Transformer 的情感分析模型：https://huggingface.co/models?sort=downloads& search=sentiment。随着时间推移，专为医学与健康研究领域定制的领域特定模型将陆续推出。目前已有一些适用于临床领域的模型，但它们不具备执行情感分析的能力。

5 致谢

作者谨向 Jamie Scott 致以诚挚谢意，感谢他为本研究所用数据集的收集、数据清洗及初步分析作出的宝贵贡献。他的奉献与辛勤工作为本项目的顺利开展发挥了关键作用。此外，作者还要感谢 Stephen Croucher教授最初提出研究慢性莱姆病文献情感的想法。

参考文献

[1] M. D. Deepa 等。用于情感分析任务的双向 Transformer 编码器表示（BERT）语言模型。《土耳其计算机与数学教育期刊（*TURCOMAT*）》，12(7):1708–1721，2021。

[2] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova。Bert：用于语言理解的深度双向 Transformer 预训练。载于*2018* 年北美计算语言学协会年会论文集：人类语言技术，第 1 卷，第 4171–4186 页，2018 年。

[3] J. J. Halperin。《慢性莱姆病：患者管理中的误解与挑战》。感染与药物耐药性，第119–128页，2015年。

[4] M. Levesque 和 M. Klohn。多线索方法：理解加拿大海洋省份莱姆病管理所面临的问题与挑战。《国际环境研究与公共卫生杂志》，16(9):1531，2019 年。

[5] S. M. Lundberg 和 S.-I. Lee。解释模型预测的统一方法。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展第 *30* 卷，第 4765–4774 页。Curran Associates, Inc.，2017 年。

[6] A. W. Rebman 和 J. N. Aucott。莱姆病治疗后综合征作为

莱姆病持续症状的模型。《医学前沿》，第 57 页，2020年。

[7]T. Wolf、L. Debut、V. Sanh、J. Chaumond、C. Delangue、A. Moi、P. Cistac、T. Rault、R. Louf、M. Funtowicz、J. Davison、S. Shleifer、P. von Platen、C. Ma、Y. Jer- nite、J. Plu、C. Xu、T. L. Scao、S. Gugger、M. Drame、Q. Lhoest 和 A. M. Rush。Transformer：自然语言处理领域的先进技术。载于论文集

*2020*年自然语言处理实证方法会议：系统演示，第38–45页，线上，2020年10月。计算语言学协会。网址
h
t
t
p
s://
www.aclweb.org/anthology/2020.emnlp-demos.6。

[8] K. H. Wong、E. D. Shapiro 和 G.K. Soffer。《面向临床免疫学家的莱姆病治疗后综合征和慢性莱姆病综述》。临床过敏与免疫学评论，62(1):264–271，2022。