arXiv:2301.00891v1 [cs.CL] 2 Jan 2023

使用语言模型理解政治极化：一个数据集与方法

Samiran Gode, [1] **Supreeth Bare,** [1] **Bhiksha Raj,** [1] **Hyungon Yoo,** [1]

卡内基梅隆大学

摘要

本文旨在利用语言模型分析美国政治体系中的政治极化，从而帮助候选人做出明智的决定。获取这些信息将有助于选民了解候选人在经济、医疗保健、教育及其他社会议题上的观点。我们的主要贡献包括一个从维基百科提取、涵盖过去 120 年的数据集，以及一种基于语言模型的方法，可帮助分析候选人的政治极化程度。我们的数据分为两部分：候选人的背景信息和政治信息，因为我们的假设是，候选人的政治观点应基于理性，而不受出生地、母校等因素影响。我们还按时间顺序将这些数据划分为 4 个阶段，以帮助了解候选人之间的政治极化是否发生变化，以及如何变化。我们对数据进行了清理，以去除偏见。为了解政治极化情况，我们首先展示 Word2Vec 和 Doc2Vec 等经典语言模型的结果，随后采用 Longformer 等更强大的技术。Longformer 是一种基于Transformer的编码器，可整合更多信息，并根据候选人的政治观点和背景，找出与每位候选人最接近的邻居。

引言

美国两大政党——共和党和民主党——之间的政治极化现象长期以来一直受到研究（（Poole and Rosenthal 1984）、（KhudaBukhsh 等人 2021））。网上许多讨论都已变得两极化（Jiang、Robertson 和 Wilson 2020），而这类讨论在网上最能引发关注（Jiang、Robertson 和 Wilson 2020）。这种政治极化可能会影响候选人当选后的决策能力（Chen、Li 和 Liu 2022）。在这类情况下，用户能够区分政治言辞并了解候选人的政治极化程度至关重要。在这项研究中，我们着手探讨以下问题：“我们能否衡量一位候选人的政治极化程度？”“我们能否衡量这种极化程度随时间发生了多大变化？”我们尝试使用基于自然语言的技术来回答这些问题，并在此过程中创建一个数据集，供研究群体用于理解美国的政治极化。尽管我们的研究聚焦于美国政治体制，但我们提出的政治极化测量方法也适用于举行类似民主选举的其他国家，可用于判断候选人的政治极化程度。我们首先尝试 Word2Vec（Mikolov 等，2013）和 Doc2Vec（Le 和 Mikolov，2014）等经典方法，探究能否利用现有数据发现政治极化并获得更多见解。我们发现，具有政治敏感性的词语（Center 2019）与其他具有政治敏感性的词语相关（Center 2019）。因此，我们转而采用基于 Transformer 构建的更新、更复杂的模型（Vaswani 等人 2017），以进一步洞察数据。随后，我们使用这些模型，尤其是 Longformer（Beltagy、Peters 和 Cohan 2020），将候选人专属数据投射到特定的嵌入空间，再利用这些数据找出每位候选人的最近邻，并提出一种衡量候选人政治极化程度的指标。

版权所有 © 2023，人工智能促进协会（www.aaai.org）。保留所有权利。

相关工作

（KhudaBukhsh 等（2021））探讨了网络上的政治极化，并使用机器翻译来解读互联网中的政治极化。（Bhatt 等（2018））讨论了极端党派网站如何影响公众舆论，并以它们能够影响 2016年美国大选中的某些事件为例。随后，作者展示了研究中所假定的某些政治偏见，即公开支持民主党或共和党，被视为网站属于自由派或保守派的指标。这篇论文对我们的研究至关重要，因为它探讨了政治分歧，并为后续使用特定特征将某个实体归类为自由派或保守派的研究奠定了基础。他们考察的特征是这些网站发布或分享内容的文本记录。对我们而言，特征则只是相关人物的维基百科页面内容。（KhudaBukhsh 等（2022））展示了电视媒体和边缘新闻网络中的政治极化，并使用基于语言模型的技术进一步理解这些现象。然而，选民中显现出的这种政治极化源自候选人。（DeSilver（2022））指出，候选人变得越来越两极化，并且多年来逐渐远离政治中间立场。通过本文，我们发布一个数据集和若干指标，以帮助我们了解候选人中是否存在政治极化，以及如何衡量这种政治极化。本研究旨在

帮助选民在选举前做出明智的决定。我们对一个划分为 4 个时代、并分为 3 个部分（主要是背景、政治和其他）的数据集采用基于语言的技术。

（Belcastro 等人，2020）证明，可以借助神经网络对政治极化进行映射。这几乎是一个基准思路，因为我们也使用注意力网络和 Longformer 模型来完成同样的任务。主要区别在于数据提取和方法论。

（Khadilkar、KhudaBukhsh 和 Mitchell，2022）深入探究了大量宝莱坞（和好莱坞）电影中的性别偏见和种族偏见。作者整合了多种知名的自然语言处理模型，并尝试创建一个自身相当稳健的模型。这项研究与以往研究的不同之处在于其样本量相当大。此外，该研究还采用了一种颇具创新性的方法，即历时词嵌入关联测试（词嵌入关联测试）。其他采用的技术包括基于计数的统计，这些统计依赖于一种广受欢迎的词典填空测试，并以 BERT 为基础模型（这是我们可以在数据注意力之后考虑的一种思路）；以及使用词嵌入关联测试进行偏见识别。最终模型结合了上述三种方法。这篇论文与我们的项目高度相关，因为它采用了与我们相似的思路。它使用上述模型来预测偏见，即进行情感预测。在我们的项目中，我们利用数据预测政治情感，并尝试将某些特征分类为分类任务的前兆。

Rajani 等人，2019 尝试提升基于语音的模型将训练期间学到的推理过程用语言表达出来的能力。该方法在常识解释数据集上使用 CAGE 框架（常识自动生成解释），将效果提升了 10%。与这些较新的模型中使用 BiDAF++（加入自注意力层）相比，该方法有所改进。它还在第二阶段主要将自然语言解释用作理由泛化，服务于情感分析。本文使用 Mturk（来自亚马逊）为数据集生成解释。CAGE 主要采用问答形式，包含 3 个选项、一个标签，以及对该标签的最佳解释。此外，研究还测试了其他影响性能的评估参数；这些参数可在我们的项目中用于验证模型或作其他用途。鉴于 CAGE 达到的较高准确率，它无疑是一个有趣的验证方案。不过，需要考虑的一个因素是，数据集类型和模型类型存在很大差异。因此，我们会对上述框架作出一些修改。

（Devlin 等人，2018）是介绍 BERT 的论文。BERT 是一种将被广泛使用的模型。该论文还展示了微调 BERT 的结果。这些结果将直接或间接地用作预训练约束或调优方法。petroni2019language

（Petroni 等人，2019 年）展示了预训练的高容量模型（如 BERT 和 ELMo）可用作知识库的能力。这主要基于以下三点观察：（1）这些模型的关系知识具有竞争力，可与能够访问特定预言机知识的自然语言处理系统相媲美；（2）BERT 在开放域问答测试中的有效性；（3）某些事实很容易学会。作者还在研究中展示了其他模型（单向和双向）的用法，即“fariseq-fconv”和“Transformer-XL”。最后，他们指出，BERT-Large 的表现优于其他模型，甚至在同一任务上能够与监督模型一较高下。

（Palakodety、KhudaBukhsh 和 Carbonell，2020 年）展示了 BERT 和类似语言模型追踪社群认知、汇总观点以及比较政党和候选人受欢迎程度的能力。这与我们的研究密切相关，因为我们打算使用 BERT 进行情感分析。作者最后指出，未来可以将语言模型用作提取数据的处理流程。

在（Hamilton、Leskovec 和 Jurafsky，2016 年）中，作者试图解决词义随语境发生语义变化的问题。他们提出了一种利用嵌入的稳健方法，并依据“从众定律”和“创新定律”对其进行评估。这些定律揭示了词频和多义性在构成语言结构基石过程中的作用。这些基石之所以至关重要，有两个原因：（1）词义变化可能对情感分析产生不利影响，进而影响结果。因此，必须适当控制词频和多义性。（2）嵌入研究是基础性的，因为我们使用的是基于嵌入的模型，尤其是 Word2Vec。

数据集描述

来源

我们的数据来自 58th 届至117th 届美国国会期间政治人物（参议员和国会议员）的个人页面。我们按时间顺序将这些数据分为 4 个阶段，每个阶段约涵盖 14 届国会。对于每位国会议员，我们按栏目抓取数据。

数据收集与处理

我们根据每届美国国会维基百科页面上的政治人物名单抓取维基百科内容。对于名单中的每位国会议员，我们都会存储其标签、政党和元数据。对于每个实例，数据包括其个人详情，以及其维基百科页面上的所有信息，并以字典形式保存，其中标题为键，内容为值。这些信息有助于下游任务中的数据清理。我们根据实验对这些数据进行标注；在本研究中，我们手动标注数据，将这些键分为三个不同类别：1）背景数据，2）政治数据，3）其他类别。在发布的数据中，我们将同时提供标注版和原始版，以便用户按需使用。维基百科页面的章节格式并不固定，每位政治人物都有不同的主要章节。例如，“早年生活与背景”可能细分为“教育经历”“职业经历”“家庭”“个人经历”等多个章节。因此，我们将这些章节统一归入“背景”类别。类似地，凡是与其政治 affiliation、选举、竞选活动以及任职期间担任的职务有关的内容，均归入一个标注类别“政治生涯”。其他类别，如奖项、争议、商业相关活动、从政后的职业生涯，则

归入“其他类别”。这样，我们便可以根据各类别中的内容，手动调整标注，从而在每个类别下仅选取相关数据。总之，仅第4阶段中，在第一轮处理里，我们将约26年（1995—2021年）间1631个实例所涉及的1656个类别合并为3个类别。这些数据仍包含人名、组织、地点、数字等信息，需要进一步清理。我们首先在数据上运行命名实体识别模型，以移除人名和组织名称。不过，我们只从政治部分移除地名。这样做是为了确保政治部分的信息不受地点信息影响。至于背景信息，我们希望了解一个人出生和成长的地点是否会影响其政治观点，因此仅保留了这类地点信息，其他地点信息均予以删除。经过命名实体识别处理后，我们再对这些信息进行处理，移除数字及其他无关的正则表达式。这样可以确保传递给其他下游任务的数据干净，并能提供无偏的答案。

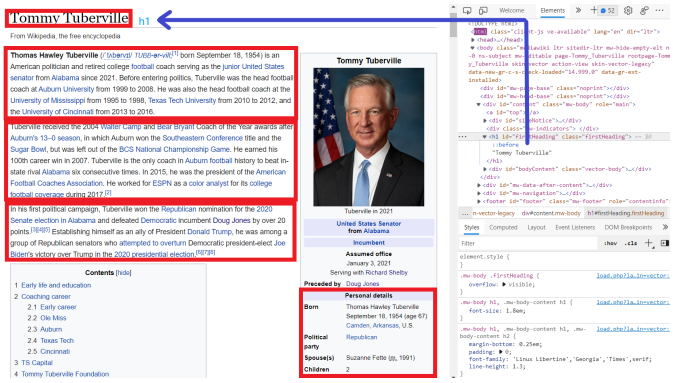

图 1：基于各标签的网页抓取

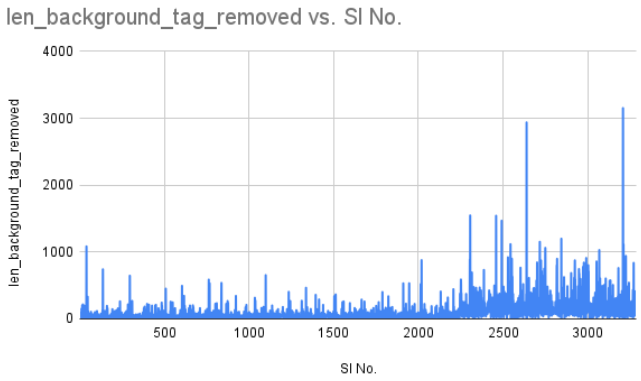

图 2：背景

语言模型

基于自然语言处理的应用一直由基于 Transformer 的语言模型主导，其中 BERT（Devlin 等人，2018）和 RoBERTa（Liu

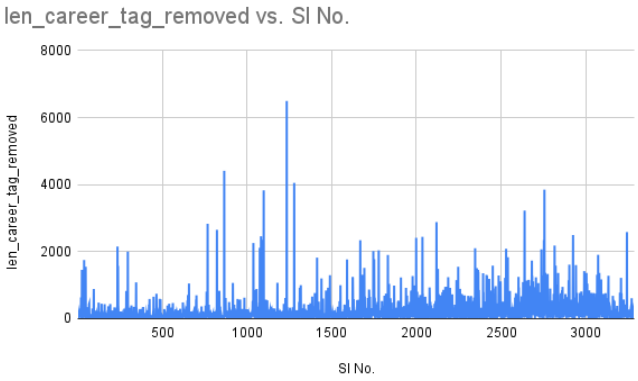

图 3：政治

等，2019）自 2018 年以来一直处于当前最先进水平，但对于我们的数据集而言，这些模型存在一个缺点：处理较长序列的能力有限，因为注意力机制的计算成本按 $O(N^2)$ 的量级增长。Longformer（Beltagy、Peters 和 Cohan，2020）及其他变体适用于这项任务，它们可接受 4096 个输入词元，而 BERT 只能接受 512 个。它通过重新设计自注意力的计算方式降低了模型复杂度。下表展示了 Longformer 在原始数据上的表现与当前最先进水平的对比。

实验

初步实验

我们的初步实验旨在深入了解数据中可能存在的模式或趋势，同时探究是否存在政治极化。我们使用 Doc2Vec（Le 和 Mikolov，2014）和 Word2Vec（Mikolov 等，2013）模型开展这些初步实验。我们使用原始数据从头构建了 Doc2Vec 模型，并将每个维基百科页面视为一篇文档。首先，我们将 Doc2Vec 模型与 K 均值聚类结合使用，在政治数据和背景数据上的分类准确率分别为 59.52% 和 61.846%。随后，我们将同一 Doc2Vec 模型与二元支持向量机分类器结合使用，在政治数据和背景数据上的准确率分别达到 72.872% 和 63.564%。这些结果汇总在下表中。Word2Vec 测试既使用了预训练模型，也使用了我们从头构建并基于所收集数据训练的模型。我们采用 Word2Vec 方法，分别针对民主党和共和党阵营中的某些词语，查找近似最近邻和精确最近邻。这种最近邻方法带来了一些有趣的发现。基于存在政治极化这一假设，我们原本预期，共和党数据与民主党数据的最近邻搜索结果会存在差异。然而，使用简单的 Word2Vec 模型时，得到的 15 个最近邻相当相似，不过某些词语的邻居排序会因政党而异。例如，对于“GUN”一词，在民主党数据中，“VIOLENCE”是第 2nd 个最近邻（使用 Spotify 的 Annoy 算法查找近似最近邻）

而在共和党案例中，同一个词是第 9th 个最近邻。同样，“CHECKS”一词在民主党人数据中是第 3rd 个最近邻，而在共和党人数据中则是第 8th 个最近邻。还有更多这样的有趣例子；结合 Doc2Vec 分类结果，这些结果证明政治极化确实存在，并且可以通过基于自然语言处理的技术进行学习。

主要分析

作为初步分析的一部分，我们使用 RoBERTa，并注意到移除“民主党”“共和党人”等词语会导致分类性能下降。这是意料之中的，因为我们失去了明显的信息，而且仅根据前 512 个标记进行分类颇具挑战。因此，我们使用 Longformer，因为它一次可以处理 4096 个标记。正如预期的那样，得分显著提升，如表中所示。
Longformer 有两个版本：“longformer-base-4096”和“longformer-large-4096”。Longformer base 相较于先前的模型 RoBERTa 以及 Doc2Vec 等更简单的模型有显著改进。不过，Longformer large 得分更高，在判断某位候选人的政党时表现最佳。我们还使用 BigBird 进行分析；它是另一种能够处理长度为 4096 个标记的模型，这些实验的结果列于表中。我们利用这些信息来了解不同词语如何影响不同得分，并将其与政治极化这一更广泛的研究目标联系起来。为此，我们计算最后一层的全局注意力分数，然后找出与 $< s >$ 标记进行自注意力计算时注意力分数最高的词语。这得出了一些有趣的结果。例如，对于共和党人 Ted Stevens，“公共”“联邦”“立法机构”“威斯康星州”等一些显而易见的词语排名较高，这符合预期，因为政治文本中的主要信息与他们的工作有关；但“堕胎”一词也出现在排名前 10% 的词语中。我们正在开展更多此类分析，相信会得到更有趣的结果。对于 BERT 等较简单模型，上述分析的效果不那么显著，因为其信息是局部的；而对 Longformer 来说，这并非易事，因为 Longformer 使用滑动窗口查看上下文。不过，Longformer架构允许某些标记具有全局注意力，而选择CLS标记后，我们便可以查看所有 4096 个标记对该词的注意力。我们还在检验另一种假设：候选人的背景也有助于识别其政治倾向。如果世界不存在政治极化，这种情况本不应发生，只有政治信息才能帮助我们进行分类；然而，正如表中所示，背景同样有显著影响。

结果

我们使用带标注的原始数据集测试了不同模型，以了解文本中的政治极化情况。我们使用政治生涯数据集和



图4：从左到右，展示民主党语料库中「移民」一词随时间变化的最近邻词。可以看到，20世纪初期，「美国人」等词与其关系密切。

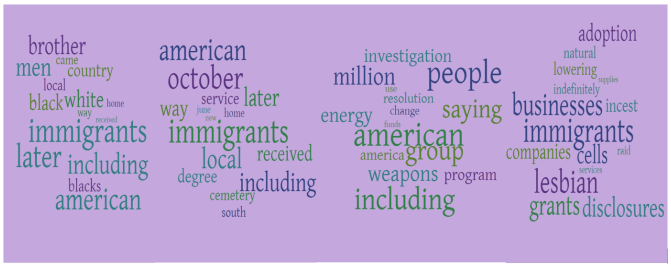

图5：共和党语料库中「移民」一词的最近邻词。在第一个时期，即20世纪初期，这些词与民主党语料库中的相似；但在最近的时期，它与其他政治敏感词相关。

背景数据集对这三种模型进行了测试，以深入了解影响政治极化的因素。所得结果如下表所示。

除了这些准确性测试外，我们还利用 Longformer 模型的注意力机制。我们找出注意力得分最高的词，并将其与我们的政治极化理论联系起来。我们还设计了一个交互式网站，帮助你了解是否存在政治极化。该网站从 Longformer 的输出中找出所选政治人物的最近邻。随后，根据最近邻中共和党人与民主党人的比例，我们估算这位政治人物的政治极化程度。下面展示了一个例子——

在图表中，x 轴表示根据数据集选出的政治人物的 20 个最近邻的排名，y 轴则显示各自的接近度分数。蓝色代表民主党人，红色代表共和党人。图中的比率是蓝点与红点的比例，因此我们的一个假设是，如果某位政治人物没有政治极化，那么仅查看背景数据时，这一比率应为 0.5（民主党人/总人数）。

模型	Data	准确率
Doc2Vec	政治	59.520
Doc2Vec	背景	61.846
Allenai/longformer-large-4096	政治	52.128
Allenai/longformer-large-4096	背景	56.383

表1：K均值分类结果

模型	Data	准确率
Doc2Vec	政治	72.872
Doc2Vec	背景	63.564
Allenai/longformer-large-4096	政治	76.862
Allenai/longformer-large-4096	背景	69.681

表 2：二元支持向量机分类结果

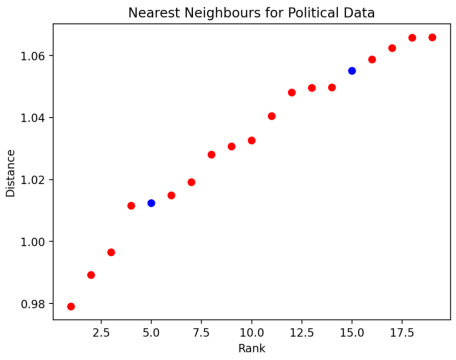

图 6：属于米奇·麦康奈尔（共和党人）的政治数据的最近邻

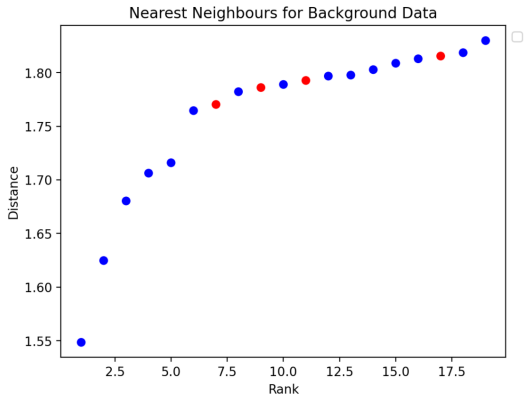

图7：属于 Ayenna Presley（民主党人）的背景数据的最近邻

上方的图10展示了米奇·麦康奈尔（共和党人）的邻居，可以清楚地看出，大多数邻居都是共和党人（红色），而民主党人只有 2 位（共 20 位）。因此可以推断，米奇·麦康奈尔的极化比率为 0.9。同样，在图11中，我们可以看到民主党人 Ayenna Presley 有 16 位邻居属于同一政党，因此背景数据的比率为 0.8。扩展这类网站的规模，并加入更多能够将成员的政治观点标示为激进或温和的指标，将有助于选民。

此外，我们展示了这组数据的价值，并希望它能帮助研究界考察候选人中的政治极化现象，以及这种现象与候选人的其他属性之间的关联。此外，我们还希望借此——了解候选人的观点，并衡量其观点具有多大的极化性。我们还希望，这组跨越多个年代的数据有助于我们理解政治理念如何随时间发生变化。

未来工作

未来，我们计划继续使用维基百科数据集，采用其他指标来衡量个人和社群的政治极化。具体而言，我们希望使用上文提到的注意力词元，考察来自候选人的给定文本中，背景词元与政治词元的比例；该文本在背景和政治两方面的分布均匀。

致谢

我们谨向 Yash Jain 和 Viraj Ranade 致谢，感谢他们作出的贡献。

参考文献

Belcastro, L.; Cantini, R.; Marozzo, F.; Talia, D.; 和 Trun- fio, P. 2020. 使用神经网络学习社交媒体上的政治极化。*IEEE Access*, 8: 47177–47187. Beltagy, I.; Peters, M. E.; 和 Cohan, A. 2020. Long-former：长文档变换器。*arXiv*预印本 *arXiv:2004.05150*。Bhatt, S.; Joglekar, S.; Bano, S.; 和 Sastry, N. 2018. 揭示党派网站生态系统。载于万维网大会*2018*配套论文集, 545–554. Center, P. R. 2019. 在政治极化时代，党派联盟之间存在明显分歧。在政治极化时代，党派联盟之间存在明显分歧。Chen, Z.; Li, Z.; 和 Liu, S. 2022. 政治极化的代价：新冠疫情期间市政债券发行方的证据。*FinanceResearch Letters*, 102781. DeSilver, D. 2022. 如今美国国会的极化，其根源可追溯至数十年前。Devlin, J.; Chang, M.-W.; Lee, K.; 和 Toutanova, K. 2018. BERT：用于语言理解的深度双向变换器预训练。*arXiv*预印本 *arXiv:1810.04805*。Hamilton, W. L.; Leskovec, J.; 和 Jurafsky, D. 2016. 历时词嵌入揭示语义变化的统计规律。*arXiv*预印本 *arXiv:1605.09096*。Jiang, S.; Robertson, R. E.; 和 Wilson, C. 2020. 探讨内容审核中的政治偏见。载于*AAAI*人工智能大会论文集, 第34卷, 13669–13672. Khadilkar, K.; KhudaBukhsh, A. R.; 和 Mitchell, T. M. 2022. 性别偏见、社会偏见与表征：BHollywood的70年。*Patterns*, 3(2): 100409. KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; 和 Mitchell, T. 2021. 我们说的不是同一种语言：通过机器翻译解读政治极化。载于*AAAI*人工智能大会论文集, 第35卷, 14893–14901.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; 和 Mitchell, T. M. 2022. 边缘新闻网络：2020年总统选举后美国新闻收视情况的动态变化。载于第*14*届*ACM*网络科学会议（*2022*）, 269–278. Le, Q.; 和 Mikolov, T. 2014. 句子和文档的分布式表示。载于国际机器学习会议, 1188–1196. 机器学习研究会论文集。Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; 和 Stoyanov, V. 2019. RoBERTa：一种稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*。Mikolov, T.; Chen, K.; Corrado, G.; 和 Dean, J. 2013. 向量空间中词表示的高效估计。*arXiv*预印本 *arXiv:1301.3781*。Palakodety, S.; KhudaBukhsh, A. R.; 和 Carbonell, J. G. 2020. 使用微调语言模型从大规模语料库中挖掘洞见。载于*ECAI 2020*, 1890–1897. IOS出版社。Petroni, F.; Rockt¨aschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; 和 Riedel, S. 2019. 语言模型能否充当知识库？*arXiv*预印本 *arXiv:1909.01066*。Poole, K. T.; 和 Rosenthal, H. 1984. 美国政治的极化。《政治学杂志》, 46(4): 1061–1079. Rajani, N. F.; McCann, B.; Xiong, C.; 和 Socher, R. 2019. 解释一下！利用语言模型进行常识推理。*arXiv* 预印本 *arXiv:1906.02361*。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; 和 Polosukhin, I. 2017. 注意力机制就是你所需要的一切。神经信息处理系统进展, 30.