arXiv:2304.01680v1 [cs.CL] 4 Apr 2023

BERT 能吃 RuCoLA 吗？用拓扑数据分析进行解释

Irina Proskurina[1], IrinaPiontkovskaya[2], Ekaterina Artemova[3][1]里昂大学,里昂 2,ERIC UR 3083,法国[2]华为诺亚方舟实验室[3]信息与语言处理中心(CIS), 慕尼黑路德维希-马克西米利安大学, 德国**通信作者：**
Irina.Proskurina@univ-lyon2.fr

摘要

本文探讨了针对可接受性分类进行微调的 Transformer 语言模型如何捕捉语言特征。我们采用自然语言处理中拓扑数据分析（TDA）的最佳实践：从注意力矩阵构建有向注意力图，从中提取拓扑特征，并将其输入线性分类器。我们提出了两个新特征：弦性和匹配数，并表明基于 TDA 的分类器优于微调基线。我们在两个数据集 COLA 和 RUCOLA 上进行了实验，[1] 涉及英语和俄语这两种类型学上不同的语言。

此外，我们提出了若干黑箱内省技术，旨在检测微调过程中语言模型注意力模式的变化，定义语言模型的预测置信度，并将各个注意力头与细粒度的语法现象相关联。

我们的研究结果有助于理解单语语言模型在可接受性分类任务中的行为，深入揭示注意力头的功能作用，并突显基于拓扑数据分析的方法在分析语言模型方面的优势。我们公开发布了代码和实验结果，以促进相关研究的进一步开展。[2]

1 引言

基于 Transformer 的语言建模（Vaswani 等人，2017年）已成为进行可接受性判断的标准方法，其结果可与人类基线相媲美（Warstadt 等人，2019年）。事实证明，预训练编码器，尤其是 BERT，在判断含有长距离依存关系的句子的可接受性时，优于其他模型（Warstadt 和 Bowman，2019年）。关于 BERT 语言模型语言学知识的研究（LM）模型的研究揭示：（1）单个注意力头可以存储句法、语义，或这两类语言信息（Jo 和 Myaeng，2020年；Clark 等人，2019年）；（2）垂直、对角和块状注意力模式可能会在不同层中频繁重复（Kovaleva 等人，2019年）；（3）微调会影响编码语言特征的方式，并往往导致部分预训练模型知识丢失（Miaschi 等人，2020年）。然而，对于英语以外语言中的语言模型语法知识，关注仍然较少。现有的跨语言探测研究表明，Transformer 语言模型的语法知识会适应下游语言；对于俄语，研究结果的解释并不容易（Ravishankar 等人，2019年）。然而，语言模型处理俄语等自由语序语言的文本时，对细粒度扰动不太敏感（Taktasheva 等人，2021年）。

本文探究了经过微调、用于俄语可接受性分类的 Transformer语言模型所捕获的语言特征。遵循可接受性分类领域的最新进展，我们使用俄语语言可接受性语料库（RUCOLA）（Mikhailov 等人,2022年），该语料库涵盖时态和词序违规、从句构建错误、不定代词使用错误以及其他相关语法现象。我们提供了一个来自 RUCOLA 的不可接受句子示例，该句子在代词使用方面存在形态学违规：使用物主反身代词“svoj”（自身的／自己的）代替第三人称代词。

（1）* 这是他职业生涯中**首次**参加世界锦标赛。
（“这是**他自己的**职业生涯中的首次世界锦标赛。”）

遵循近期提出的基于拓扑数据分析（TDA）的方法来研究语言

[1] <escape>[1]</escape>英语中的芝麻菜（rocket salad）<escape>[2]</escape>https://github.com/upunaprosk/la-tda

可接受性（LA）任务（Cherniavskii 等人，2022年），我们根据注意力矩阵构建有向注意力图，并将图的特征视为模型习得的语言特征。我们将可接受性分类任务的现有研究扩展至俄语，并展示基于拓扑数据分析的方法在该任务中的优势。我们的主要贡献如下：*(i)* 我们采用基于拓扑数据分析的方法，研究语言模型在俄语和英语可接受性分类任务中的单语行为；*(ii)* 我们引入新的拓扑特征，并取得优于此前基线的结果；*(iii)* 我们提出一种新的基于拓扑数据分析的方法，用于衡量大配置和基础配置下预训练语言模型与微调后的语言模型之间的距离。*(iv)* 我们确定注意力头在俄语和英语 LA 任务中的作用。

我们的初步假设是，不同语言的注意力图结构存在差异，尤其是在含有形态、句法和语义违规的句子中。我们通过比较注意力图的特征来分析模型之间的关系。据我们所知，我们的研究是最早尝试使用基于拓扑数据分析的方法，分析在俄语和英语可接受性分类语料库上微调的单语语言模型差异的研究之一。

2 相关工作

可接受性分类。早期研究采用统计机器学习方法、基于规则的系统和上下文无关文法进行可接受性分类（Cherry 和 Quirk，2008；Wagner 等人，2009年；Post，2011年）。其他方法采用阈值评分函数来估计句子的可能性（Lau 等人，2020年）。近期研究主要关注无处不在的 Transformer 语言模型判断可接受性的能力（Wang 等人，2018年）、探究其语法习得情况（Zhang 等人，2021年），以及评估语言生成中的语义正确性（Batra 等人，2021年）。在本项目中，我们开发可接受性分类方法，并将其应用于英语和俄语这两种不同语言的数据集。

自然语言处理中的拓扑数据分析（TDA）。近期研究利用 TDA 探索语言模型的内部运作机制。Kushnareva 等人(2021年) 从注意力图中提取 TDA 特征，用于构建人工文本检测方法。Colombo 等人(2021年) 提出了 BaryScore，这是一种自动文本生成评估指标，基于 Wasserstein 距离和重心。Chauhan 和 Kaul(2022年) 开发了一种评分函数，用于捕捉高维隐藏表示的同调，并以预测测试准确率为目标。我们扩展了 Cherniavskii 等人(2022年) 为可接受性分类提出的持久特征集合，并对这些持久特征如何影响分类器性能进行了广泛分析。

语言模型如何通过微调发生变化？ 关于微调如何影响语言模型的内部运作，现有研究主要分为两类：（i）子词表示捕捉了什么；（ii）注意力头发挥什么功能作用？实验技术包括：分析源模型与微调后的检查点权重之间的相似性（Clark 等人，2019年）、训练探测分类器（Durrani 等人，2021年）、计算特征重要性分数（Atanasova 等人，2020年）、对子词表示进行降维（Alammar，2021年）。研究发现有助于通过修改损失函数改进微调流程（Elazar 等人，2021年），并为解释语言模型的预测提供技术（Danilevsky 等人，2020年）。我们的方法通过将特定中心语的持久特征与细粒度语言现象相关联，揭示注意力头的语言能力。

3 方法论

我们遵循 Warstadt 等人，2019年 并将 LA任务视为一个监督分类问题。我们对 Transformer语言模型进行微调，以近似将输入句子映射到目标类别的函数：可接受或不可接受。

3.1 提取的特征

给定输入文本，我们从 Transformer语言模型中提取输出注意力矩阵，并遵循Kushnareva 等人，2021年 据此计算三类持久特征。

拓扑特征是注意力图的性质。我们给出了一个基于图 1中注意力矩阵构建的注意力图示例。注意力图的邻接矩阵$A = (a_{ij})_{n\times n}$ 由注意力矩阵 $W = (w_{ij})_{n\times n}$使用预定义的阈值获得。

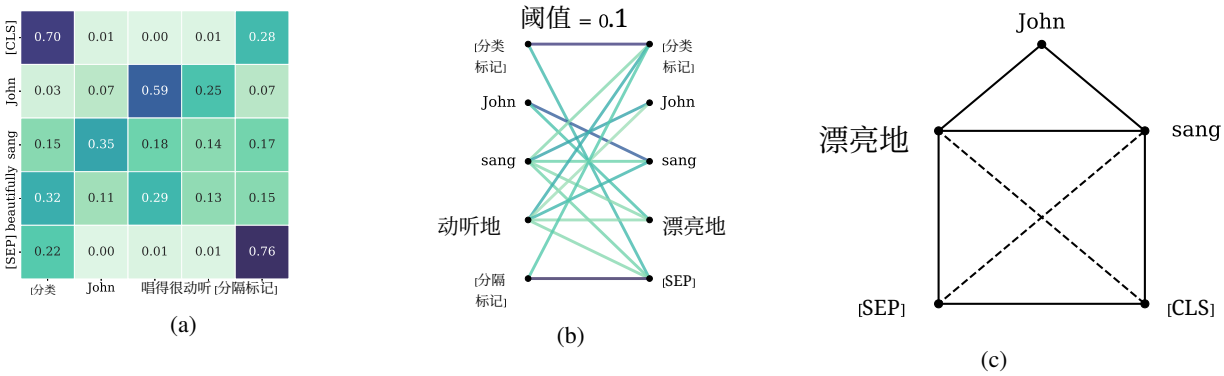

图 1：COLA 句子*"John sang beautifully"*的注意力图 (a) 以及对应的二分图 (b) 和注意力图 (c)。这些图的构建阈值为 0.1。

thr:

$$a_{ij} = \begin{cases} 1 & \text{if } w_{ij} \geq thr \\ 0 & \text{otherwise,} \end{cases}$$

其中，w_{ij} 是词元i 和 j 之间的注意力权重，n 是输入序列中的词元数量。每个词元都对应一个图节点。有向注意力图的特征包括强连通分量的数量、边数、简单环数量以及顶点平均度。无向图的性质包括前两个贝蒂数：连通分量的数量和简单环的数量。我们提出了无向注意力图的两个新特征：匹配数和弦性。匹配数是图中的最大匹配大小，即不共享节点的最大边集。

考虑图示于 图 1a 的注意力矩阵，以及基于二分图（图 1c）并以 0.1 为阈值构建的简单无向注意力图（图 1b）。该注意力图的匹配数为 2。该图中一个最大匹配的示例如下，由一组边构成：{(*John* - 唱歌), ([SEP]- [CLS])}。该匹配是最大的，因为不存在其他边能够连接尚未匹配的节点（词元）；已有匹配涉及 4 个节点（词元）。弦性是一种二元特征，用于表示注意力图是否为弦图；也就是说，注意力图是否不含长度大于 3 的诱导环。例如，图 1c所示的图是弦图，因为其中没有边数超过 3 的诱导环。如果图中没有虚线边（弦），就会存在一个长度为 4 的环 [SEP]-优美地-唱歌-[CLS]-[SEP]，这意味着该图不是弦图。

我们预期这些新特征能够表达输入文本的句法现象。弦性特征可能包含有关主语—动词—宾语三元组的信息。最大匹配可能对应于句子片段（从句、状语、分词、引导短语等）的匹配。

从条形码提取的特征包括 $0/1$ 维条形码的描述性特征，并反映连通分量和边在整个过滤过程中的存续情况（消亡与诞生）。

模式距离特征衡量注意力矩阵与预定义注意力模式的单位矩阵之间的距离，例如，对序列首个词元的注意力 [CLS] 以及对最后一个词元[SEP] 的注意力、对前一个和后一个词元以及对标点符号的注意力（Clark 等人，2019年）。我们使用公开可用的实现来计算特征。[3]

3.2 实验框架

数据我们使用了两种类型学特征迥异的语言中公开可用的 LA 基准：俄语（RuCoLA；Mikhailov 等人，2022年）和英语（CoLA；Warstadt 等人，2019年）。这两个选定的语料库均包含域内和域外数据，以及从语言学出版物中收集的句子；每个句子都被标注为可接受或不可接受。不可接受句子还标注了其中违反的句法、形态和语义现象。RuCoLA 此外还涵盖由生成式语言模型合成的数据。我们给出了从观测语料库中提取的可接受句子示例（2a，3a），以及

[3] https://github.com/danchern97/tda4atd

语义违规句子的示例（2b，3 b).

(2) a. 狗咬了猫。 b. *那只**既安稳又毛茸茸的**猫睡着了。(3) a. 在古埃及，猫是神圣的动物。（“猫在古埃及是神圣的动物。”）
b. *在扶手椅**外面**的一张小地毯上，一只猫躺着。（“一只猫躺在扶手椅外的一张小地毯上。”）

表4 (附录A) 汇报了所用语料库的统计信息。对于分类别评估，我们使用 RuCoLA 的错误标注；对于 CoLA，我们使用细分语法现象的标注来归类错误句子。更多细节见表5(附录A)。

模型 我们的基线模型架构、微调和评估脚本取自 Transformers 库（Wolf 等人，2020年）。实验中，我们使用以下区分大小写的单语 Transformer 语言模型：(1) base 规模的 En-BERT[4] (Devlin 等人，2019年) 和 Ru-BERT，[5](2) large 规模的 En-RoBERTa[6] (Liu 等人，2019年) 和 Ru-RoBERTa。[7] 为估计微调的影响，我们比较两类模型：权重冻结的预训练语言模型（冻结模型），以及在训练集上微调的语言模型。Transformer 语言模型在域内训练数据上微调 5 个 epoch，批次大小为 32，超参数采用数据集作者确定的最优设置。为缓解类别不平衡问题，我们使用加权交叉熵损失。微调细节见表6 (附录A)。

TDA 分类器我们从 Transformer 语言模型中提取第 3.1 节列出的一系列持久性（TDA）特征，并将其作为训练特征输入线性分类器。我们使用主成分分析（PCA）降低特征空间的维度。接下来，我们在降维后的特征空间上训练逻辑回归分类器，并调整类别权重。我们遍历一系列逆正则化参数值 $C \in \{10^{-3}, 10^{-2}, 0.1\}$ 以及主成分数量# $PC \in [10, 20 \dots 200]$。我们将 200 设为主成分网格的上限，以确保潜在特征的数量至少比域内开发集（IDD）或域外开发集（OODD）的规模少两倍。我们调整超参数，以最大化分类器在 IDD 集上的性能。我们通过报告在以下数据上训练的分类器结果，比较两组特征的性能：（i）Kushnareva 等人，2021年提出的基础 TDA 特征（称为 TDA）以及（ii）添加了两个新特征的 TDA 特征（称为 TDA_{ext}）。

3.3 评估

性能指标遵循 Warstadt 等人，2019年，我们使用准确率（准确率）和马修斯相关系数（马修斯相关系数）来衡量性能。马修斯相关系数用作寻找超参数、评估训练后的模型以及调整决策阈值的主要性能指标。

微调效应 我们使用两种方法估计预训练语言模型与微调后的语言模型之间注意力权重的变化。首先，我们遵循Hao 等人， 2020年并采用 Jensen-Shannon（JS）散度：

$$D_{JS}(M_t||M_0) = \frac{1}{N}\frac{1}{H}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{W}\sum_{i=1}^{K} D_{JS}(W_t^h(token_i)||W_0^h(token_i))$$

其中， M_t 和 M_0 分别是微调后的模型和冻结模型， N 是句子数量， H 是注意力头的数量（基础配置语言模型为$H = 12$，大型语言模型为 $H = 24$），K是句子 n中的词元数量，W_t^h（$token_i$）是模型 M_t中注意力头 h 在词元 i 处的注意力权重。

其次，我们将注意力图之间的差异估计为跨注意力头的 TDA_{ext} 特征之间的平均相关距离：

$$D_{TDA}(M_t, M_0) = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{F}\sum_{f=1}^{F} D_{corr}(V_{tf}^h, V_{0f}^h)$$

其中， F 是特征的数量， V_{tf}^h 是特征 f的取值，基于从模型 M_t中提取的注意力矩阵W_t^h计算得出。

4 结果

4.1 可接受性分类

表 1 报告了 LA 分类结果。使用 TDA 特征训练的线性分类器提升了

[4]hf.co/bert-base-cased[5] hf.co/sberbank-ai/ruBert-base[6] hf.co/roberta-large [7] hf.co/sberbank-ai/ruRoberta-Large

模型	微调后的语言模型				冻结语言模型			
	IDD		OODD		IDD		OODD	
	准确率	马修斯相关系数	准确率	马修斯相关系数	准确率	马修斯相关系数	准确率	马修斯相关系数
RUCOLA								
Ru-BERT	80.3	0.420	75.1	0.438	62.4	0.079	54.7	0.112
+ TDA	80.1	0.440	75.1	0.447	76.5	0.314	62.3	0.253
+ TDA $_{ext}$	80.1	0.478	73.2	0.440	76.7	0.331	62.6	0.270
Ru-RoBERTa	83.5	0.530	79.3	0.530	72.8	0.313	58.1	0.241
+ TDA	85.0	0.581	**81.0**	**0.584**	77.0	0.374	**64.7**	0.343
+ TDA $_{ext}$	**85.7**	**0.594**	80.1	0.558	**77.2**	**0.391**	64.2	**0.358**
CoLA								
En-BERT	85.0	0.634	82.0	0.561	62.6	0.039	64.3	0.124
+ TDA	85.6	0.649	81.4	0.548	77.0	0.484	68.4	0.335
+ TDA$_{ext}$	**88.2**	**0.726**	81.0	0.556	81.4	0.543	73.1	0.369
En-RoBERTa	87.3	0.692	**84.9**	**0.637**	74.0	0.317	75.0	0.362
+TDA	86.3	0.680	83.5	0.620	81.2	0.543	**78.5**	0.464
+ TDA$_{ext}$	87.3	0.695	83.1	0.604	**83.1**	**0.604**	77.3	**0.476**

表 1：基准测试基于特征集训练的单语语言模型和线性分类器的可接受性分类结果。**IDD**=域内开发集。**OODD**=域外开发集。TDA$_{ext}$=TDA 特征+弦性和匹配数。最佳分数以粗体显示，次佳分数带下划线。

Transformer语言模型的性能；这一趋势在所有模型中都一致，俄语语言模型的马修斯相关系数最高提升了 +0.252，而 En-BERT 的提升幅度更为显著，达到 +0.504。所提出的弦性和匹配数特征很有用，有助于提升性能，这证明它们能够捕捉语言信息。

与基础语言模型不同，大型冻结语言模型在微调之前就已展现出语法知识。基础语言模型的马修斯相关系数分数在零附近波动，而大型语言模型的马修斯相关系数至少达到 0.3。

这一观察与近期研究的发现一致：预训练的大型 En-RoBERTa 无需进一步微调，便能在词汇复杂度预测等任务中取得有竞争力的分数（Rao 等人，2021年）。

与此同时，TDA分类器的表现略优于微调后的模型，俄语的马修斯相关系数最高提升了 +0.064，英语则提升了+0.092。我们认为，微调可能会导致语言模型丧失通用语法能力，并遗忘微调数据集中未出现的语言现象（Miaschi 等人，2020年）。因此，从微调后的模型中提取的特征可能需要通过非线性模型进行全面的特征选择，以缓解特征冗余问题。用于 RuCOLA 的 TDA 分类器取得了与基线语言模型相当的分数。然而，对于 COLA，结合 En-RoBERTa 的 TDA$_{ext}$ 分类器优于基准——

行。我们报告了表 7 和表 8、附录B.1中的 OOD 测试数据分类结果。

4.2 对违反类别的敏感性

接下来，我们分析 TDA分类器在不同违规类别上的召回率提升。表 2报告了 Ru-BERT 和 En-BERT 基线模型以及 TDA分类器在 IDD 和 OODD数据集上针对 5 种语法违规的平均得分。TDA分类器在不可接受句子上的表现优于语言模型；这一上升趋势在两种语言中都成立，但在可接受句子上的表现有所下降。

与英语相比，在 Ru-BERT 特征上训练的 TDA$_{ext}$ 分类器对句法违规更加敏感，整体召回率达到 76.6；也就是说，与微调后的 Ru-BERT 相比，召回率提高了约 20 个百分点。至于其余语法类别，TDA$_{ext}$分类器大幅优于微调后的 Ru-BERT，尤其是在包含词级形态学违规的句子中，Ru-BERT 的召回率提高了一倍以上。

接下来，我们人工分析微调后的 Ru-BERT和我们的分类器TDA$_{ext}$ 在俄语 OODD句子中的错误。首先，我们比较那些被 Ru-BERT错误分类、但被TDA$_{ext}$ 分类器正确分类的不可接受句子。我们发现，OODD句子中的错误跨度相对较短，最多包含三个词元。具体而言，这些句子中最常见的违规现象包括不存在的词（母语者常常会误用这类词）（4a，构词错误“ekhaj”）、局部倒装语序（4b）或无意义表达（4c）。两个模型常见的错误预测包括：混合多种语法现象、存在远距离一致关系违规以及标点符号错误复杂的长句。

(4) a. * A ty **ekhaj** pryamo k direktoru teatrov.（“**你应该得去**剧院主任那里。”）
b. * V etom lesu**vodyatsya volki**.（“这片森林里**有狼出没**。”）c. * Oni chitali moi zhaloby **na se-bya**.（“他们读了我的抱怨，**针对他们自己**。”）

从 ID 到 OOD 的领域偏移引入了 ID 数据中不存在的新型不可接受现象。总体而言，OOD

模型	可接受	幻觉	形态学	语义	句法
Ru-BERT	92.1	53.9	20.0	25.0	55.7
+TDA$_{ext}$	80.6	73.9	53.9	46.6	76.6
En-BERT	94.3	68.5	69.4	63.0	55.6
+TDA$_{ext}$	84.5	78.8	82.5	76.3	73.0

表2：基准测试的整体类别级召回率。

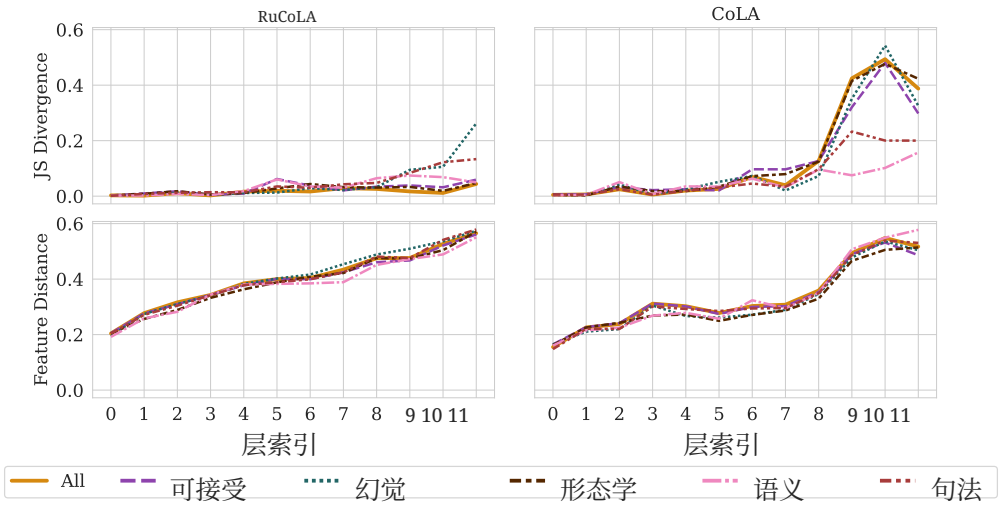

图 2：冻结的 Ru-BERT 和 En-BERT 与微调后的 Ru-BERT 和 En-BERT 之间逐层的特征距离及注意力分数的 JS 散度。

数据的得分低于 ID 数据（表2，表9，附录B.1）。因此，语言模型无法很好地泛化到未见过的不可接受现象，对未见过的语言属性也知之甚少。

4.3 微调效应

我们研究语言模型微调的动态，并测量冻结模型和微调后的语言模型在 OODD 子集上提取的 TDA$_{ext}$特征之间的逐层距离（§3.3）。图2展示了 Ru-BERT 和 En-BERT 的逐层特征距离及 JS 散度（图 3，大型模型见附录B.2）。总体而言，我们发现特征之间的距离从底层到较高层稳步增大；而对于英语语言模型，最显著的变化仅出现在最后四层。这一观察表明，En-BERT 与 Ru-BERT 的微调动态存在明显差异。

对于两种语言，特征距离的趋势都不同于 JS 散度，尤其是在前六层。这表明，TDA$_{ext}$ 特征可用于检测较低层中的细微变化，而这些变化使用 JS 散度时难以充分体现。例如，在较低的预设阈值下，基于 TDA 的距离对注意力权重的微小变化很敏感，即使较大的注意力权重保持不变也是如此。JS 散度无法捕捉这类情况。

不同违规类别之间的特征距离分布均匀。可接受句子与不可接受句子的趋势几乎一致，不过 JS 散度仍存在明显差异。对于俄语模型，与其他类别相比，含有句法违规和幻觉的句子在较高层中的 JS 散度更为明显。相比之下，英语的 JS 散度表明，对于含有语义和句法违规的句子，注意力模式与冻结的 En-BERT 更为一致；对于可接受句子和其他句子，峰值则出现在倒数第二层。与基础配置的语言模型类似，特征差异在所有层中稳步增大；而对于英语，主要变化出现在较高层。

错误类型	句子	特征	Head
形态学	食谱**薯片** s 奶酪, **木薯** i 鸡蛋. (“奶酪木薯鸡蛋**cheps**食谱。”)	$c_{thr0.25}$	(9,5)
句法	开始了**修建新的行政区**。 (“**新区建设**已启动。”)	$c_{thr0.1}$	(9,5)
语义	昨天两点钟商店**关门**。 (“昨天两点钟商店**正在关门**。”)	[CLS]	(11,0)

表 3：用于按错误类型判断 RUCOLA 不可接受句子的最重要 Ru-BERT TDA$_{ext}$ 特征示例。 c = 图中的简单环数量， thr =用于构建注意力图的阈值，[CLS] =到[CLS]标记的距离。

4.4 注意力头重要性

我们借助持久特征探究语言现象：利用线性分类器中学到的特征权重（附录 B.3）。特征权重越高，该特征对最终预测的贡献就越大。我们汇总来自各个中心语的特征：中心语的重要性根据重要特征的数量确定。我们定义两类中心语：（1）对真阳性和真阴性预测（即正确预测）贡献最大的中心语，称为一致注意力头；（2）对假阴性和假阳性预测（即分类器的错误）贡献最大的中心语，称为不一致注意力头。首先，我们考察每个独立中心语的重要性。图 4 、附录 B.4 展示了中心语对最终预测的重要程度。En-BERT 和 Ru-BERT 的（1）类中心语呈现相似模式：对 Ru-BERT 最有用的特征位于中间层至较高层。对于 En-BERT，这些特征则大多集中在最后两层。

接下来，我们根据违规类别计算特征重要性。中间层的中心语对检测英语和俄语中的句法及形态学违规贡献更大。除了少数例外，类型（2）的中心语与类型（1）的中心语并不重叠；这些例外分别是 Ru-BERT 最后一层的中心语 10 和 En-BERT 最后一层的中心语 0。根据类型（2）中心语的数量来看，Ru-BERT 在区分带有幻觉的句子与可接受句子方面遇到的困难最大。这可能是由多种原因造成的：（i）训练期间没有见过带有幻觉的句子；（ii）带有幻觉的句子通常在形式上组织良好，但在语义上不正确，因此没有可依赖的表层或句法线索可依赖的。

接下来，我们确定哪些句子对 TDA分类器最具挑战性，也就确定了相应语言模型面临的挑战，因为 TDA 特征是从其注意力图中提取的。为此，我们将语言模型的置信度定义为预测可接受类和不可接受类时特征权重绝对值之和。分数越低，语言模型就越困惑，越多注意力头倾向于与预期预测意见不一。我们将置信度分数最低的句子视为具有挑战性的句子。最具挑战性的句子篇幅较长，包含多个分句和术语或命名实体，例如5 中的不可接受句子。为求完整，我们对 CoLA 句子进行了相同分析，并给出了一个最令 TDA 分类器困惑的句子示例（$_{ext}$ 6）。结果吻合良好。最具挑战性的句子包含长距离依存关系和命名实体。

(5) * 埃塔群体发现**(不正确地)认为，梅里马克河的北部分水岭** 靠近如今被称为**温尼珀索基湖**的地方，该湖位于**新罕布什尔州**。（“该群体（**不恰当地）发现，梅里马克河**的北部分水岭靠近如今被称为**温尼珀索基湖**的地方，该湖位于**新罕布什尔州**。”）
(6) * 古尔德演奏**巴赫钢琴曲**，远不如罗斯用羽管键琴演奏的版本令我满意。

最后，我们探究句子级别的特征贡献。基于拓扑数据分析的方法能够解释每个句子的预测。为此，每个特征的贡献（=重要性）等于特征值乘以

线性分类器学习到的权重。我们观察到俄语和 Ru-BERT 的不可接受句子中存在以下模式：

1. 模式距离特征似乎有助于对包含词级错误的不可接受句子进行分类，包括拼写、标点符号和一致性错误；

2. 拓扑特征和从条形码中提取的特征对更复杂的语法现象具有同等贡献。

表 3列出了不可接受句子的示例及其特征重要性值。在俄语中，由阈值 0.1 或 0.25 得出的弦性、匹配数、简单环数量和平均顶点度，往往对预测不可接受句子具有重要作用。同样，平均顶点度数对英语和 En-BERT 具有最强的判别能力。在语言模型的不同层中均存在重要特征。对于英语，最重要的特征提取自最后一层；而对于俄语，这些特征最早出现在第 6 层。

然而，在可接受句子与不可接受句子的注意力图差异方面，我们发现两种语言都存在以下共同特征。不可接受句子的注意力图在最低和最高阈值下具有更多的连通分量。在最高阈值下，这些连通分量由单个词元组成；而在最低阈值下，它们则由少数几个词元组成。这意味着，不可接受句子中注意力图的数值彼此之间差异不大。相反，对于可接受句子，模型倾向于将最大的注意力权重分配给单个词元，而该词元通常是句子的中心动词。从 TDA 特征值的角度来看，这种效应可以通过特征值与目标类别之间的相关系数的符号来体现。因此，在阈值为 0.5 时，平均顶点度特征明显呈现出向正相关转变的趋势（图 5）。

总而言之，这种分析有助于更好地解释分类器的预测。由于持久特征归属于各个中心语，我们可以追踪每个中心语的作用和重要性。对语言现象进行细粒度标注，有助于我们将特定的语言技能与各个中心语联系起来。

5 结论

本文采用并改进了可接受性分类方法，借鉴了拓扑数据分析（TDA）中的最佳实践。我们使用英语和俄语的数据集 COLA 和 RUCOLA，分别展示了所开发方法在这两种类型学差异显著的语言中的应用。具体而言，我们提出了两个新特征：弦性和匹配数，并将基于 TDA 的分类器性能与微调进行比较。基于 TDA 的分类器能够提升预训练语言模型的性能。

基于 TDA 的分类器相比语言模型微调具有优势，因为它们更易于解释，并有助于深入探究语言模型的内部工作机制。为此，我们引入了一种基于 TDA 特征的距离度量方法，用于检测语言模型在微调过程中注意力模式的变化。与广泛使用的 Jensen-Shannon 散度不同，这种距离度量方法甚至能够敏锐地检测到语言模型底层中发生的细微变化。此外，更重要的是，我们展示了 TDA 特征如何揭示注意力头的功能作用。我们根据重要性，对有助于产生正确预测和错误预测的注意力头进行比较。通过这种方式，我们发现了一些在不可接受的句子中存储有关词序、词语派生和复杂语义现象信息的注意力头，以及一些关注可接受句子的注意力头。

给定句子后，我们根据特征的贡献评估预测置信度。我们确定了语言模型预测置信度较低的句子集合，并发现这些句子通常由多个分句组成，而且经常包含命名实体。最后，我们发现，在英语和俄语的不可接受句子的注意力图中，频繁出现一种独特模式。

我们希望，研究结果能够阐明语言模型在俄语和英语中的性能，帮助理解其微调动态以及注意力头的功能作用。我们期待自然语言处理从业者将TDA应用于其他语言和下游问题。

局限性

可接受性判断数据集可接受性判断数据集以语言学文献作为不可接受句子的来源。这种方法受到两方面的批评：（i）

可接受性判断的可靠性与可复现性（Gibson 和 Fedorenko，2013年；Culicover 和 Jackendoff，2010；Sprouse 和 Almeida，2013；Linzen 和 Oseki，2018年），（ii）代表性，因为语言学家的判断可能无法反映说话者往往会产生的错误（D.abrowska，2010）。

计算复杂度 所提出特征的计算复杂度为线性。对于弦性特征，我们依赖于线性 $O(|E|+|V|)$ 时间算法的实现（Tarjan 和 Yannakakis，1984年），其中 $|E|$ 和 $|V|$ 分别是边数和节点数。我们使用线性复杂度的贪心算法 $O(|E|)$ 来寻找最大匹配。使用指数时间算法计算简单环时（最坏情况下），我们将约束设为 500，以便提前停止。我们认为，超过该值后，简单环特征所包含的信息较少。 Kushnareva 等人， 2021年 讨论了其余特征的时间复杂度。

致谢

我们感谢 Laida Kushanareva、Daniil Cherniavskii、Vladislav Mikhailov、Serguei Barannikov、Alexander Bernstein 和 Dmitri Piontkovski 在项目早期阶段提出的意见，并感谢 Max Ryabinin 提供处理 RUCOLA 数据集的脚本。

参考文献

J Alammar。2021年。Ecco：一个用于解释 Transformer 语言模型的开源库。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议：系统演示论文集，第249–257页，线上举行。计算语言学协会。

Sam Bowman、Alex Warstadt、Amanpreet Singh。2018年。Cola：域外开放评估。

Pepa Atanasova、Jakob Grue Simonsen、Christina Lioma 和 Isabelle Augenstein。2020年。文本分类可解释性技术的诊断研究。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第3256–3274页，线上举行。计算语言学协会。

Soumya Batra、Shashank Jain、Peyman Heidari、Ankit Arun、CatharineYoungs、XintongLi、PinarDonmez、Shawn Mei、Shiunzu Kuo、Vikas Bhardwaj、Anuj Kumar，Michael White。2021年。用于神经自然语言生成的自适应可接受性分类器。载于《*2021*年自然语言处理经验方法会议论文集》，第682–697页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Jatin Chauhan 和 Manohar Kaul。2022年。BERTops：从拓扑学视角研究 BERT表示。载于 *2022*年国际神经网络联合会议（*IJCNN*），第1–8页。IEEE。

Daniil Cherniavskii、Eduard Tulchinskii、Vladislav Mikhailov、Irina Proskurina、Laida Kushnareva、Ekaterina Artemova、Serguei Barannikov、Irina Pi- ontkovskaya、Dmitri Piontkovski 和 Evgeny Bur- naev。2022年。通过考察注意力图拓扑进行可接受性判断。载于计算语言学协会会议论文集：*EMNLP 2022*，第 88–107 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Colin Cherry 和 Chris Quirk。2008年。通过潜在支持向量机进行判别式句法语言建模。载于美洲机器翻译协会第八届会议论文集：研究论文，第65–74页，美国怀基基。美洲机器翻译协会。

Kevin Clark、Urvashi Khandelwal、Omer Levy 和 Christopher D. Manning。2019年。BERT关注什么？对BERT注意力的分析。载于*2019*年 *ACL BlackboxNLP*研讨会论文集：分析和解释自然语言处理神经网络，第276–286页，意大利佛罗伦萨。计算语言学协会。

Pierre Colombo、Guillaume Staerman、Chloé Clavel 和 Pablo Piantanida。2021年。透过Wasserstein重心的视角进行自动文本评估。载于*2021*年自然语言处理实证方法会议论文集，第10450–10466页，线上会议及多米尼加共和国蓬塔卡纳。计算语言学协会。

Peter W Culicover 和 Ray Jackendoff。2010年。仅靠定量方法还不够：对Gibson和Fedorenko的回应。认知科学趋势，6(14)：234–235。

Ewa D.abrowska。2010 年。朴素直觉与专家直觉：关于可接受性判断的实证研究。《语言学评论》。

Marina Danilevsky、Kun Qian、Ranit Aharonov、Yan- nis Katsis、Ban Kawas 和 Prithviraj Sen。2020年。 自然语言处理可解释人工智能现状综述。收录于 计算语言学协会亚太分会首届会议暨第十届自然语言处理国际联合会议论文集，第447–459页，中国苏州。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。 BERT：用于语言理解的深度双向 Transformer 预训练。收录于 计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Nadir Durrani、Hassan Sajjad 和 Fahim Dalvi。2021年。迁移学习如何影响深度自然语言处理模型中的语言学知识？ 载于 计算语言学协会：*ACL-IJCNLP 2021* 年论文集，第 4947–4957 页，线上举行。计算语言学协会。

Yanai Elazar、Nora Kassner、Shauli Ravfogel、Abhi- lasha Ravichander、Eduard Hovy、Hinrich Schütze 和 Yoav Goldberg。2021年。测量并改进预训练语言模型的一致性。《计算语言学协会会刊》，9:1012–1031。

Edward Gibson 和 Evelina Fedorenko。2013年。句法和语义研究中对定量方法的需求。《语言与认知过程》，28(1-2):88–124。

Yaru Hao、Li Dong、Furu Wei 和 Ke Xu。2020年。探究 BERT 微调的学习动态。载于 第*1*届计算语言学协会亚太分会会议暨第*10*届自然语言处理国际联合会议论文集，第87–92页，中国苏州。计算语言学协会。

Jae-young Jo 和 Sung-Hyon Myaeng。2020年。基于 Transformer 的神经语言模型中注意力头的作用与运用。收录于第*58*届计算语言学协会年会论文集，第3404–3417页，线上举行。计算语言学协会。

Olga Kovaleva、Alexey Romanov、Anna Rogers 和 Anna Rumshisky。2019年。 揭示 BERT 的黑暗秘密。载于 *2019*年自然语言处理经验方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第 4365–4374 页，中国香港。计算语言学协会。

Laida Kushnareva、Daniil Cherniavskii、Vladislav Mikhailov、Ekaterina Artemova、Serguei Barannikov、Alexander Bernstein、Irina Piontkovskaya、Dmitri Piontkovski 和 Evgeny Burnaev。2021年。通过考察注意力图拓扑检测人工生成文本。载于*2021*年自然语言处理实证方法会议论文集，第 635–649 页，线上会议及多米尼加共和国蓬塔卡纳。计算语言学协会。

Jey Han Lau、Carlos Armendariz、Shalom Lappin、Matthew Purver 和 Chang Shu。2020年。无色的绿色思想能睡得多么狂躁？语境中的句子可接受性。计算语言学协会汇刊，8:296–310。

Tal Linzen 和 Yohei Oseki。2018年。跨语言可接受性判断的可靠性。《*Glossa*：普通语言学期刊》，3(1)。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Alessio Miaschi、Dominique Brunato、Felice Dell'Orletta 和 Giulia Venturi。2020年。神经语言模型的语言特征分析。载于第*28*届国际计算语言学会议论文集，第745–756页，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Vladislav Mikhailov、Tatiana Shamardina、Max Ryabinin、Alena Pestova、Ivan Smurov 和 Ekaterina Artemova。2022年。RuCoLA：俄语语言可接受性语料库。载于*2022*年自然语言处理实证方法会议论文集，第5207–5227页，阿联酋阿布扎比。计算语言学协会。

Matt Post。2011年。使用树替换语法推导判断语法性。载于计算语言学协会第*49*届年会论文集：人类语言技术，第217–222页，美国俄勒冈州波特兰。计算语言学协会。

Gang Rao、Maochang Li、Xiaolong Hou、Lianxin Jiang、Yang Mo 和 Jianping Shen。2021年。SemEval-2021任务1中的 RG PA：一种基于上下文注意力的 RoBERTa 词汇复杂度预测模型。载于第*15*届国际语义评测研讨会（*SemEval-2021*）论文集，第623–626页，线上举行。计算语言学协会。

Vinit Ravishankar、Memduh Gökırmak、Lilja Øvrelid 和 Erik Velldal。2019年。深度预训练上下文编码器的多语言探测。载于首届自然语言处理深度学习研讨会论文集，第37–47页，芬兰图尔库。林雪平大学电子出版社。

Jon Sprouse 和 Diogo Almeida。2013年。句法中数据的实证地位：对 Gibson 和 Fedorenko 的回应。语言与认知过程，28(3):222–228。

Ekaterina Taktasheva、Vladislav Mikhailov 和 Eka- terina Artemova。2021年。在芝麻街摇晃句法树：通过可控扰动进行多语言探测。第一届

多语言表征学习研讨会，第191–210页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Robert E Tarjan 和 Mihalis Yannakakis。1984年。用于测试图的弦性、测试超图的无环性以及选择性约简无环超图的简单线性时间算法。《SIAM计算杂志》，13(3)：566–579。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。Attention is All You Need。神经信息处理系统进展，30。

JoachimWagner、JenniferFoster 和 Josefvan Gen- abith。2009年。判断语法性：句子分类实验。《*Calico*期刊》，26(3):474–490。

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy 和 Samuel Bowman。2018年。GLUE：用于自然语言理解的多任务基准及分析平台。载于 *2018*年 *EMNLP Workshop BlackboxNLP*：自然语言处理神经网络的分析与解释论文集，第 353–355 页，比利时布鲁塞尔。计算语言学协会。

Alex Warstadt 和 Samuel R. Bowman。2019年。利用可接受性判断对预训练句子编码器进行语言学分析。*arXiv*：计算与语言。

Alex Warstadt、Amanpreet Singh 和 SamuelR. Bow- man。2019年。神经网络可接受性判断。计算语言学协会会刊，7：625–641。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformers：自然语言处理的最新进展。载于*2020*年自然语言处理实证方法会议论文集：系统演示，第38–45页，线上。计算语言学协会。

Yian Zhang、Alex Warstadt、Xiaocheng Li 和 Samuel R. Bowman。2021年。你何时需要数十亿词的预训练数据？载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第1112–1125页，线上举行。计算语言学协会。

A 实验设置

	CoLA	RUCOLA
语言	英语	俄语
数据类型	Real	真实数据, 合成数据
α	0.86	0.89
# 训练句子数	8551	7869
# 开发集句子数	527	983
# 测试集句子数	516	1804
%	70.5	71.8

表4：语言可接受性语料库的统计数据。 α = 平均标注者一致率。% = 可接受句子的百分比。

语法特征	错误类型
多词/漏词	幻觉
语义违规	语义
屈折变化/一致性违规	形态学
其他	句法

表5：按错误类型汇总的 COLA 特征。Infl/Agr =屈折变化与一致性。其他=COLA标注中存在于不可接受句子里的其他语法违规现象。

模型	学习率	权重衰减
En-BERT	$3 \cdot 10^{-5}$	0.01
En-RoBERTa	$2 \cdot 10^{-5}$	10^{-4}
Ru-BERT	$3 \cdot 10^{-5}$	0.1
Ru-RoBERTa	10^{-5}	10^{-4}

表 6：用于微调 Transformer 的超参数值。

B 实验结果

B.1 语言可接受性分类

模型	专家 准确率 马修斯相关系数	机器 准确率 马修斯相关系数
Ru-BERT	77 0.37	75 0.44
+ TDA $_{ext}$	75 0.39	72 0.42
Ru-RoBERTa	84 0.55	80 0.56
+ TDA $_{ext}$	83 0.53	80 0.56

表 7：单语语言模型和线性分类器在 RUCOLA 域外测试集上的语言可接受性分类结果（按源模型划分）。[8]

模型	MCC
En-BERT	0.509
+ TDA $_{ext}$	0.469
En-RoBERTa	0.608
+ TDA $_{ext}$	0.616

表 8：单语语言模型和线性分类器在 CoLA 域外测试集上的可接受性分类结果（AlexWarstadt，2018年）。

模型	可接受	幻觉	形态学	语义	句法
RUCOLA IDD					
Ru-BERT	93.9	-	12.5	24.0	56.0
+TDA$_{ext}$	86.2	-	56.2	45.0	75.4
Ru-RoBERTa	95.9	-	50.0	37.0	70.9
+TDA$_{ext}$	96.3	-	31.2	34.0	72.4
RUCOLA OODD					
Ru-BERT	90.3	53.9	26.6	25.9	55.4
+TDA$_{ext}$	75.0	73.9	51.6	48.1	77.7
Ru-RoBERTa	90.9	64.3	54.7	42.0	75.5
+TDA$_{ext}$	89.9	63.9	53.1	39.5	71.4
COLA IDD					
En-BERT	94.8	65.0	69.0	72.2	61.2
+TDA$_{ext}$	87.9	77.5	86.2	83.3	82.4
En-RoBERTa	94.8	72.5	88.9	75.9	64.7
+TDA$_{ext}$	87.3	75.0	79.3	88.9	70.6
COLA OODD					
En-BERT	93.8	72.0	69.7	53.8	50.0
+TDA$_{ext}$	81.0	80.0	78.8	69.2	63.5
En-RoBERTa	93.5	76.0	87.9	76.9	56.2
+TDA$_{ext}$	83.1	80.0	81.8	92.3	63.5

表9：各基准测试在 IDD 和 OODD数据集上的类别级召回率。

[7]https://rucola-benchmark.com [8] https://rucola-benchmark.com

B.2 微调效应

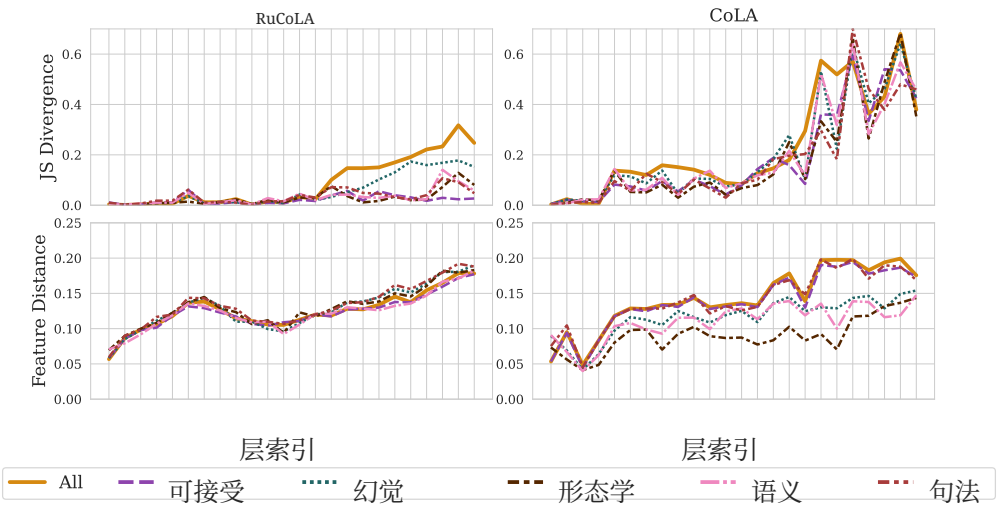

图 3：冻结版与微调后的 Ru-RoBERTa 和 En-RoBERTa 各层的特征距离及注意力分数的詹森-香农散度。

B.3 特征重要性

考虑一个采用 L1 正则化的线性分类器，则句子 i 的输出概率为：

$$p_i \sim \exp(X_{0i}^T C^T w + c),$$

其中，X_{0i} 为输入的 TDA 特征，C 为主成分矩阵，w^T 为决策函数中主成分系数构成的向量，c 为添加的偏置。$C^T w$ 为对预测的特征贡献。

B.4 注意力中心语的作用

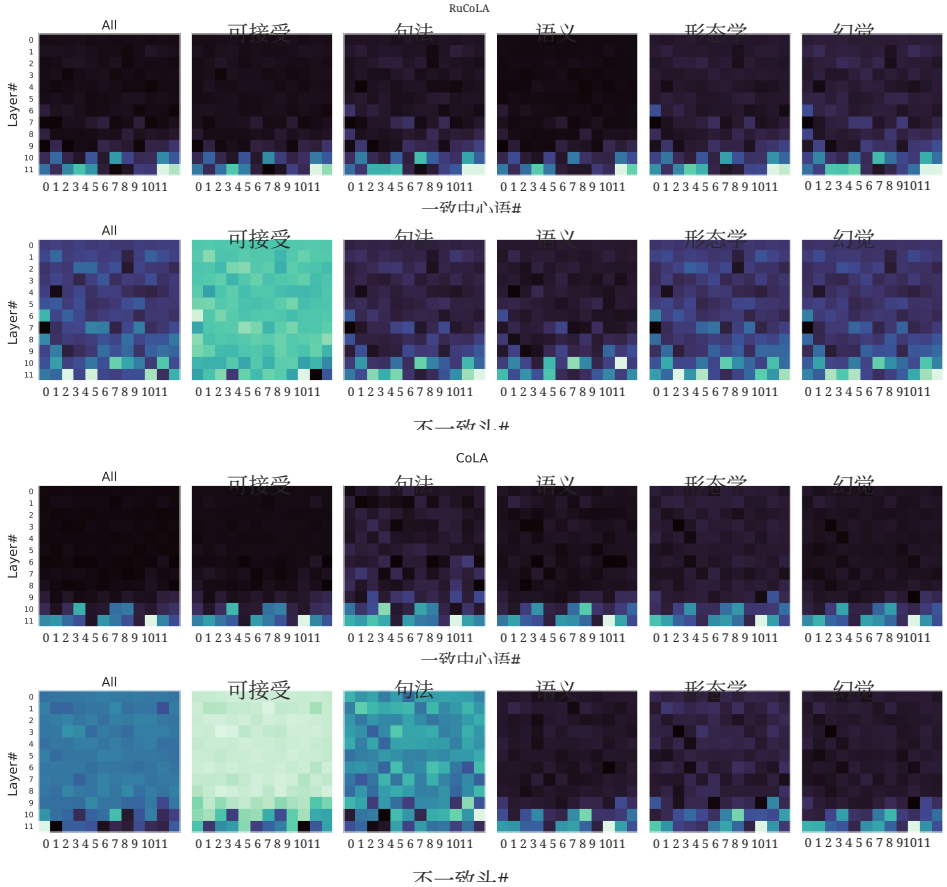

图4：TDA$_{ext}$ 分类器中各特征相对于数据集的平均特征权重。TDA$_{ext}$ 分别从微调后的Ru-BERT和En-BERT中提取。一致中心语的特征有助于正确预测。不一致头的特征会导致错误预测。颜色越亮，表示平均特征权重越高。

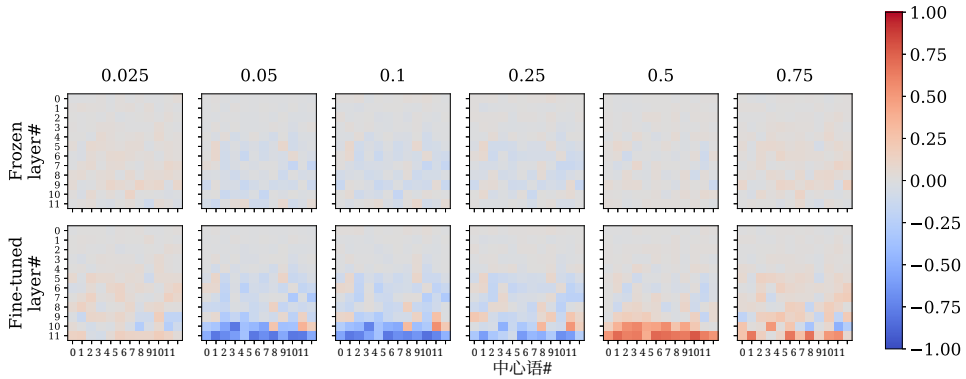

图5：冻结和微调后的Ru-BERT中，平均顶点度特征与目标标签之间的相关系数，按用于构建注意力图的阈值划分。