QUST团队在SemEval-2023 任务3中的研究：检测在线新闻体裁、框架构建和说服技巧的单语与多语方法全面研究

Ye Jiang中国青岛科技大学
ye.jiang@qust.edu.cn

arXiv:2304.04190v1 [cs.CL] 9 Apr 2023

摘要

本文介绍了QUST团队参与SemEval2023任务3的情况。在任务早期阶段，首先通过对多数类进行欠采样来评估单语模型。随后，使用类别权重和样本权重的组合对预训练多语言模型进行微调。进一步研究了两种不同的微调策略：任务无关策略和任务相关策略。所有实验均采用10折交叉验证，多语方法优于单语方法。所提交的系统[1]在子任务 1中意大利语和西班牙语（零样本）上取得了第二好的成绩。

1 引言

任务3（Piskorski 等人，2023）要求参与者开发算法，以在如表1所示的多语言设置中自动检测新闻体裁、框架构建和说服技巧。本任务涵盖六种不同语言，包括英语、法语、德语、意大利语、波兰语和俄语。此外，西班牙语、希腊语和格鲁吉亚语这三种惊喜语言也被纳入最终测试阶段，以开展零样本学习场景。

该任务的主要挑战之一在于，模型必须具备应对训练数据不平衡且不足问题的能力。在子任务1 中，标签“re- porting”、“意见”和“讽刺”的样本数量分别为269、878 和 87。与“意见”等其他多数类别相比，“讽刺”样本寥寥无几，训练模型很容易在多数类别上过拟合，而无法从少数类别中充分学习。例如，许多批次中都不会包含“讽刺”样本，因此在训练阶段，梯度所包含的信息会减少，从而导致模型在预测少数类别时的表现不佳。均值-

[1]https://github.com/zgjiangtoby/SemEval2023_QUST

与此同时，先前的研究（Leite 等人，2020；Zhang 等人，2021）表明，训练数据的规模可能会显著影响模型性能。然而，子任务 1 中的训练样本总数为 1234，这不足以有效地从头训练深度学习模型。

为解决上述问题，我们首先在单语设置下，对采用多数类欠采样的传统机器学习方法和深度学习方法进行评估。然后，对预训练多语言模型进行微调，以学习不同语言之间的跨语言特征。具体而言，微调是利用预训练语言模型（PLM）执行多种下游任务的常用方法。通常，这种方法会将原始输出层替换为新的任务专用层，然后对整个模型进行微调。为了建模不同语言新闻文章的特征表示，我们对XLM-RoBERTa（XLM-R）（Conneau 等人，2020）进行微调，因为它能够处理任务3中的所有语言，并且通常优于 mBERT 等其他模型（Kenton 和 Toutanova，2019）。

此外，本研究还计算样本权重和类别权重，以应对数据不平衡。样本权重使训练批次能够较好地涵盖数据集。类别权重则通过提高少数类的类别权重，同时降低多数类的权重，来惩罚对少数类样本的错误分类。

本研究在三个子任务上采用了两种微调方法：任务无关和任务相关。任务无关策略是一种标准方法，即针对每个子任务分别对预训练语言模型进行微调。本研究还采用了任务相关策略，将预训练语言模型从子任务 1 微调至子任务3，因为所有训练样本的内容都相似，只有标签不同。

为进行评估，本文采用 10 折交叉

文章编号 (段落编号)	Text	标签		
		T1	T2	T3
710376094 (4)	……名为3200的小行星法厄同……被归类为“可能有害”由小行星中心...	意见	安全和辩护，质量生活的	重复
23114 (17)	法国国家统计与经济研究所的研究以及德语的研究和法国人口研究所还确认社会流动阶梯运行……对于本土出生者(第四代和加)...	报告	公平和平等	已加载 Language
2672（3）	两千六百万公民关在家里。被禁足。哎居家隔离者。这些居民在上海……正接受娱乐主管...一次极其严格的封锁措施	讽刺	政策处方和评估，生活质量	已加载 Language 语言，名称呼叫-标注

表 1：数据集中的随机示例。T1（子任务1）、T2（子任务2）和 T3（子任务3）分别是文档级多类别、文档级多标签和句子级多标签任务，段落编号（段落编号）仅在子任务3中有效。

验证（CV），全面分析了不同的单语和多语方法，并研究使用样本权重和类别权重来解决数据不平衡问题的有效性。实验结果表明，在任务相关设置下，使用样本权重和类别权重对预训练语言模型进行微调，可以显著提升模型性能。此外，任务相关微调策略略优于任务无关方法，表明最终微调后的模型能够学习所有子任务之间的共享信息。

2 系统描述

提交的系统采用标准的微调方法，通过添加线性分类器，将$[CLS]$嵌入映射为输出类别上的未归一化概率向量。为解决数据不平衡问题，类别权重使交叉熵损失能够为少数类分配更高的权重。因此，分配的权重能够对少数类施加更高的惩罚，算法也就可以着重减少少数类的错误。

此外，训练集中“讽刺”样本仅占 7%。这可能导致训练批次中只有少量甚至没有少数类样本。为解决这一问题，样本权重旨在计算每个样本在整个数据集中的权重，然后为每个样本分配权重，以确保多数类的权重较小，反之亦然。除类别权重外，子任务2和子任务3也会计算样本权重。

在采用所有正向策略进行训练后，应用早停，以保存每一折中的最佳模型。最终，从10折交叉验证中选出的表现最好的3个模型进行集成，用于每个子任务的最终预测。

2.1 加权微调

类别权重用于调整交叉熵损失，这是用于衡量预测概率分布与真实标签分布之间差异的标准损失函数。

类别权重是一个与每个类别标签相对应的值向量，并且会与每个样本的交叉熵损失相乘。具体而言，类别权重可以为少数类分配更多权重，这意味着损失函数会因错误分类少数类样本而对模型施加更大的惩罚。最终，模型能够学会更加关注少数类，并提升其在不平衡数据上的表现。形式化地说，给定总数为 n训练样本，类别权重 C_w^j 对于类别 j可以简单计算如下：

$$C_w^j = \frac{n}{c \times n_j} \tag{1}$$

其中， c 表示类别数量，n_j 是类别 j中的样本总数。

类别权重可以与每个样本的交叉熵损失相乘，以调整损失函数。具体而言，设 y_c 是类别 j的真实标签，且 p_j 是预测概率，则类别 j 的交叉熵损失为：

$$L_j = -C_w^j \times y_j \times log(p_j) \quad (2)$$

总损失是所有类别损失之和。

类别加权能够缓解数据不平衡问题的影响。然而，鉴于少数类的规模相当小，训练批次中很可能只有少量甚至没有少数类样本，从而导致模型收敛速度变慢。

为解决这一问题，加权随机采样器能够重新采样训练数据，从而使每个批次都包含来自各个类别的相对均衡数量的样本。加权随机采样器根据每个样本的类别频率为其分配权重，然后按照权重从数据集中随机选择样本。

2.2 任务无关与任务相关

任务无关策略是一种标准方法，即在每个子任务上分别微调预训练语言模型。具体而言，预训练语言模型使用预训练权重进行初始化，然后根据每个子任务的特定数据和标签进行训练。任务无关策略有助于预训练语言模型适应每个子任务的领域和任务，但不会利用子任务之间的共享信息或层次结构。

任务相关策略是一种替代方法，它将预训练语言模型从子任务 1 微调到子任务3，因为所有训练样本的内容都相同，只是标签不同。这意味着，预训练语言模型首先使用预训练权重进行初始化，然后在子任务 1 的数据和标签上进行训练。之后，预训练语言模型使用子任务 1 的权重作为初始权重，在子任务 2 的数据和标签上进一步训练。同样，预训练语言模型使用子任务 2 的权重作为初始权重，在子任务3 的数据和标签上进一步训练。任务相关策略能够帮助预训练语言模型将从先前子任务中学习到的知识和特征迁移到后续子任务中。

	T1	T2	T3
最小值	88	67	3
最大值	7747	7747	1069
平均值	763	761	47

表 2：每个子任务中的最小、最大和平均值令牌数。

3 实验设置

3.1 数据

子任务1 和子任务2中新闻文章的总数分别为1234和 1238。子任务 3 将每篇文章划分为段落，因此段落总数为 20207。三个子任务中的词元数量的最大值、最小值和平均值计算结果如表 2所示。本文讨论的所有实验均采用 10 折交叉验证，展示的结果为平均值。具体而言，对于每个子任务，数据集被划分为10 折，并且任务无关策略和任务相关策略使用相同的折。对于每一折，将其中一折用作测试集，其余九折用作训练集。

3.2 预处理

数据中移除了标点符号、链接、转义字符（例如“\n”）和数字，并将字符串转换为小写。所有新闻文章都会被截断或填充至最多512个词元，不同语言被合并到一个JSON文件中。在子任务3中，添加了一个“None”类别，用于表示没有标签的段落。

在数据版本三（v3）中，组织方发布了开发集的标签。直观而言，扩大训练数据的规模有助于提升模型性能。因此，v3 合并了所有语言的训练数据和开发数据，以进行 10折交叉验证。子任务 1 的结果采用宏平均 F1 评估，其他子任务则采用微平均 F1 评估。

3.3 模型配置

在构建最终系统之前，我们对多个基线模型进行了比较。鉴于训练数据的规模相对较小，我们首先在单语设置下评估传统机器学习方法。所有传统方法均基于词频-逆文档频率向量构建，并对多数类进行欠采样。然后，深度学习方法包括卷积神经网络（Jiang 等，2019）、循环神经网络（Lin 等，

模型	en	fr	de	it	po	ru
SVM	0.36	0.37	0.31	0.16	0.38	0.14
LG	0.42	**0.48**	0.42	**0.29**	**0.67**	**0.29**
RF	**0.49**	0.32	0.33	0.27	0.50	0.24
BERT卷积神经网络	0.21	0.37	0.37	0.25	0.47	0.23
BERT循环神经网络	0.19	0.35	0.35	0.19	0.49	0.21
BERT自注意力	0.26	0.31	0.28	0.13	0.37	0.22
RBERT-CNN	0.41	0.43	**0.44**	**0.29**	0.58	0.25
RBERT-RNN	0.29	0.35	**0.44**	0.23	0.55	0.21
RBERT-Self	0.24	0.35	0.41	0.23	0.55	0.17

表 3：单语模型在v2 子任务1上的宏平均 F1。SVM、LG和 RF 分别表示支持向量机、逻辑回归和随机森林。卷积神经网络、循环神经网络和自注意力模型均基于直接从BERT 和 RBERT（RoBERTa）编码得到的 $[CLS]$ 嵌入构建。

2017) 和自注意力（Jiang 和 Wang，2022）分别基于单语预训练语言模型的 $[CLS]$ 嵌入构建。传统方法和深度学习方法均使用 *imblearn* （Lemaître 等，2017）和*skorch*（Tietz 等，2017）。最后，还对微调后的预训练语言模型的任务无关方法和任务相关方法进行了评估。

最终系统基于预训练XLM-RoBETRa，并使用 huggingface 库构建。AdamW（Loshchilov 和 Hutter，2017）优化器的学习率设为 3e-5，训练轮数为30。在每一折中，提前停止容忍度为 5 个训练轮次，并且仅保存得分最高的模型。每个子任务中均保存排名前三的最佳系统，最终预测通过集成[2] 排名前三的模型的所有预测结果得出。

4 结果

4.1 单语方法

评测分两个阶段进行（即已发布数据的 v2 和 v3 版本）。在 v2 阶段，评测旨在 выяс明，鉴于各语言的训练数据规模相对较小，单语方法是否有效。首先采用传统机器学习方法和深度学习方法评估单语模型，如表 3所示。v2 阶段仅评估子任务 1，以便快速进行“热身”，并找出模型与数据之间可行的方法。

总体而言，传统机器学习方法优于深度学习方法。

[2]系统通过多数投票进行集成。

语言	XLM-R(v2)			XLM-R（v3）		
	T1	T2	T3	T1	T2	T3
en	0.51	0.50	0.47	0.51	0.52	0.49
fr	0.62	0.41	0.30	0.62	0.42	0.29
de	0.64	0.63	0.34	0.65	0.65	0.37
it	0.75	0.54	0.30	0.78	0.54	0.31
po	0.50	0.59	0.15	0.51	0.59	0.13
ru	0.47	0.30	0.20	0.49	0.27	0.21

表4：基于任务相关微调策略，通过每项任务的10折交叉验证获得的 XLM-RoBERTa（XLM-R）平均值（即 T1 的宏平均 F1，以及 T2 和 T3 的微平均 F1）。

例如，将词频-逆文档频率与逻辑回归分类器简单结合，并对多数类进行欠采样，在六种语言中有四种（即 fr、it、po 和 ru）取得了最佳成绩。与此同时，采用随机森林分类器的类似组合在 v2 阶段的开发集排行榜上取得了最佳成绩。

对于深度学习方法，所有模型首先使用预训练语言模型[3] 对新闻文章进行编码，并获得 $[CLS]$ 嵌入。随后，将 $[CLS]$嵌入作为多核卷积神经网络、注意力循环神经网络和自注意力网络的输入。一方面，深度学习方法的整体性能不如传统方法。一个可能的原因是，相对于训练集较小的规模而言，深度学习模型包含的参数数量较多，从而导致过拟合。另一方面，在子任务2中，使用 RoBERTa 可以取得优于 BERT 的性能。此外，在所有子任务中，卷积神经网络模型通常都优于注意力循环神经网络和自注意力模型。

4.2 多语言方法

尽管单语方法在单种语言中取得了良好的性能，但其在所有语言中的表现并不一致。为了捕捉所有语言之间的共享信息，并训练一个能够同时处理所有语言的多语言模型，如表4所示，在 v3 阶段使用了预训练 XLM-R。

多语方法在 v2 阶段和 v3 阶段均进行了评估。在 v2 阶段，基于发布的 v2 数据集进行 10折交叉验证，以此微调 XLM-R。为应对微调过程中的不平衡数据，样本权重和

[3]单语设置中所有预训练语言模型的完整列表见附录A。

类别权重会被计算。加权随机采样器与样本权重相结合，以相对平衡训练批次。类别权重与训练损失相乘，使模型更加关注少数类。

因此，与 v2 阶段的单语方法相比，性能有了显著提升（见 XLM-R(v2) T1 列），这表明某些语言（例如 it、fr 和 en）能够从其他语言中提取有意义的特征。鉴于不同子任务中的新闻文章内容相同，但标签不同，因此在子任务 1 上得到的最佳模型会进一步在子任务 2和子任务3上进行微调。

在 v3 阶段，开发集发布的标签使 10 折交叉验证中的训练样本得以扩充。因此，除 v3 阶段的训练样本数量增加外，仍采用相同的微调方法。XLM-R（v3）的结果表明，训练样本的扩充带来了小幅性能提升。最后，将 v3 阶段表现最好的 3 个 XLM-R 模型进行集成，以生成最终预测。

4.3 消融实验

为探究所提出方法中各组成部分的有效性，我们进行了消融研究，如5表所示。结果表明，使用类别权重、样本权重和任务相关的微调，能够显著提升模型在所有子任务上的表现。

具体而言，移除类别权重后，宏平均 F1 的平均值下降了 14%。快速人工检查预测文件还发现，大多数新闻文章都被预测为“意见”，这是训练数据中占比最大的多数类。移除样本权重后，得分也会进一步下降。这可能是因为某些训练批次不包含任何少数类样本，且信息量较低的特征在训练阶段得到了更新。随后，移除任务相关策略后，XLM-R（v3）w/o td 就变成了一个未采用上述任何策略的普通 XLM-R 模型。这使得微调策略变为任务无关。结果表明，任务相关微调（即 XLM-R（v3）w/o sw）能够学习任务之间的共享信息，其表现优于任务无关微调（即 XLM-R（v3）w/o td）。

方法	T1	T2	T3
XLM-R（v3）	0.51	0.52	0.49
XLM-R（v3）不含类别权重	0.37	n/a	n/a
XLM-R（v3）不含停用词	0.25	0.31	0.20
XLM-R（v3）不含 td	0.24	0.31	0.19

表 5：微调后 XLM-R 模型的消融研究。cw、sw 和 td 分别表示类别权重、样本权重和任务相关。请注意，子任务2和子任务3中未使用类别权重。

4.4 错误分析

在错误分析中，XLM-R(v2) 仅使用带有10折交叉验证的v2数据训练集进行微调，并使用最佳模型预测v3开发数据的标签。XLM-R(v2)中每个组件的混淆矩阵如图1所示。

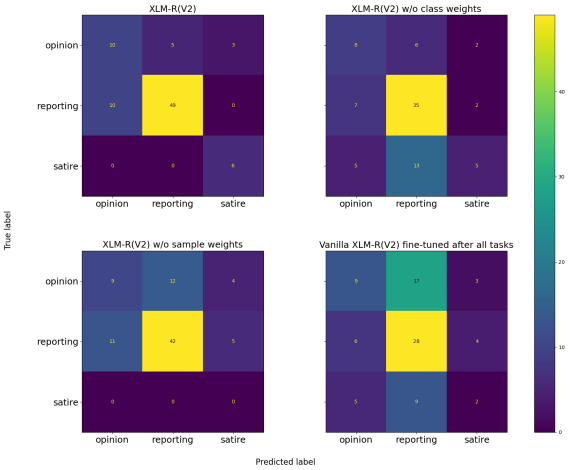

图 1：英语子任务 1 中 XLM-R(v2) 各组件的混淆矩阵。

XLM-R(v2) 在少数类“讽刺”和多数类“报告”上的正确预测数量均为最高。然而，移除类别权重和样本权重后，其性能显著下降。移除了样本权重的 XLM-R(v2) 无法正确预测少数类“讽刺”，这表明当数据高度不平衡时，标准微调无法做出正确预测。为了评估子任务之间共享信息的有效性，最后一个 XLM-R(v2) 移除了类别权重和样本权重，并直接通过三个子任务进行微调。结果，最后一个 XLM-R(v2) 对“报告”的正确预测数量最少，且“报告”与

Lan	基线			QUST（排名）		
	T1	T2	T3	T1	T2	T3
en	0.29	0.35	**0.20**	**0.51**（第10名）	**0.51**（第6名）	0.14（第22名）
fr	0.57	0.33	**0.24**	**0.62**（第9名）	**0.45**（第 11 名）	0.21（第19名）
de	**0.63**	0.49	**0.32**	**0.63**（第10名）	**0.62** (第7名)	0.15（第19名）
it	0.39	0.49	**0.40**	**0.77** (第 2 名)	**0.50**（第10名）	0.21（第19名）
po	0.49	**0.59**	**0.18**	**0.53**（第 11 名）	**0.59**（第11名）	0.10（第19名）
ru	0.40	0.23	**0.21**	**0.47** (第9名)	**0.25**（第12名）	0.10（第19名）
es	0.15	0.12	**0.25**	**0.55**（第2名）	**0.37**（第13名）	0.12（第17名）
gr	0.17	0.35	**0.08**	**0.49**（第10名）	**0.41**（第9名）	0.05（第15名）
ka	0.26	0.26	**0.14**	**0.54** (第7名)	**0.31**（第12名）	0.09（第15名）

表6：官方任务3测试集排行榜。与基线相比，较优分数以粗体显示。

“意见”之间的错误分类数量最多。然而，它在少数类“讽刺”上获得了两次正确预测，这表明共享信息能够提供有用的特征来处理少数类。

4.5 官方测试集上的结果

如表 6 所示，QUST团队在 T1 官方测试集中意大利语和西班牙语（零样本）任务上排名第二。此外，除德语和波兰语外，所提出的方法在所有语言中的表现均优于 T1 和 T2 基线。值得注意的是，所提出的方法在 T3 中表现明显较差，所有提交结果均未达到 T3 基线。可能的原因有两点：1）所有早期阶段的评估均基于 T1 的模型性能，因此结果无法充分反映 T3 的表现。2）T1 和 T2 都是文档级分类任务，而 T3 是句子级任务，因此与 T3 相比，T2 从任务相关微调中获益更多。

5 结论

本文详细介绍了用于检测新闻体裁、新闻框架构建和说服技巧的单语与多语方法。为解决不平衡数据问题，在任务早期阶段首先对多数类别进行欠采样。传统机器学习模型和深度学习模型在单语设置下进行了评估。为了学习所有语言之间的共享信息，在任务后期阶段构建了一个多语言模型。通过使用预先计算的类别权重和样本权重对XLM-R 进行微调，以应对不平-衡数据。此外，还分别评估了两种微调策略，即任务无关和任务相关策略。所提出的方法在意大利语和西班牙语（零样本）子任务 1 中是排名第二的最佳系统。在未来的工作中，次要-

类别样本可以从其他语言回译，以扩充数据量。

参考文献

Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。2020 年。大规模无监督跨语言表示学习。载于计算语言学协会第 *58* 届年会论文集，第 8440–8451 页，线上举行。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018。BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv* 预印本 *arXiv:1810.04805*。

Ye Jiang、Johann Petrak、Xingyi Song、Kalina Bontcheva 和 Diana Maynard。2019。SemEval-2019 任务 4 中的 Team Bertha von Suttner：使用 ELMo 句子表示卷积网络检测高度党派化新闻。载于第 *13* 届语义评估国际研讨会论文集，第 840–844 页。

Ye Jiang 和 Yimin Wang。2022。面向文本分类的主题感知层次化多注意力网络。国际机器学习与控制论期刊，第 1–13 页。

Jacob Devlin、Ming-Wei Chang、Kenton 和 Lee Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。载于 *NAACL-HLT* 论文集，第 4171–4186 页。

Yuri Kuratov 和 Mikhail Arkhipov。2019。针对俄语的深度双向多语言 Transformer 适配。*arXiv* 预印本 *arXiv:1905.07213*。

Joao A Leite、Diego F Silva、Kalina Bontcheva 和 Carolina Scarton。2020。面向巴西葡萄牙语的社交媒体有毒语言检测：新数据集与多语言分析。*arXiv* 预印本 *arXiv:2010.04543*。

Guillaume Lemaître、Fernando Nogueira 和 Chris-tos K. Aridas。2017。Imbalanced-learn：一个用于应对机器学习中不平衡数据集困境的 Python 工具箱。机器学习研究期刊，18(17):1–5。

Zhouhan Lin、Minwei Feng、Cicero Nogueira dos Santos、Mo Yu、Bing Xiang、Bowen Zhou 和 Yoshua Bengio。2017。一种结构化自注意力句子嵌入。载于国际学习表征会议。

Yinhan Liu、Myle Ott、NamanGoyal、JingfeiDu、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。

RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Ilya Loshchilov 和 Frank Hutter。2017。解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*。

Benjamin Minixhofer、FabianPaischer 和 Navid Rek- absaz。2022年。WECHSEL：用于单语语言模型跨语言迁移的子词嵌入的有效初始化。载于 《*2022* 年北美计算语言学协会分会会议论文集：人类语言技术》，第 3992– 4006 页，西雅图，美国。计算语言学协会。

Jakub Piskorski、Nicolas Stefanovitch、Giovanni DaSanMartino和 PreslavNakov。2023 年。Semeval-2023任务3：在多语言环境下检测在线新闻中的类别、框架构建和说服技巧。载于 第十七届语义评估国际研讨会论文集，Se- mEval 2023，多伦多，加拿大。

StefanSchweter.2020. Europeana BERT 和 ELECTRA 模型。

Marian Tietz、Thomas J. Fan、Daniel Nouri、Benjamin Bossan 和 skorch 开发者。2017。*skorch*：兼容 *scikit-learn*、封装 *PyTorch* 的神经网络库。

Tianyi Zhang、Felix Wu、Arzoo Katiyar、Kilian Q Weinberger 和 Yoav Artzi。2021。重新审视少样本 BERT 微调。发表于国际学习表征会议。

单语设置下的预训练语言模型

单语实验中使用的 Hugging Face 模型名称如下：

Lan	BERT 模型名称
en	bert-base-cased(Devlin 等,2018)
fr	dbmdz/bert-base-法语-europeana-cased(Schweter, 2020)
de	bert-base-german-cased
it	dbmdz/bert-base-italian-cased(Schweter, 2020)
po	dkleczek/bert-base-polish-uncased-v1
ru	DeepPavlov/rubert-base-cased(Kuratov 和 Arkhipov，2019)

Lan	RoBERTa 模型名称
en	roberta-base(Liu 等人，2019)
fr	benjamin/roberta-base-wechsel-french(Minixhofer 等人，2022)
de	benjamin/roberta-base-wechsel-german(Minixhofer 等人，2022)
it	flax-community/robit-roberta-base-it
po	clarin-pl/roberta-polish-kgr10
ru	blinoff/roberta-base-russian-v0