如何为文本数据增强选择“优质”样本

Xiaotian Lin[1], Nankai Lin[2], Yingwen Fu[1], Ziyu Yang[1] 和 Shengyi Jiang[1, 3(✉)]

[1] 广东外语外贸大学信息科学与技术学院，中国广东省广州市[2] 广东工业大学计算机科学与技术学院，中国广东省广州市[3] 广州多语种智能处理重点实验室，广东外语外贸大学，中国广东省广州市 jiangshengyi@163.com

摘要。基于深度学习的文本分类模型需要大量标注数据，才能获得有竞争力的性能。遗憾的是，标注大规模语料库既耗时又费力。为解决这一问题，许多研究尝试使用数据增强来扩充语料库规模。然而，数据增强可能会产生一些噪声增强样本。目前，自然语言处理领域尚无研究探索增强样本的样本选择。本文提出一种新颖的自训练选择框架，利用两个选择器从数据增强结果中筛选高质量样本。具体而言，我们首先使用基于熵的策略和模型预测来选择增强样本。考虑到上一步中一些高质量样本可能被错误过滤，我们提出从词语重叠和语义相似度两个角度将其召回。实验结果表明，我们的框架有效且简洁。

关键词：数据增强、样本选择、置信度测量、自训练

1 引言

文本分类是自然语言处理（NLP）的基础，已广泛应用于情感分析和问答等诸多领域。一般而言，训练具有较强泛化能力的分类器通常需要大量标注数据。然而，为大规模语料库添加标注既耗时又费力。为解决这一问题，人们提出了多种数据增强策略，以增加数据量并提升模型的泛化能力。

遗憾的是，充足的数据规模并不意味着模型性能良好。数据增强策略可能会破坏文本的语义信息，并引入噪声样本。这些噪声样本往往会对模型产生负面影响。因此，选择数据增强生成的高质量样本至关重要。然而，据我们所知，目前尚无研究

Xiaotian Lin、Nankai Lin 和 Yingwen Fu 是共同第一作者。他们合作完成了这篇论文，并作出了同等贡献。

专注于自然语言处理领域的数据增强样本选择。尽管其他领域已有一些样本选择方法，但这些方法侧重于采用基于置信度的策略去除噪声样本，却未考虑教师模型不确定性。

为解决上述问题，本文提出了一种基于自训练的新型样本选择框架。所提出的框架包含两个筛选步骤，并采用多种筛选策略，从数据增强生成的样本中筛选出高质量样本。具体而言，我们首先利用基于熵的策略和模型预测来选择增强样本。考虑到上述步骤中一些质量较高的样本可能会被错误过滤，我们从词汇相似度和语义相似度两个角度对这些样本进行重新召回。

总而言之，本文的主要贡献如下。
（1）据我们所知，我们首次关注自然语言处理领域中数据增强的样本选择自然语言处理领域中数据增强所面临的问题。
（2）我们提出了一种新型样本选择框架，可以从多个维度选择样本从多个维度。
(3) 我们的模型可以自动训练，且只需进行少量超参数调优工作，无需进行归一化和概率建模。
(4) 实验结果表明，我们的框架有效。此外，我们深入分析了影响数据增强样本质量的原因。

2 相关工作

2.1 文本数据增强

数据增强是应对数据匮乏问题的一个高度活跃的研究领域[1]。例如，Ma 等人[2] 提出使用回译方法扩充三个用于文本分类的中文数据集，然后通过深度分类模型对这些数据集进行训练和预测。Tamming [3] 实现并测试了三种数据增强方法：同义词替换、回译和上下文增强。此外，Wei 和 Zou [4] 提出了一种名为 EDA 的简单而有效的数据增强技术，以提升分类性能，其中包括同义词替换、随机插入、随机交换和随机删除操作。基于 EDA 技术，Karimi 等人[1] 提出了 AEDA （更简易的数据增强）技术，以进一步提升分类性能，该技术仅包括在输入序列中插入各种标点符号。

预训练语言模型（PLM）已被证明能够在不同的自然语言处理任务中带来显著提升。因此，一些研究人员也选择使用预训练语言模型进行数据增强。基于 Transformer，Kumar 等人 [6] 研究了不同类型的 PLM，并证明在文本序列前添加类别标签，是为数据增强对 PLM 进行条件化的一种简单而有效的方法。Hu 等人[7] 利用结合条件语言模型的强化学习，该方法在训练时通过将正确标签附加到输入序列来实现。

此外，在深度学习模型领域，越来越多的学者利用对抗学习方法生成增强样本。Li 等人[8]

提出了一种高质量且有效的方法（BERT-ATTACK），利用以 BERT 为代表的预训练掩码语言模型生成对抗样本 [9]。Yang 等人 [10] 提出了一种新颖的生成式数据增强技术，以实现在低资源环境下准确且稳健的学习。具体而言，他们利用预训练语言模型生成合成样本，并筛选出信息量最大且最多样化的增强样本。

2.2 样本选择

尽管数据增强在一定程度上缓解了模型训练对数据的需求，但低质量的增强样本可能会对模型性能产生负面影响。因此，筛除噪声增强样本尤为重要。目前，已有许多研究致力于样本选择。

Cao 等人 [11] 提出使用不确定性感知自训练框架（UAST）量化模型不确定性，以选择伪标签样本。Rizos 等人 [12] 提出了一种通过利用底层神经网络的不确定性估计来改进自训练的方法。他们利用蒙特卡洛（MC）丢弃，从未标注样本池中选择样本。为应对神经序列标注模型面临的标签稀缺挑战，Wang 等人 [13] 开发了自训练和元学习技术，以利用少量标签训练神经序列标注器。

目前，自然语言处理领域尚无研究探讨增强样本的样本选择。与现有的样本选择研究不同，我们更关注样本选择与数据增强策略的结合。

3 框架

如图 1 所示，我们的框架由四个模块组成：(1) 自训练模块；(2) 编码器模块；(3) 样本选择模块；(4) 样本召回模块。

我们使用 RoBERTa 作为编码器来获取文本表示。此外，我们利用蒙特卡洛 dropout 计算模型不确定性，并采用基于熵的策略，根据模型不确定性估计选择模型不确定的样本进行自训练。最后，在样本召回模块中，我们从词语重叠和语义流畅度两个角度，自动召回一些置信度较低的高质量样本。

3.1 自训练模块

自训练是一种半监督学习方法，通过扩展有标签训练集，为未标注数据生成伪标签；如今，它已广泛应用于样本选择领域。

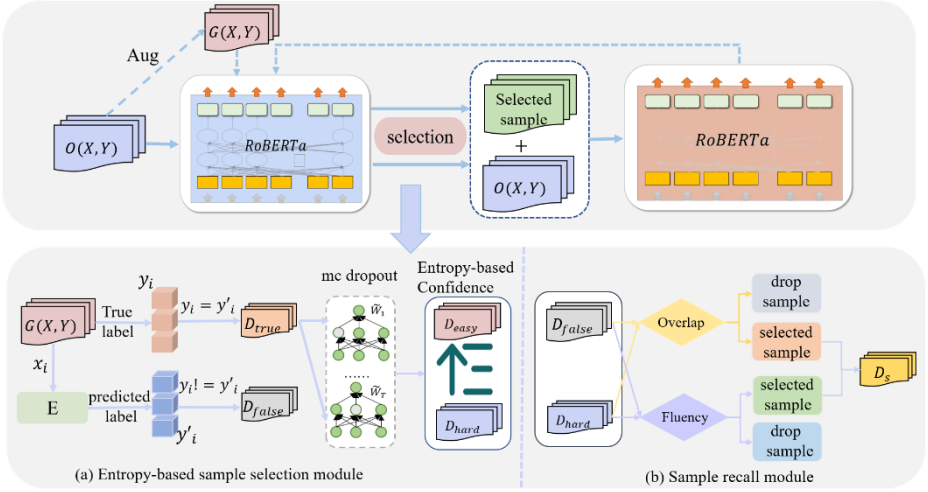

图1. 所提出的框架结构。

因此，我们采用自训练技术对数据增强生成的样本进行预测和筛选。形式化地说，给定一个包含 N samples$D_l = \{x_l^i, y_l^i\}_{i=1}^N$ 的原始语料库，以及一个包含 M samples $D_g = \{x_g^i, y_g^i\}_{i=1}^M$ 的增强语料库，该模块的实现如下：

● 步骤1：在原始语料库 D_l上使用交叉熵（CE）损失训练教师模型 $RoBERTa_T$ 。● 步骤 2：使用教师网络为增强语料库 D_g 预测伪标签 D_p 。 ● 步骤 3：使用多种筛选策略筛选增强语料库 D_g ，从而得到增强语料库的一个子集 S_g。 ● 步骤4：训练学生模型 $RoBERTa_{Stu}$ ，同时使用 S_g 和 D_l。 ● 步骤5：将当前学生模型视为教师模型，返回步骤2，并重复步骤2至4，直至模型收敛。

3.2 编码器模块

RoBERTa（稳健优化的BERT预训练） [15], 是BERT的一种变体，在预训练时仅保留掩码语言建模（MLM）任务。输入句子的首尾位置分别添加特殊标记 [CLS] 和[SEP]。对于输入句子中的每个词元，其输入表示由对应的词元嵌入、片段嵌入和位置嵌入相加构成。

本文采用 RoBERTa 作为编码器来训练教师模型和学生模型。对于分类任务，通常使用 [CLS] 的最终隐藏状态来表示句子。因此，对于第 i 个句子，其句子表示计算如下：

$$S_i = RoBERTa(a_i, b_i, c_i)$$

其中 a_i、 b_i、 c_i 是词元嵌入、segmentation, and position

3.3 样本选择模块

由于一些噪声样本是通过数据增强生成的，模型在自训练过程中往往会逐渐偏向噪声数据，从而影响模型性能。以往的研究主要采用基于置信度的方法来移除噪声样本，却未考虑教师模型的不确定性。受这些方法的启发，我们提出了一种两阶段样本选择方法来解决这一问题，该方法包括干净数据选择和置信度评估。

干净数据选择。为验证预测标签的正确性，我们引入一项二分类任务，旨在筛除预测标签错误的增强数据。我们使用教师模型预测增强样本的标签，并将预测标签与其真实标签进行比较。形式化地，给定一个增强样本 x_i^g 及其真实标签 y_i^g，我们首先使用带线性分类器的编码器预测其概率分布 $f(x_i^g, \theta)$，并获得伪标签 $\tilde{y}_i^g$。

$$f\left(x_i^g, \theta\right) = softmax\left(W x_i^g + b\right)$$
$$\tilde{y}_i^g = argmax\, f\left(x_i^g, \theta\right)$$

然后，我们将 th 和真实标签视为干净样本，并将它们放入集合 D_{clean}；其余样本则放入集合 D_{noisy}。

e samples with the same predictions

置信度评估。筛除预测错误的样本后，我们基于概率分布，采用基于熵的置信度度量来量化样本在 D_{clean} 上的置信度。为更好地评估样本置信度，我们首先利用蒙特卡洛 dropout，通过带有 dropout 层的 T 次前向传播来预测增强样本的标签。例如，给定一个增强样本 x_i，其在第 t 次 dropout p_{ic}^t 中属于类别 c 的概率如下：

$$p_{ic}^t = p(y = c|x_i) = softmax\left(\widetilde{W}_t x_i + B\right)$$

然后，根据来自 T 次传递的 p_{ic} gate 预测，并利用基于熵的置信度度量生成样本置信度，其计算方式如下：

ng to the probability distribution, we aggre

$$H(p_i) = -\frac{1}{C}\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T} p_{ic}^t \log\left(p_{ic}^t\right)$$
$$w_{easy}^i = 1 - H(p_i)$$
$$w_{hard}^i = H(p_i)$$

其中，w_{easy}^i 和 w_{hard}^i respectively donate the easy and hard 分别是第 i 个样本的置信度。最后，我们对所有样信度进行排序，得到两个集合 D_{easy} 和 D_{hard}。

3.4 样本召回模块

D_{noisy} 和 D_{hard} 中的一些高质量样本可能会被错误地过滤掉，而这些样本能够在一定程度上提升模型性能。因此，我们提出从词语重叠和语义流畅度两个角度重新召回这些样本。

词语重叠。词语重叠从词语的角度衡量文本相似度。它可以有效筛选出与原始句子含义相似的样本

。本文使用杰卡德系数计算增强样本与原始样本之间的词语重叠。假设 x_l^i 和 x_g^i 分别表示第 i 个样本中的原始样本和增强样本，则词语重叠 $J(x_l^i, x_g^i)$ 可按如下方式计算：

$$J(x) = J(x_l^i, x_g^i) = \frac{|x_l^i \cap x_g^i|}{|x_l^i \cup x_g^i|} = \frac{|x_l^i \cap x_g^i|}{|x_l^i| + |x_g^i| - |x_l^i \cap x_g^i|}$$

语义流畅度 [16]，旨在对语法错误纠正任务 [17, 18] 至关重要。因此，我们提出一种指标来评估增强数据的语义流畅度。形式上，给定一个原始样本 $x_l^i = \{w_1^i, w_2^i, \dots, w_n^i\}$，其中包含 n 个词，以及其增强样本 $x_g^i = \{u_1^i, u_2^i, \dots, u_n^i\}$，其中包含 m 个词，我们首先使用在 One-Billion[1] 语料库上训练的语言模型 ResLSTM [19] 计算它们的困惑度 $P(\cdot)$。

$$P(x_l^i) = \sqrt[n]{\prod_{j=1}^{n} \frac{1}{p(w_j^i | w_1^i w_2^i \dots w_{j-1}^i)}}$$

$$P(x_g^i) = \sqrt[m]{\prod_{j=1}^{m} \frac{1}{p(u_j^i | u_1^i u_2^i \dots u_{j-1}^i)}}$$

语义流畅度 F_i ency i for the -th sentence is defined 根据困惑度计算得出。

$$F(x) = F(x_l^i, x_g^i) = P(x_l^i) - P(x_g^i)$$

之后，为了重新获取 high-quality samples from $D_{noisy} + D_{hard}$，我们使用 D_{easy} 自动计算阈值，以减轻归一化和概率建模的工作量。具体而言，我们计算 D_{easy} 上样本的平均杰卡德系数 J_{avg} 和流畅度 F_{avg}，并将其作为阈值来筛选重新获取的样本。

$$R_{cutoff} = \{x \in D_{noisy} + D_{hard} \mid J(x) > J_{avg} \wedge F(x) < F_{avg}\}$$

4 实验

4.1 数据集

斯坦福情感树库 (SST) [20] 是最受欢迎的情感分类数据集之一，由三个部分组成：SST-1、SST-2 和 SST-5。SST-1 和 SST-2 是二分类数据集，而 SST-5 包含五种情感标签，用于细粒度情感分类任务。本文在 SST-5 上进行了大量实验。如表 1 所示，该数据集原始的训练集、验证集和测试集分别包含 8543、1100 和 2209 个样本。我们直接使用这些数据来训练和测试模型。此外，我们采用两种数据增强方法扩充训练数据。(1) 简易数据增强 (EDA) [4]。这种

[1] https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark

数据增强方法包括同义词替换 (SR)、随机插入 (RI)、随机交换 (RS) 和随机删除 (RD)。每种方法都会使原始训练集的规模增加一倍。最终，我们构建了一个规模是原始训练集 4 倍的新训练集。(2) nlpaug 工具[2]。我们从该工具中随机选取四种数据增强方法，对每个原始样本进行增强。这样可以生成另一个规模是原始训练集 4 倍的新训练集。

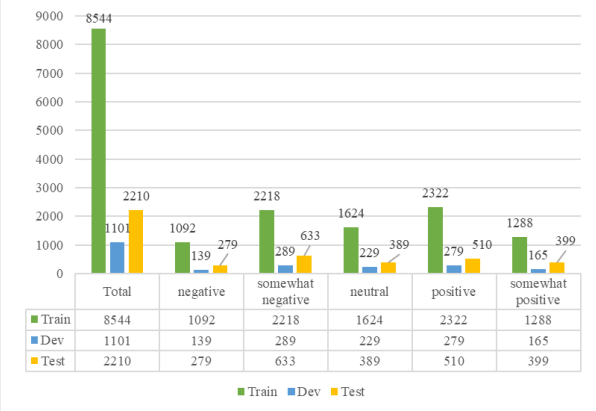

图2. 数据集分布。

4.2 实验设置

我们使用 PyTorch[3] 和 Transformers[4] 框架实现模型。在训练过程中，我们使用 Adam 作为优化器，并报告采用不同随机种子运行 3 次的结果。我们比较了不同预训练语言模型的性能，包括 RoBERTa-Base[5], BERT-base[6], XLM-RoBERTa[7] , 和 mBERT[8], , 以选择最佳基础模型。最终，我们选择 RoBERTa-Base 作为基础模型。所提出框架和基线模型的超参数如表1所示。

至于评估指标，我们使用 scikit-learn[9] 实现的准确率来评估模型，这与以往的研究一致[21][22][23][24][25][26][27][28][29]。

表1. 参数值。

模块	参数	数值
ResLSTM	嵌入维度	256
	词元长度	256
	批次大小	32

[2] https://github.com/makcedward/nlpaug[3]
https://github.com/pytorch/pytorch[4]
https://github.com/huggingface/transformers[5]
https://huggingface.co/roberta-base[6]
https://huggingface.co/bert-base-uncased[7]
https://huggingface.co/xlm-roberta-base[8]
https://huggingface.co/bert-base-multilingual-uncased[9]
https://scikit-learn.org/stable/

	训练轮次	20
	学习率	0.1
	LSTM层	2
编码器模块	最大长度	32
	学习率	5e-5
	批次大小	32
	训练轮次	15
	丢弃	0.5
	权重衰减	0.001
	优化函数	Adam
样本选择模块	结合 EDA 的熵阈值	0.2
	结合 nlpaug 的熵阈值	0.25
	T	10

5 结果与分析

5.1 主要性能

为了验证我们框架的有效性，我们将模型与多种先进的基线方法进行比较。结果如表 2 所示。与使用相同基础模型的其他模型相比，我们的模型取得了具有竞争力的结果，在 EDA 和 nlpaug 中的准确率分别达到 56.6 和 56.0。这些结果证明了我们筛选策略的有效性。

表 2. 主要性能。

模型	准确率
BERT Large [21]	55.5
BCN+ELMo [22]	54.7
byte mLSTM7 [23]	54.6
BCN+字符+CoVe [21]	53.7
BERT Base [24]	53.2
Star-Transformer [25]	53.0
RoBERTa-Base	54.2
SELFEXPLAIN [26]	54.3
MS-Transformer [27]	53.5
SentiX [28]	55.6
ACNN-BERT [29]	53.7
我们使用 EDA 的方法	**56.5**
我们使用 nlpaug 的方法	56.0

5.2 不同预训练语言模型的效果

表 3. 不同预训练语言模型的效果。

模型	准确率
RoBERTa-Base	54.2
BERT-base	53.1
mBERT	48.4
XLM-RoBERTa	51.8

不同预训练语言模型作为编码器的结果如表 3 所示。显然，RoBERTa-Base 模型的表现最佳，因此我们最终选择它作为基础模型。

5.3 不同数据增强方法的有效性

为探究不同数据增强方法的有效性，我们分别对每种数据增强方法进行实验，并将其与使用所有数据增强方法训练的模型，以及未使用任何数据增强方法训练的模型进行比较。如表 4 所示，EDA(SR)[10] 和 EDA(RS)策略似乎能略微提升模型性能。EDA(SR)策略不会改变文本的基本语义信息。此外，EDA(RS)策略仅交换词元顺序，这种变化对模型性能的影响不大 [30]。相反，EDA(RI)和 EDA(RD)策略会损害模型性能，我们推测原因在于这些策略往往会改变文本的语义信息。此外，我们可以看到，使用全部增强数据训练模型并未带来显著提升，这也表明了我们框架的有效性。

表 4. 不同数据增强方法的有效性。

模型	准确率
RoBERTa-Base	54.2
使用 EDA（完整数据）的 RoBERTa-Base	54.4
使用 nlpaug（完整数据）的 RoBERTa-Base	53.6
使用 EDA(SR) 的 RoBERTa-Base	54.8
使用 EDA(RI) 的 RoBERTa-Base	53.6
RoBERTa-Base 搭配 EDA(RS)	54.6
RoBERTa-Base 搭配 EDA(RD)	52.9

5.4 各模块的有效性

[10] EDA(·) 表示仅使用一种策略的 EDA 方法。

表 5. 各模块的有效性。

模型	增强数据数量	Acc
BERT Base	-	54.2
使用 EDA 的 BERT Base	34176	54.4
+ 基于熵的样本选择模块	584	55.5
+ 召回模块	23765	56.6

此外，我们验证了框架中各个模块的有效性。表 5 的结果表明，每个模块都有助于提升模型性能。尤其是，与 RoBERTa-Base 模型相比，加入样本选择模块后，模型的准确率提高了 1.3%。由于可靠性模块会过滤掉大量有损模型性能的噪声样本，因此它能在一定程度上提升模型性能。此外，召回模块也能带来一定的性能提升，这证明了召回模块的有效性和重要性。由于召回模块能够召回一些高质量样本，以便进一步提升性能。

5.5 不同阈值的有效性

我们探究了不同策略中阈值的最佳组合。具体而言，我们依次探究了不同模块中各项策略的最佳阈值。如图3所示，熵策略在 EDA 中的阈值为 0.25、在 nlpaug 中的阈值为 0.3 时表现最佳。

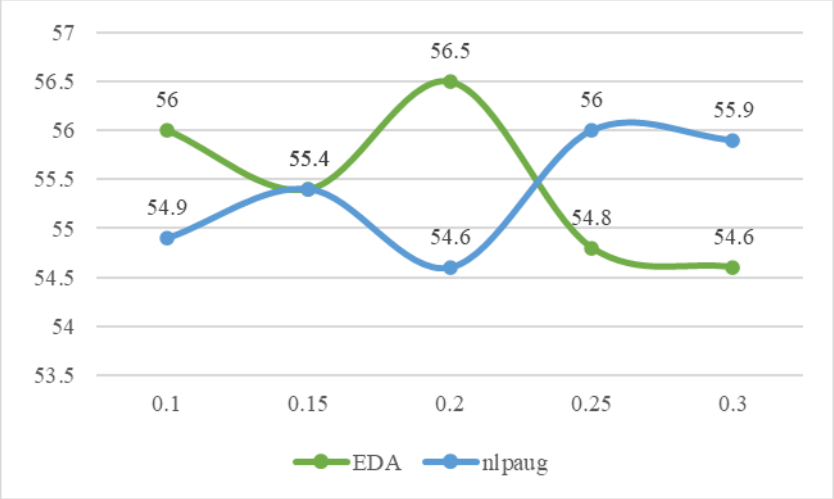

图3. 我们的框架结构。

表6. 增强数据的数量。

策略	增强数据的数量
同义词替换（SR）	5227

随机插入（RI）	6093
随机交换（RS）	6265
随机删除（RD）	6180
总计	23765

为了探究不同数据增强方法对模型性能的影响，我们以 EDA 的结果为例，统计各项操作最终保留的样本数量，以确定哪项操作生成对模型有益的增强样本的能力最弱。最终，经过两个模块的联合筛选，我们共获得 23,765 个高置信度样本，其中采用 SR策略的样本召回率最低，仅召回 5,227 个样本。召回的增强数据数量见表6。

6 结论

本文提出了一种用于数据增强技术的样本选择框架，该框架易于使用且合理有效。针对现有数据增强方法存在的问题，我们提出的框架引入了多种筛选策略，从增强样本中筛选出高质量样本，从而有效提升模型性能。未来，我们将继续提升模型性能，并将该框架应用于其他领域。

7 声明

7.1 利益冲突

作者声明，我们不存在任何与所提交的工作相关、可能构成利益冲突的商业或关联利益；本研究不涉及人类参与者和/或动物。

7.2 数据可用性

数据可应要求提供。

参考文献

1. Karimi, A., Rossi, L., Prati, A.：AEDA：一种更简便的文本分类数据增强技术。载于：计算语言学协会年会论文集：EMNLP，第2748–2754页。（2021）2. Ma, J. 和 Li, L.：利用回译进行中文文本分类的数据增强。载于：《物理学杂志：会议系列》，第1651卷（1），012-039（2020）。3. Tamming, D.：文本分类任务的数据增强。滑铁卢大学（2020）。

4. Wei, J. 和 Zou, K.：EDA：用于提升文本分类任务表现的简易数据增强技术。载于：2019年自然语言处理实证方法会议与第9届自然语言处理国际联合会议论文集，第6382–6388页。（2019）。5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. 和 Polosukhin, I.：Attention Is All You Need。神经信息处理系统进展，第30卷（2017）。6. Varun, K., Ashutosh, C. 和 Eunah, C.：利用预训练Transformer模型进行数据增强。载于：口语系统终身学习研讨会论文集，第18-26页。（2020）。7. Hu, Z., Tan, B., Salakhutdinov, R. R., Mitchell, T. M., Xing, E. P.：学习用于数据增强和加权的数据操作。载于：第33届神经信息处理系统国际会议论文集，第15764–15775页。（2021）。8. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.：BERT-ATTACK：使用BERT对BERT发起对抗攻击。载于：2020年自然语言处理实证方法会议论文集，第6193–6202页。（2020）。9. Devlin, J., Chang, M. W., Lee K., Toutanova K.：BERT：用于语言理解的深度双向Transformer预训练。载于：NAACLHLT 2019论文集，第4171–4186页。（2019）。10. Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., Bras, R.L., Wang, J.P., Bhagavatula, C., Choi, Y. 和 Downey, D.：用于常识推理的生成式数据增强。载于：计算语言学协会年会论文集：EMNLP，第1008–1025页。（2020）。11. Cao, P., Zuo, X., Chen, Y., Liu, K., Zhao, J., Bi, W.：面向事件时间关系抽取的置信度感知自训练。载于：第30届ACM信息与知识管理国际会议论文集，第2900-2904页。（2021）。12. Rizos, G., Hemker, K., Schuller, B.：以增强防范：深度学习中用于仇恨言论分类的短文本数据增强。载于：第28届ACM信息与知识管理国际会议论文集，第991-1000页。（2019）。13. Wang, Y., Mukherjee, S., Chu, H., Tu, Y., Wu, M., Gao, J., Awadallah, A. H.：面向少样本神经序列标注的自适应自训练。载于：第27届ACM SIGKDD知识发现与数据挖掘大会论文集，第1737–1747页。（2021）。14. Amini, M. R., Feofanov, V., Pauletto, L., Devijver, E., Maximov, Y.：自训练：综述。CoRR。（2022）。15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis M., Zettlemoyer, L. 和 Stoyanov, V.：RoBERTa：一种稳健优化的BERT预训练方法。CoRR，（2019）。16. Lau, J. H., Clark, A., Lappin, S.：无监督预测可接受性判断。载于：第53届计算语言学协会年会与第7届自然语言处理国际联合会议论文集，第1618–1628页。（2015）。17. Napoles, C., Sakaguchi, K., Tetreault, J.：无可比性：语法错误纠正中的无参考评估指标。载于：2016年自然语言处理实证方法会议论文集，第2109-2115页。（2016）。

18. Napoles, C., Sakaguchi, K., Tetreault, J.: JFLEG：用于语法错误纠正的流利度语料库和基准。在：计算语言学协会欧洲分会第 15 届会议论文集，第 229-234 页。（2017）。19. Kim, J., El-Khamy, M., Lee, J.: 残差长短期记忆网络：用于远场语音识别的深度循环架构设计。CoRR。（2017）。20. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C.: 基于情感树库的语义组合性递归深度模型。在：2013 年自然语言处理实证方法会议论文集，第 1631-1642 页。（2013）。21. Munikar, M., Shakya, S., Shrestha, A.: 使用 BERT 进行细粒度情感分类。在：2019 年人工智能推动商业与社会转型会议，第 1-5 页。（2019）。22. Peters, M. E., Neumann, M., Iyyer, M., Iyyer, M., Clark, C., Lee, K., Zettlemoyer, L.: 深度上下文化词表示。在：2018 年计算语言学协会北美分会：人类语言技术会议论文集，第 2227-2237 页。（2018）。23. Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., Arora, S.: 自选式嵌入：低成本且有效的语义特征向量归纳方法。在：计算语言学协会第 56 届年会论文集，第 12-22 页。（2018）。24. McCann, B., Bradbury, J., Xiong, C., Socher, R.: 从翻译中学习：上下文化词向量。在：第 31 届神经信息处理系统国际会议论文集，第 6297–6308 页。（2017）。25. Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., Zhang, Z.: Star-Transformer。在：2019 年计算语言学协会北美分会：人类语言技术会议论文集，第 1315-1325 页。（2019）。26. Rajagopal, D., Balachandran, V., Hovy, E.H., Tsvetkov, Y.: SELFEXPLAIN：一种用于神经文本分类器的自解释架构。在：2021 年自然语言处理实证方法会议论文集，第 836–850 页。计算语言学协会，线上及多米尼加共和国蓬塔卡纳（2021）。https://doi.org/10.18653/v1/2021.emnlp-main.64。27. Qi, L., Zhang, Y., Yin, Q., Liu, T.: MS-Transformer：将多种结构先验引入统一的 Transformer，以编码句子。《计算机语音与语言》。72，101304（2022）。https://doi.org/https://doi.org/10.1016/j.csl.2021.101304。28. Zhou, J., Tian, J., Wang, R., Wu, Y., Xiao, W., He, L.: SentiX：一种面向跨领域情感分析的情感感知预训练模型。在：第 28 届计算语言学国际会议论文集，第 568–579 页。国际计算语言学委员会，西班牙巴塞罗那（线上）（2020）。https://doi.org/10.18653/v1/2020.coling-main.49。29. Sadr, H., Nazari Soleimandarabi, M.: ACNN-TL：将基于注意力的卷积神经网络与迁移学习和上下文化词表示相结合，以提升情感分类性能。《超级计算期刊》。78，10149–10175（2022）。https://doi.org/10.1007/s11227-021-04208-2 30. Wang, Y., Xie, H., Fang, S., Qu, Y., Zhang, Y.: PERT：一种用于场景文本去除的渐进式基于区域的网络。CoRR。（2021）。