文章编号：2095-2163(2023)05-0197-07 中图分类号：TP391 文献标志码：A

面向主流价值观的文本质量评价研究

崔丁洁，徐 冰
（哈尔滨工业大学 计算学部，哈尔滨 150001）

摘 要：针对面向主流价值观的文本质量评价这一全新且较为复杂的任务，本文依据主流价值观对文本质量进行定义，构建了一个面向主流价值观的文本质量评价数据集。为了缓解人工标注数据的压力以及解决域内数据获取困难的问题，提出了一个基于无监督数据增强框架的文本质量评价方法。实验证明，在数据量较小时，能显著提升模型性能。为了获取更多数据，自主构建了一个大规模中文微博检索库，通过检索对数据集进行扩充。最终模型的 *F*1 值达到 86.2%，相比 BERT 提升 1.22%。

关键词：文本质量评价；主流价值观；半监督学习

Research on text quality evaluation oriented to mainstream values

CUI Dingjie, XU Bing
(Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 More and more user generated content on the network provides a new window and channel for the publicity of mainstream values. Aiming at the new and complex task of text quality evaluation oriented to mainstream values, this paper defines text quality according to mainstream values, and constructs a text quality evaluation data set oriented to mainstream values. In order to alleviate the pressure of manually labeling data and solve the problem of difficult data acquisition in the domain, this paper proposes a text quality evaluation method based on unsupervised data enhancement framework. Experiments show that the performance of the model can be significantly improved when the amount of data is small. In order to obtain more data, we independently built a large-scale Chinese microblog retrieval database to expand the data set through retrieval. The *F*1 value of the final model reached 86.2%, which is 1.22% higher than BERT.

【Key words】 text quality evaluation; mainstream values; semi-supervised learning

0 引 言

随着互联网的发展，网络用户的沟通方式发生了明显变化，越来越多的用户喜欢通过网络论坛、博客、微博、社交网站等网络平台浏览、发布和转发消息，以此与其他用户进行交流。网络上出现了越来越多用户生成的内容，逐渐形成了草根创作、广泛参与、多元互动的网络传播新局面。这不仅拓展了文化产品生产、传播的深度和广度，也为主流价值观传播提供了新的路径。

主流价值观是国家文化软实力的重要体现，其传播的广度和践行的深度直接影响着国家意识形态安全和社会稳定。因此，传播主流价值观是中国各大媒体需要承担的责任与义务。

然而，用户生成内容良莠不齐、信息过载等现象，为主流价值观的传播带来了挑战。如何承担社会责任，传播积极向上的主流价值观，营造良好的网络舆论氛围，成为国家和社交媒体平台共同关注的问题。

面向主流价值观的文本质量评价任务需要对文本质量从正能量、主流价值观等角度进行定义。将其定义为一个五分类问题，即将面向主流价值观的文本质量划分为 1～5 个等级。这一研究和文本的情感分类存在差别，积极的情感并不一定代表正能量的价值观。如：某明星发帖称‘日本血统真的好酷，穿和服走在雪里感觉好高贵！’，这一帖子表达

基金项目：国家重点研发计划(2020YFB1406902)。

作者简介：崔丁洁(1998-)，女，硕士研究生，主要研究方向：自然语言处理、文本质量评价；徐 冰(1975-)，女，博士，副教授，主要研究方向：自然语言处理、多模态情感分析。

通讯作者：徐 冰 Emial：hitxb@ hit.edu.cn

收稿日期：2022-06-03

的情感是积极的,但却明显不符合‘爱国’这一社会主义核心价值观。

庞大的用户生成内容是新时代的产物,但也承担着一定的社会责任,以正能量作品暖人心、聚民心。通过主流价值观来驾驭算法,减少泛娱乐化、低俗类内容传播,增加符合社会主义核心价值观的内容,让算法服务于主流价值导向。

1 相关工作

面向主流价值观的文本质量评价是一个全新的任务,同时也是一个富有新时代中国特色的任务。除了缺乏统一技术框架之外,如何获取大规模的域内数据、如何标注数据,以及如何利用未标注数据也是文本质量评价任务一大难点。因此,这一任务主要与文本增强和半监督学习两个研究方向存在重合。

1.1 文本增强研究现状

文本增强主要分为无条件增强和条件增强两种方法。

1.1.1 无条件增强方法

由于不需要强制引入标签信息,无条件增强方法既可以对标注数据进行增强,又可以针对无标注数据进行增强。主要包括词汇/短语替换、随机噪音注入和混合交叉方法。在对标注数据进行增强后,不会改变数据的标签,但可能会造成文本主旨发生变化,带来一定的噪音。

对于词汇/短语的替换方法:文献[1]提出基于词典从文本中选择词汇或短语进行同义词替换;文献[2]基于词向量在嵌入空间中找寻相邻词汇进行替换;文献[3]根据 TF-IDF 分值,对非核心词进行替换。

对于随机噪音注入方法:文献[3]根据 Uni-gram 词频分布进行采样,从而随机插入一个词汇;文献[1]除了进行同义词替换外,同时采用上述随机插入词汇、随机交换词汇或交换句子、随机删除词汇或句子等随机注入噪音。文献[4]提出了一种应用于图像领域的表示增强方法(Mixup)。借鉴 Mixup 的思想,文献[5]提出了 wordMixup 和 sentMixup,将词向量和句向量进行混合;文献[6]利用交叉增强方法将相同极性的文本进行交叉。

此外,回译也是一种应用非常广泛的无条件增强方法。该方法基于机器翻译技术,文献[3]中就采用了回译技术进行数据增强。此外,对抗训练方法对模型鲁棒性的提升也是基于数据增强原理的。但是不同于 CV 领域 GAN 生成对抗进行数据增强[7],NLP 中通常在词向量上添加扰动并进行对抗训练。

1.1.2 条件增强方法

条件增强方法需要强制引入“文本标签”信息到模型中,再产生数据。随着 BERT 等预训练语言模型在 NLP 领域取得巨大成功,近来许多研究者对预训练语言模型用做文本增强进行了有益尝试。

文献[8]利用条件变分自编码模型进行增强。文献[9]基于 LSTM 进行双向语言模型预训练,将标签信息融入网络结构进行微调,使替换生成的词汇与标签信息兼容一致。在此基础之上,文献[10]基于 BERT 进行微调,将段嵌入转换为融入标签指示的标签嵌入。文献[11]基于 GPT-2 将标签信息与原始文本拼接,当作训练数据进行微调,同时采用一个判别器,对生成数据进行了过滤降噪。

1.2 半监督学习研究现状

半监督学习方法是指利用少量标注数据和大量无标注数据进行学习。相关研究主要着力于如何针对未标注数据构建无监督信号,与监督学习联合建模。简单来说,就是如何在损失函数中添加针对未标注数据相关的正则项,使模型能够充分利用大量的未标注数据不断迭代,最终增强泛化性能。半监督学习方法主要有熵最小化和一致性正则两种方法。

文献[12]提出 Π-Model 和 时间集成(Temporal Ensembling)。Π-Model 对无标注数据输入进行两次不同的随机数据增强,并通过不同 dropout 输出得到结果,引入一致性正则到损失函数中。时间集成采用时序融合模型,避免同一个训练步进行两次前向计算,从而提高训练速度。文献[13]提出的 Mean Teacher 模型认为采用在训练步骤上的平均模型会比直接使用单一模型权重更精确,于是对时间集成方法进行改进,对模型参数而不是预测结果进行平均。文献[14]提出的虚拟对抗训练(Virtual Adversarial Training,VAT)仍然采用一致性正则,采取对抗训练的方式添加噪音,不同于传统的有监督学习下的对抗训练,其没有标签信息,而是构建一个虚拟标签,并根据这个虚拟标签计算对抗扰动方向。Google 在文献[3]中提出了无监督数据增强方法(Unsupervised Data Augmentation, UDA),也采用一致性正则,同时结合了熵最小化正则:对无监督信号构建人工标签,使其趋近于 One-Hot 分布。此外,还直接计算了熵损失。将人工标签与增强后的预测

标签共同构建一致性正则，并计算损失时采用基于置信度的训练信号退火(TSA)方法防止对标注数据过拟合。

MixMatch[15]同样来自 Google，与 UDA 类似，同样结合了熵最小化和一致性正则。对标注数据进行一次增强，对于未标注数据作 K 次弱增强输入模型得到 average 后的概率。并将无标注数据得到的人工标签与标注数据混合在一起并进行 MixUp[16]操作，进而得到增强后的无标注数据以及标注数据。ReMixMatch[17]是 MixMatch 原作者对自己工作的改进，一方面进行了分布对齐，另一方面，引入强增强，将弱增强后的人工标签与强增强后的预测标签共同构建一致性正则。FixMatch[18]结合了 UDA 和 ReMixMatch，舍弃了 sharpen 操作和 UDA 的训练信号退火、ReMixMatch 的分布对齐和旋转损失等，直接利用 Pseudo-Label 构建人工标签。

以上方法大多引入了一致性正则，其关键在于如何注入噪声，一个好的模型对于输入扰动的任何细微变化也都应具有鲁棒性。所以半监督学习经常和文本增强方法结合。半监督学习方法能充分挖掘未标注数据中潜在的价值，最终增强泛化性能。在少样本场景下甚至可以比肩充分样本下的监督学习模型性能，而在充分样本场景下，性能仍然继续提升。

2　基于无监督数据增强框架的文本质量评价方法

2.1　模型结构

模型的整体框架来源于 UDA，其结构如图 1 所示：

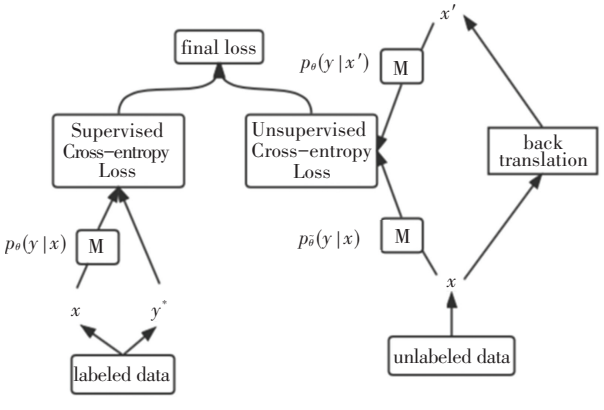

图 1　UDA 模型结构

Fig. 1　UDA model structure

图中：M 表示一个模型，可以在给定 x 的条件下预测 y 的分布。本文采用 BERT-base。

UDA 模型的输入包括有标签数据和无标签数据。对于带有标签的数据，模型 M 可以得到其预测的标签分布；对于无标签数据，采用反向翻译方法进行数据增强，x' 表示经数据增强的无标签数据。

模型的总损失 = 标签数据的交叉熵损失(Cross-entropy loss)(有监督) + λ^* 无标签数据的一致性损失(无监督)。总损失公式如式(1)：

$$\min_\theta J(\theta) = -\sum y^* \log p_\theta(y \mid x) + \lambda E_{x\in U} E_{x'\sim q(x'\mid x)}[D_{KL}(p_{\tilde{\theta}}(y \mid x) \parallel p_\theta(y \mid x'))] \tag{1}$$

其中，$q(x' \mid x)$ 表示数据增强变换，x' 由 x 经数据增强得到；θ 是模型参数；$\tilde{\theta}$ 是 θ 的复制。

由于实验数据集存在严重的不平衡问题，在实验中采用 Focal loss 代替上文中的 Cross-entropy loss，Focal loss 通过改变正类、负类的权重，使其能应用于不平衡的分类中，如式(2)：

$$L_{Focal} = -\sum \alpha_c (1 - p_\theta(y \mid x)_c)^\gamma \log p_\theta(y \mid x)_c \tag{2}$$

其中，α_c 表示第 c 类样本的权重，$p_\theta(y \mid x)_c$ 表示第 c 类样本的概率值。

2.2　数据集构建

2.2.1　获取数据集

基于 scrapy 框架，自主开发爬虫工具，在人民网强国论坛板块下爬取1 887条评论数据，部分评论数据见表 1。

表 1　人民网数据集示例

Tab. 1　Examples of people's daily online dataset

序号	评论
1	赞成！
2	“春节是民俗活动和非遗实践最集中的时期，要让年味更浓，让生活更美，让乡愁得到慰藉。”
3	学习文章，感悟：厚重！
4	“乡愁”依然、自然、天然、释然是您……对一山一水，一方水土的改良与成全！预祝您：阖家欢喜，心身健康，家庭幸福，节日愉快！

由于爬虫获取的公开数据都是符合主流价值观的，而本文的研究工作需要获取反例，即不符合主流价值观的数据。经调研，采用 SemEval 2019 Task 6 攻击性语言检测数据集(Offensive Language Identification Dataset，OLID)。该数据集收集了 14 120条推特，并对有无攻击性进行了标注。部分数据见表 2，数据集统计信息见表 3。

表 2 OLID 数据集示例

Tab. 2 Examples of OLID datasets

序号	Twitter	lable
1	And this from the clown that should be in prison?	OFF
2	Go home you're drunk!!!	OFF
3	You are correct.	NOT
4	I'M SO FUCKING READY	NOT

注:表中 OFF 表示有攻击性,NOT 表示无攻击性。

表 3 OLID 数据集统计信息

Tab. 3 OLID dataset statistics

label	样本数
OFF	4 660
NOT	9 460

从 OLID 数据集中随机选择 2 500 条攻击性数据与 613 条非攻击性数据,并将其翻译成中文,再结合爬取的人民网评论,构成最终数据集。数据集统计信息见表 4。

表 4 最终数据集统计信息

Tab. 4 Final dataset statistics

	正能量		负能量
来源	人民网	OLID 数据集	OLID 数据集
数量	1 887	613	2 500
总数	2 500		2 500

注:正能量表示符合主流价值观,负能量表示不符合主流价值观

2.2.2 数据集的人工标注

数据标注的标准见表 5。

表 5 数据集标注标准

Tab. 5 Data set labeling standards

label	等级	描述	例子
1	非常差	违禁内容(包括色情、暴力、涉政、恐怖、辱骂、歧视等)	(1)他疯了! (2)干得好,特朗普总统先生!!
2	差	低质量内容(包括消极言论、广告等)	(3)真可怜! (4)现在自由主义者快乐!
3	一般	客观描述	(5)少量酒精有兴奋作用 (6)阳历也有闰月,闰月年 2 月份 29 天。
4	好	符合主流价值观但观点表述简单	(7)实在是高! (8)很抱歉听到你失去朋友的消息
5	非常好	符合主流价值观,且观点表述有理有据。	(9)防风险,保安全依然是监管的重点,加大违规违法的处罚力度。

从表中可以看出,本文面向主流价值观的文本质量评价研究和情感分析存在明显差别。如:例(2)、例(4)虽然表达了积极的情感,但却是负能量的价值观;例(8)虽然表达了消极情感,但却是正能量的价值观。

从收集的数据中选取 585 条数据作为种子进行人工标注,标注后的初始数据分布见表 6。

表 6 初始数据集统计信息

Tab. 6 Initial dataset statistics

标签	1	2	3	4	5	总计
数量	76	99	74	197	139	585

为了统一不同人标注带来的主观性和误差,采用十折交叉验证进行数据纠错。纠错后的数据分布见表 7。

表 7 纠错后数据集统计信息

Tab. 7 Dataset statistics after error correction

标签	1	2	3	4	5	总计
数量	73	97	41	223	137	585
比例/%	12.8	17.0	17.2	39.1	24.0	

2.2.3 基于自训练的数据集自动标注

利用以上人工标注数据作为训练集,基于 self-training 对其余数据进行自动标注。实现过程如下:

Algorithm 1 self-training

1: Initialize:

2: 初始的有标签数据集作为初始的训练集 $(X_{train}, y_{train}) = (X_l, y_l)$

3: while 还有无标签样本 do

4: 利用 (X_{train}, y_{train}) 训练分类器 C_{int}

5: 利用 C_{int} ,对无标签数据集 X_u 中的样本进行分类

6: 以某一置信度阈值选出最有把握的样本 (X_{conf}, y_{conf})

7: 从 X_u 中去掉 (X_{conf}, y_{conf})

8: 将 (X_{conf}, y_{conf}) 加入到有标签数据集中,$(X_{train}, y_{train}) \leftarrow (X_l, y_l) \cup (X_{conf}, y_{conf})$

9: End while

根据观察,选取 0.7 作为置信度阈值,基于

BERT 进行数据迭代标注。针对数据集严重不平衡的问题，分别采用 Focal loss 和重采样方案，将 3 个模型标注结果有差异的并集，由人工进行再标注，最终得到的数据集分布见表 8。

表 8　最终数据集统计信息

Tab. 8　Final dataset statistics

label	1	2	3	4	5	总计
数量	940	1 207	109	1 412	728	4 396

2.3　基于检索的数据集扩充方法

半监督学习方法需要获取大规模的域内数据。然而，在许多场景下收集大规模域内数据非常困难。为了解决这一问题，采用检索的方法进行域内数据的扩充。

利用句子编码器对数据集中的句子进行编码，得到其向量表示，将每一条数据的向量表示作为检索向量，在大规模语料库中进行检索，以得到相似的句子。为保证检索到的句子可以作为域内数据，从而减少通用语料对下游特定任务的噪声干扰，每次检索只抽取 Top－K 个句子，并且对抽取的句子需满足能取得较高的置信度。

3　实验结果与分析

3.1　数据集

采用本文 2.2 节构建的数据集，其中训练集包含3 956条数据，测试集包含 441 条数据。另外，为了构建大规模检索库，本文收集了 9 个来源于微博的数据集，其中包括公开的数据集 Weibosenti100k，以及来自人民网数据平台的数据集：新浪微博数据集_凤凰周刊_202110－11、新浪微博数据集_头条新闻_202110－11、新浪微博数据集_环球时报_202110－11 等。将以上数据集进行清洗和去重，共获得 1 905 039条数据，采用上述数据扩充方法，共检索得到6 331条数据。部分数据见表 9。

从表中例子可以看出，检索到的扩充数据和原数据在语义上存在明显的相似性。如：第三条，原数据和扩充数据都和抗疫相关。另外，扩充得到的数据和原数据的主流价值观质量标签也是相似的。于是，除了无监督的数据增强方法，本文也尝试将 query 的标签赋予检索得到的数据，进行有监督的数据增强。

表 9　检索得到的部分数据及其匹配分数

Tab. 9　Some of the retrieved data and their matching scores

原数据	扩充数据	匹配分数
加油	加油心心心心心	1
祝福祖国	祝祖国繁荣昌盛心国泰民安	0.917 4
随着近期疫情的复杂变化，针对这一形势防控措施升级，希望人们减少流动，提倡大家就地过春节；不图聚会热闹，全家健康快乐又轻松，不给家人增添麻烦；这样一来，让在外上班	每次看到这种逃避防疫检查的行为，都会想到抗疫情况最急迫的时候，大家都自觉居家不外出，积极配合抗疫工作。随着政府的各种措施和有效控制，现在疫情已经没有之前那么严重，但情况依然严峻。	0.911 8
这些人真是疯了！自由主义者多年来一直在攻击白人，而且情况越来越糟。他们所推动的道路将造成另一场内战，而这场战争不会有复苏，他们太愚蠢了，看不到这一点。	好多人的政治常识真的少的可怕，比如美国搞中国是因为这届外交部工作不行，这种来自于学历不低的成年人的言论我都听到过。可偏偏还就这些人不喜欢现行制度，向往西方民主。你说有这些人存在，怎么敢完全放开言论自由	0.920 6

3.2　实验设置

3.2.1　面向主流价值观的文本质量评价实验设置

回归层的 dropout rate 为 0.1。训练过程采用 AdaGrad，初始学习率为1^{-10}，在训练过程中预热学习率以加速模型收敛。Focal loss 中，将 γ 值设置为 2。由于显卡内存所限，带标签数据的 batch size 取 16 或者 32 中较好的结果，无标签数据的 batch size 随两种数据的比例变化而变化。实验结果通过准确率（*Accuracy*）、精确率（*Precision*）、召回率（*Recall*）、*F*1 值进行评估。

3.2.2　数据集扩充实验设置

由于构建的检索库较大，直接检索非常耗时，于是数据集扩充实验基于 ANYQ 框架进行。对 ANYQ 框架源代码进行改写，只保留问题分析和检索模块。以全部训练集作为 query，对 query 和检索库均基于百度开源的 LAC2 分词工具进行分词，对检索库添加基于 PaddleSimAdapter 的语义表示模型，配置 SimNet 语义检索。每次检索只抽取 Top－10 个句

子,同时满足置信度>0.7。

3.3 结果分析

为了验证本文提出的数据增强方法的性能,选择如下几种模型并设计了相应内容进行对比实验,实验结果见表 10。

表 10 主要实验结果对比

Tab. 10 Main results

模型	*Accuracy*	*Precision*	*Recall*	*F*1
BERT	0.877 3	0.863 8	**0.850 7**	0.856 5
BERT_DA	0.829 5	0.816 3	0.829 1	0.822 2
BERT+无标签数据	**0.884 1**	**0.909 0**	0.818 3	0.849 8
BERT_UDA	0.881 8	0.903 7	0.837 5	**0.862 0**

注:表中黑体表示最高值,下划线表示高于 BERT(不进行数据增强)的结果。

其中:BERT 模型仅使用带标签的训练集,将文本作为 BERT 的输入,将[CLS]对应位置的输出作为评论表示输入分类层中,优化 Focal loss 损失。BERT_DA 模型在基于检索的数据集扩充方法中,将每一条训练数据作为查询条件进行检索,将 query 的标签赋予检索得到的数据,从而进行有监督的数据增强。BERT+无标签数据模型增加无标签训练集(6 331 条)作为输入,并对无标签数据应用熵最小化损失,从而达到数据增强效果。BERT_UDA 模型即本文提出的方法。

由表中数据可见:在本文构建并标注的训练集上,其 BERT_UDA 方法在准确率、精确率、*F*1 值 3 个指标上都超过了 BERT。与一般的无监督数据增强方法(BERT+无标签数据)相比,BERT_UDA 在 *F*1 值上提高 1.22%,表明了本文采用数据增强方法的有效性。其次,虽然 BERT_UDA 相比 BERT 的 *F*1 值提升了 0.55%,但效果并不明显,可能是构建的检索库不够大所致(获取到的无监督扩充数据数量仅为原数据的 1.6 倍)。BERT+无标签数据与 BERT 相比,准确率有所提升,但 *F*1 值却有所下降;BERT_DA 相比 BERT 在各项评估指标上都有明显下降。究其原因可能是因为引入了大量噪音,这表明数据增强也有可能会降低模型的性能。

3.4 无监督损失函数权重分析

为了验证总损失中无监督损失函数的权重 λ 对实验结果的影响,选取 $\lambda = \{0, 0.5, 0.7, 1\}$ 进行实验。实验结果见表 11。

实验结果表明,当 λ 取 0.5 时, *F*1 值最高。

表 11 无监督损失函数权重 λ 的影响

Tab. 11 Influence of unsupervised loss function weight λ

无监督损失函数权重 /λ	*F*1
1	0.851 9
0.7	0.859 0
0.5	**0.862 0**
0	0.856 5

3.5 无监督数据增强框架有效性分析

在不使用扩充数据集的情况下,仅在训练集中随机选取一部分数据作为标注数据,其余作为未标注数据。在损失函数中,将无标签数据的一致性损失函数权重设置为 1。实验结果见表 12。

由表中数据可见:BERT 和 BERT_UDA 的分类效果随着带标签数据比例的增大而提高,且在任何比例的带标签数据中,BERT_UDA 的表现均优于 BERT。由此表明,BERT_UDA 可以从无标签数据中学到知识,尤其在仅仅使用 10%的训练数据(396 条)时,BERT_UDA 的提升达到 4.77%。

表 12 *F*1 值实验结果

Tab. 12 Results of *F*1 value

模型	训练集中带标签数据的比例							
	10%	20%	30%	40%	50%	60%	70%	80%
BERT	0.685 5	0.741 3	0.752 5	0.763 1	0.804 8	0.823 8	0.824 5	0.842 3
BERT_UDA	0.733 2	0.741 4	0.789 8	0.793 8	0.815 6	0.837 2	0.839 5	0.849 8

3.6 基于检索的数据集扩充方法有效性分析

分别采用原数据和检索得到的数据基于 BERT_UDA进行实验,对数据集扩充的有效性进行分析,各实验所用模型和数据集见表 13。

表 13 基于检索的数据集扩充方法有效性分析实验

Tab. 13 Effectiveness analysis experiment of retrieval-based dataset expansion method

描述	训练集	模型
BERT	从人工标注的数据集中随机抽取约 400 条作为带标签数据	BERT
BERT_UDA-indomain	+从人工标注的数据集中随机抽取约3 600条作为无标签数据	BERT_UDA
BERT_UDA-retrieval	+从扩充数据集中随机抽取约3 600条作为无标签数据	BERT_UDA