利用半监督生成对抗网络检测孟加拉语虚假评论

Md. Tanvir Rouf Shawon[§]、G. M. Shahariar[§]、Faisal Muhammad Shah、Mohammad Shafiul Alam 和 Md. Shahriar Mahbub Ahsanullah 科学与技术大学，孟加拉国达卡 电子邮箱：{shawontanvir, shahariar_shibli, faisal, shafiul, shahriar}.cse@aust.edu

摘要—本文探讨了半监督生成对抗网络（GAN）在微调预训练语言模型方面的潜力，旨在利用少量标注数据，将孟加拉语虚假评论与真实评论分类。随着社交媒体和电子商务的发展，检测虚假或欺骗性评论变得日益重要，有助于保护消费者免受虚假信息误导。任何机器学习模型在识别虚假评论时都会面临困难，尤其是对于孟加拉语这类低资源语言。我们证明，所提出的半监督 GAN-LM架构（在预训练语言模型之上构建生成对抗网络） 是一种可行的孟加拉语虚假评论分类方案。实验结果表明，即使只有 1024 条标注样本，采用半监督生成对抗网络（SSGAN）的 BanglaBERT 仍取得了 83.59% 的准确率和 84.89% 的 F1分数；就准确率而言，其表现分别比其他预训练语言模型——BanglaBERT Generator、Bangla BERT Base 和 Bangla-Electra 高出近 3% 、 4% 和 10%。实验使用了一个人工标注的美食评论数据集，其中包含从多个社交媒体群组收集的共 6014 条真实和虚假评论。对于因缺少标注数据而难以识别虚假评论或解决其他分类问题的研究人员，所提方法或许能提供一种解决方案。

关键词— 生成对抗网络、BERT、SS-GAN、虚假评论、孟加拉语评论、分类、孟加拉语 BERT、ELECTRA

I. 引言

社交媒体平台使任何人都可以在任何时候，对任何服务、商品、品牌或企业提交不受限制的评论或批评。由于没有任何约束，人们利用社交媒体不公正地宣传特定的服务、产品或商品，或贬低竞争对手。客户在阅读社交媒体上的评论后，决定是否购买某种特定产品，或改变自己的想法。这些评论提供了评论者亲身经历的第一手视角，并且是任何潜在客户的宝贵参考。例如，在决定是否去某家餐厅之前，人们会阅读之前用餐者对菜单项目或服务的看法。他们会根据评论中的反馈来决定是否光顾这家餐厅。评论被认为是传达对产品和服务的正面或负面意见的真实方式，因此任何伪造评论的尝试

[§]作者贡献相同。

在评论中包含虚假或欺骗性信息被视为不道德。虚假评论检测具有挑战性，因为评论往往可以以看似真实的方式撰写，使其难以与真实评论区分开来。此外，虚假评论通常是为了操纵公众舆论而创作的，这使得准确检测它们更加困难。更进一步，评论中使用的语言可能因用户的背景和文化语境而存在很大差异，这使虚假评论检测任务变得更加复杂。已有大量研究致力于创建用于虚假评论检测的计算模型，即传统的基于机器学习的方法 [1]，基于深度学习的方法 [2], 基于生成对抗网络的方法 [3] 以及基于 Transformer 的方法 [4]。近期研究大多集中于监督学习技术，而这类技术需要标注数据；在识别虚假在线评论方面，标注数据十分稀缺，因为人工准确标注欺诈性评论极其困难，甚至几乎不可能。虽然已有一些针对英语、波斯语和罗马乌尔都语等语言的研究，但针对孟加拉语的研究尚不存在。因此，为了利用人工收集并标注的孟加拉语虚假评论文本数据集识别虚假评论，我们在本研究中提出了一种包含五个预训练语言模型的半监督生成对抗网络架构。

II. 相关工作

Croce 等人 [5] 提出的半监督 GAN-BERT 架构，最初由 Raihan 等人 [6] 用于在孟加拉语上微调预训练的 BERT，以便利用少量带标签的样例对孟加拉语文本进行分类。他们评估了 GAN-Bangla-BERT 在两项孟加拉语下游任务（仇恨言论和假新闻检测）上的表现，并与 Bangla-Electra 和 Bangla BERT Base 进行了比较。在多项下游任务中，研究人员采用了以针对特定任务和语言预训练的语言模型为编码器的半监督 GAN-BERT 架构。Ta 等人 [7]使用 GAN-BERT 识别西班牙语社交媒体帖子中的暴力和辱骂内容。然而，在将噪声向量输入生成器网络之前，他们使用了不同于原始 SS-GAN 架构的随机率对其进行了调整。Zaharia 等人 [8] 将罗马尼亚语 BERT 与 SS-GAN 结合，用于罗马尼亚语方言识别；Yusuf 等人 [9] 则使用 SS-GAN 微调 ARBERT 和 MARBERT，以识别阿拉伯语方言。Colo´n-Ruiz 和 Segura- Bedmar [10] 采用了一个 BERT 模型，后接娱乐主管

发表于并收录于“2023年第五届自然语言处理国际会议”。© 2023 IEEE。允许将本材料用于个人用途。对于任何其他用途，无论是在当前还是未来的媒体中，均须获得 IEEE 的许可，包括为广告或促销目的重印或再版本材料、创作新的合集、转售或重新分发至服务器或列表，或在其他作品中重用本作品中受版权保护的任何部分。

用于药物评论情感分析的双向长短期记忆网络（Bi-LSTM）与SS-生成对抗网络。Auti等人[11] 采用了BioBERT与SS-生成对抗网络，用于将药品合规文本和不合规文本进行分类，且表现最佳。最近，Athirai等人[12] 提出了一种对抗训练策略，利用生成式预训练Transformer 2（GPT-2），通过少量有标签数据和大量无标签数据来识别虚假评论；该模型还能够以适当的困惑度生成合成的虚假和非虚假评论，从而为训练提供更多有标签数据。

III. 背景研究

A. 半监督生成对抗网络

生成建模是一种无监督学习问题，也称为“对抗学习”。在这一过程中，模型会自动识别并学习输入数据中的模式或分布，并可能生成与原始数据相同的新合成样本。生成对抗网络（GAN）[13] 是一种用于训练生成模型的架构。半监督生成对抗网络（SS-GAN） [14] 是对 GAN 框架的改进，可同时利用有标签和无标签数据。GAN 的训练过程涉及生成器模型和判别器模型。生成器模型会生成一批样本（伪造样本）。随后，将生成的样本与原始样本提供给判别器模型，由其判定样本是真实的（原始数据）还是伪造的（生成数据）。在 SS-GAN 中，判别器不仅判断样本是否由模型生成，还会为每个样本分配一个目标类别。生成器和判别器模型在对抗环境中进行训练，但判别器采用（n + 1）类目标进行训练。在将样本进行二元表示分类（真实或伪造）并归入（n+1）th 类的同时，判别器还会将真实样本归入（1, ..., n）类中的一个类别。判别器使用有标签数据进行训练，而无标签输入数据以及生成器生成的样本则用于帮助学习数据的表示。无标签数据越多，模型就越能捕捉数据表示。尽管 SS-GAN 通常用于计算机视觉领域，Croce 等人 [5]指出，它们也可用于在文本特征上微调语言模型，这种方法称为“GAN-BERT”。作者在 BERT [15]，之上采用了 SS-GAN 架构，从而能够将任何预训练语言模型用作 SS-GAN 架构中的编码器进行微调，因此本研究将该架构称为“GAN-LM”。

B. 孟加拉语预训练语言模型

尽管为孟加拉语等资源匮乏的语言构建专用语言模型颇具挑战，但近年来，一些预训练语言模型凭借在多种下游任务上取得的最先进（SOTA）结果而备受关注，下面将对此进行简要介绍。

1) 基于 BERT：来自 Transformer 的双向编码器表示（BERT）[15] 基于一个深度学习模型构建。其中每个输入和输出元素都相互关联，元素之间的权重会根据这种关系动态确定。BERT 的新颖之处在于其双向训练能力：语言模型不再只学习某个词之前或之后的词，而是根据该词周围的词来学习其上下文。Bangla BERT Base[1] 是一个基于掩码语言建模的预训练孟加拉语模型，其定义与 BERT 中所述一致。

表 I：孟加拉语语言模型的配置

预训练 语言模型	**Type**	参数	嵌入 **Size**
Bangla BERT Base	BERT-base	110M	768
BanglaBERT	ELECTRA-base	110M	768
BanglaBERT 生成器	ELECTRA-base	34M	768
sahajBERT	ALBERT-large	18M	128
Bangla-Electra	ELECTRA-small	14M	128

2）基于 ELECTRA：高效学习能够准确分类词元替换的编码器（ELECTRA）的预训练任务 [16] 基于识别输入序列中被替换的词元。训练判别器模型以识别损坏序列中哪些词元被替换，同时训练生成器模型，以预测被掩码词元的原始词元。该设置与生成对抗网络训练系统相似，但其并非对抗性的，因为生成器并未经过训练来尝试欺骗判别器。Bangla-Electra [2] 是一个以替换词元检测（RTD）目标进行预训练的 ELECTRA模型，BanglaBERT [17] 是一个 ELECTRA判别器模型，而BanglaBERT Generator [17] 是一个以孟加拉语语料库中大量数据上的掩码语言建模（MLM）目标进行预训练的 ELECTRA 生成器模型

3) 基于 ALBERT：轻量级BERT（ALBERT）[18] 通过采用原始 Transformer 的相同编码器段架构，并做出三项关键改进，证明了更好的语言模型并不总是意味着更大的模型：因式分解的嵌入参数、跨层参数共享，以及以句子顺序预测（SOP）取代下一句预测（NSP）。sahajBERT[3] 是一个针对孟加拉语协同预训练的 ALBERT 模型，采用掩码语言建模（MLM）和句子顺序预测（SOP）目标。表 I 展示了本研究使用的孟加拉语预训练语言模型的各种配置。

IV. 所提方法

本研究提出的方法分为以下四个步骤。所提出的工作流程与 [5] 和 [6] 的主要区别并不在于架构，而在于语言专用 BERT 模型的使用。[6] 中描述的方法仅使用了 Bangla BERT Base。

[1] https://github.com/sagorbrur/bangla-bert[2] https://huggingface.co/monsoon-nlp/bangla-electra

[3] https://huggingface.co/neuropark/sahajBERT

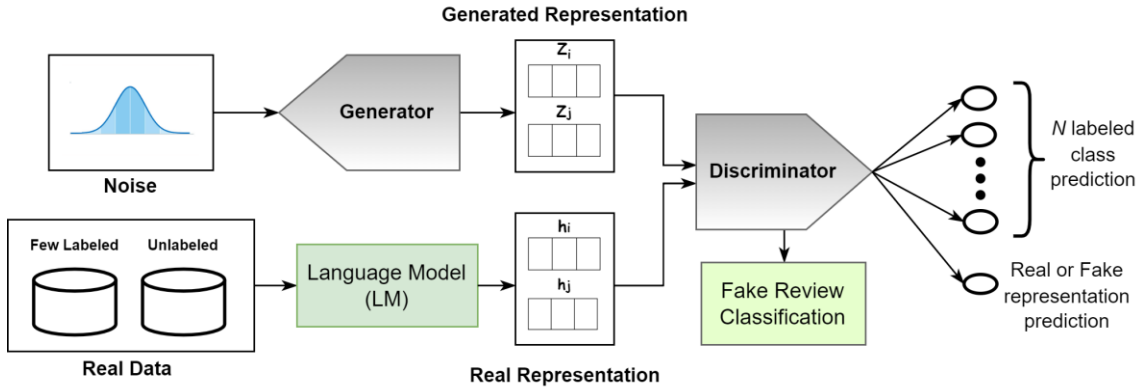


图1：Pr 用于采用半监督GAN架构对语言模型进行微调的所提出的方法

其架构，而我们采用了几种广泛可用的孟加拉语预训练语言模型，例如 BanglaBERT、BanglaBERT Generator、Bangla-Electra 和 sahajBERT。

步骤 1）文本规范化：采用 [19] 中指定的孟加拉语文本清理模块对整个数据集进行预处理。文本规范化包括处理多个空白、替换 URL、表情符号和 Unicode，以及修正单引号或双引号。

步骤 2）数据划分：数据集被划分为三部分：标注样本、未标注样本和测试样本。数据划分依据表 II 中列出的数值进行。在所有实验中，未标注样本数量为 512，而标注样本数量逐步增加。

步骤 3）使用半监督 GAN 微调语言模型：图 1 所示的半监督 GAN-LM 架构与 Croce 等人 [5] 的架构类似，包含一个生成器和一个判别器，二者均为多层感知机（MLP），本步骤将其与第 III-B 节讨论的可用孟加拉语预训练语言模型作为编码器一起使用。生成器模型以服从高斯分布的 100 维噪声向量作为输入，输出维度等于语言模型嵌入大小的表示向量。将上一步的标注样本和未标注样本作为输入提供给语言模型，输出为句子嵌入，即维度等于生成表示的真实分布向量。

表 II：实验数据分布

标注样本数 未标注样本数 测试样本数		
32	512	512
64	512	512
128	512	512
256	512	512
512	512	512
1024	512	128

生成的分布向量和真实的分布向量作为输入提供给判别器。判别器不仅学习将真实实例的表示分类到 (1, ..., n) 个类别之一，还通过在 $(n + 1)^{th}$ 类中设置 0 或 1 的值，来识别表示是真实还是伪造的。对于本文使用的半监督 GAN-LM架构，我们采用了与 [5] 中所述相同的对抗训练目标和损失函数。在反向传播过程中，语言模型的权重与判别器的权重一同调整，以改进有标签和无标签数据的中间表示。训练后，使用预训练语言模型进行推理，并移除生成器。

步骤 4）使用性能指标进行评估：使用测试样本衡量微调后语言模型的性能。为衡量性能并进行比较，采用准确率、精确率、召回率和 F1分数等评估指标。

V. 实验结果与评估

A. 数据集描述

本研究使用的数据集共包含 6014 条以孟加拉语撰写的虚假和真实评论，这些评论收集自一些公开可访问的 Facebook 群组，群组中的孟加拉语使用者会定期分享对食品的评论。我们手动收集了群组成员发布的有关各类餐厅食品和服务的帖子，并将其标记为真实或虚假。对三位不同人员所做的标注进行多数投票后，我们分别收集到 871 条虚假评论和 5015 条真实评论。

B. 超参数设置 对于所有实验，

我们使用了批量大小为 16、判别器和生成器的学习率均为 5e-5、以 AdamW 作为优化器、以 Softmax作为激活函数；BanglaBERT、BanglaBERT 生成器、Bangla BERT Base、Bangla-Electra 和 sahajBERT 的训练轮次数量分别为 7、18、25、13 和 26。

C. 实验结果与分析

我们使用 1024 个标注样本对 BanglaBERT 进行微调，得到的准确率为 65.62%。对于二分类任务而言，这一准确率非常低，因此我们决定采用 GAN-LM架构。表III 展示了微调模型与五种采用半监督生成对抗网络架构的不同孟加拉语语言模型之间的性能比较，并在图2中以图形形式呈现了测试准确率随训练轮次的变化。

。**BanglaBERT** 在我们的数据集上采用 GAN-LM架构时，性能优于所有其他语言模型。从表III可以看出，BanglaBERT 的准确率最高，达到 83.59%，分别比 BanglaBERT Generator 和 Bangla BERT Base 高出近 3% 和 4%，比 Banla-Electra 高出 10%（使用 1024 个标注样本时）。F1-分数为 84.89%，进一步证明了该模型在所有模型中表现最佳，在两种情况下都能避免错误预测。此外，仅使用 32 个和 64 个样本时，其准确率也分别达到 69.53% 和 71.09%。

表III：不同语言模型的性能比较

模型	数量标注样本	准确率	精确率	召回率	**F1 –** 分数
微调后 孟加拉语 BERT	1024	**0.65625**	0.68000	0.54838	0.60714
GAN 孟加拉语 BERT	32	**0.69531**	0.71179	0.64427	0.67635
	64	**0.71093**	0.76650	0.59684	0.67111
	128	0.73047	0.73469	0.71146	0.72289
	256	0.75391	0.71381	0.83795	0.77091
	512	0.76563	0.78298	0.72727	0.75410
	1024	**0.83593**	0.84286	0.85507	0.84892
GAN 孟加拉语 BERT 生成器	32	**0.65234**	0.65060	0.64032	0.64542
	64	0.67578	0.66540	0.69170	0.67830
	128	0.71680	0.70930	0.72332	0.71624
	256	0.77539	0.75746	0.80237	0.77927
	512	0.78711	0.78571	0.78261	0.78416
	1024	**0.80468**	0.82353	0.81159	0.81752
GAN 孟加拉语 BERT Base	32	0.55469	0.55760	0.47826	0.51489
	64	0.64258	0.62868	0.67589	0.65143
	128	0.67188	0.66798	0.66798	0.66798
	256	0.68555	0.69328	0.65217	0.67210
	512	0.73633	0.72519	0.75099	0.73786
	1024	**0.79687**	**0.79452**	**0.84058**	0.81690
GAN 孟加拉语-ELECTRA	32	0.55273	0.55405	0.48617	0.51790
	64	0.58398	0.60638	0.45059	0.51701
	128	0.65234	0.64591	0.65613	0.65098
	256	0.65234	0.61610	0.78656	0.69097
	512	0.68359	0.64444	0.80237	0.71479
	1024	**0.73437**	**0.72727**	0.81159	0.76712
GAN 简单 BERT	32	0.51172	0.50319	0.93676	0.65470
	64	**0.68750**	0.72683	0.58893	0.65066
	128	0.66016	0.67873	0.59289	0.63291
	256	0.72852	0.70956	0.76285	0.73524
	512	0.73438	0.71910	0.75889	0.73846
	1024	0.46094	N/A	N/A	N/A

图2a展示了 BanglaBERT 模型的稳定性：标注样本数从 32 增至 1024 时，其准确率持续上升，且不像其他语言模型那样出现明显波动。在使用 1024 个标注样本时，参数量仅为 34M 的 BanglaBERT 生成器在其他语言模型中取得了第二高的准确率。它仅比 BanglaBERT 低 3.12%，但准确率仍高于其他模型。在使用 32 和 64 个样本时，其准确率也相当可观，分别为 65.23% 和 67.57%。BanglaBERT 生成器的测试准确率相对于

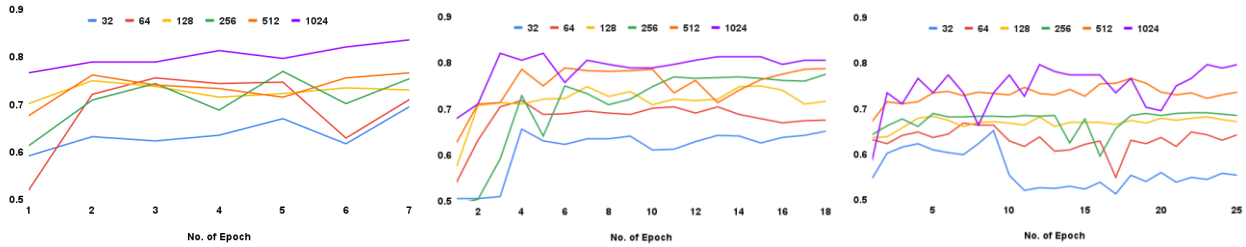
图2b中训练轮次的变化清楚地展现了该模型凭借较少参数所具备的全部潜力。尽管 Bangla BERT Base 的参数量更多（110M），其表现却几乎与 BanglaBERT 生成器相同。该模型的准确率为 79.68%，与 BanglaBERT 生成器的准确率极为接近，但其精确率较低，表明在使用 1024 个标注样本时，它比 BanglaBERT 生成器更容易产生假阳性预测。图2c展示了该模型的不稳定性：测试准确率在最多 25 个训练轮次内起伏较大。Bangla-Electra 在所有模型中参数量最少（14M），但考虑到其参数规模，它的表现相当突出。该模型使用 1024 个标注样本时准确率达到 73.43%，但精确率为 72.72%、召回率为 81.15%，这表明它在识别假阳性案例方面存在相当大的困难。随着标注样本数增加，该模型出现了幻觉，将大量虚假数据预测为真实数据。标注数据极少时，该模型表现不佳。尽管它在少量标注数据下不如其他语言模型，但随着训练轮次增加，其测试准确率平稳提升，这一点可从图2d中看出。sahajBERT 的表现也令人意外，因为它在少量标注数据下取得了尚可的准确率。使用 64 和 256 个标注样本时，其准确率分别为 68.75% 和 72.85%，与其他模型相当接近；但标注样本数增至 1024 时，它的表现不佳。当数据集中虚假评论与真实评论的表征差异较小时，该模型参数量较少可能导致其无法准确检测测试数据。图2e解释了该模型的行为，展示了测试准确率如何随训练轮次变化而波动。它在使用 32 个标注样本时表现较差；但训练 15 轮后，在使用 64 至 512 个标注样本时表现良好；训练 26 轮后，准确率趋于稳定。

VI. 结论与未来工作

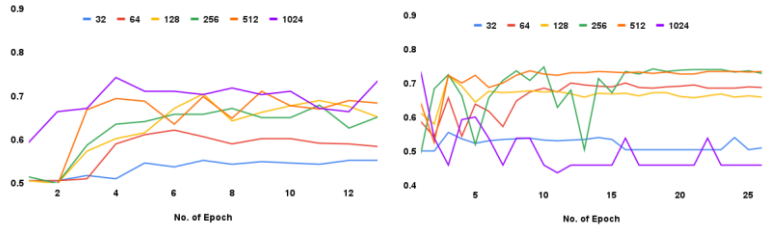
采用半监督GAN架构微调语言模型，是利用少量标注数据对孟加拉语虚假评论进行分类的可行方案。实验结果表明，即使只有 1024 个标注样本，我们提出的方法在以半监督方式微调 BanglaBERT 和 Bangla BERT Base 后，仍能更好地对评论进行分类。尽管参数较少，Bangla-Electra、SahajBERT 和 BanglaBERT 生成器也取得了不错的结果。探索如何利用生成对抗网络生成孟加拉语评论，将是一个值得关注的发展方向。

致谢

本研究工作是在“AUST内部研究资助”资助的“孟加拉语虚假评论：基准数据集与检测系统”项目下开展的。



(a) BanglaBERT (b) BanglaBERT Generator (c) Bangla BERT Base



(a) Bangla-Electra (b) sahajBERT

图2：测试准确率 y 与实验模型的训练轮次（分别使用 32、64、128、256、512 和 1024 个标注样本）的对比 s

参考文献

[1] Jindal、Nitin 和 Bing Liu。“垃圾评论与分析。”载于《2008 年国际网络搜索与数据挖掘会议论文集》，第219-230 页。2008 年。[2] Shahariar、G. M.、Swapnil Biswas、Faiza Omar、Faisal MuhammadShah 和 Samiha Binte Hassan。“使用深度学习检测垃圾评论。”载于《2019 年IEEE 第十届年度信息技术、电子与移动通信会议（IEMCON）》，第 0027-0033 页。IEEE，2019 年。[3] Aghakhani、Hojjat、Aravind Machiry、Shirin Nilizadeh、ChristopherKruegel 和 Giovanni Vigna。“使用生成对抗网络检测欺骗性评论。”载于《2018 年 IEEE 安全与隐私研讨会（SPW）》，第89-95 页。IEEE，2018 年。[4] Kennedy、Stefan、Niall Walsh、Kirils Sloka、JenniferFoster 和 Andrew McCarren。“事实还是虚构？语境化垃圾评论检测。”arXiv 预印本，arXiv:2010.15296（2020 年）。[5] Croce、Danilo、Giuseppe Castellucci 和RobertoBasili。“生成对抗网络-BERT：利用少量标注样本进行稳健文本分类的生成对抗学习。”载于《计算语言学协会第58届年会论文集》，第 2114-2119 页。2020 年。[6] Tanvir、Raihan、Md Tanvir Rouf Shawon、Md Humaion Kabir Mehedi、Md MotaharMahtab 和 Annajiat Alim Rasel。“基于生成对抗网络-BERT 的孟加拉语文本分类方法：使用少量标注样本。”载于《分布式计算与人工智能国际研讨会》，第20-30 页。Springer，Cham，2023 年。[7] Ta、HoangThang、Abu Bakar Siddiqur Rahman、LotfollahNajjar 和AlexanderGelbukh。“生成对抗网络-BERT：从社交媒体中检测攻击性和暴力事件的对抗学习。”载于《伊比利亚语言评测论坛论文集（IberLEF 2022）》，CEUR 研讨会论文集。CEUR-WS.org。2022 年。[8] Zaharia、George-Eduard、Andrei-Marius Avram、Dumitru-ClementinCercel和Traian Rebedea。“通过对抗学习和知识蒸馏进行罗马尼亚语 BERT方言识别。”载于《第八届面向相似语言、变体和方言的自然语言处理研讨会论文集》，第 113-119 页。2021年。[9] Yusuf、Mahmoud、MarwanTorki 和NagwaM.El-Makky。“使用少量标注样本和生成对抗网络进行阿拉伯语方言识别。”载于《亚洲及太平洋地区计算语言学协会第二届会议暨自然语言处理第十二届国际联合会议论文集》，第 196-204 页。2022 年。

[10] Col o´n-Ruiz，Cristo´bal 和 Isabel Segura-Bedmar。“用于药品评论情感分析的半监督生成对抗网络。”（2021）。[11] Auti，Tapan、Rajdeep Sarkar、Bernardo Stearns、Atul Kr Ojha、Arindam Paul、Michaela Comerford、Jay Megaro、John Mariano、Vall Herard 和 John Philip McCrae。“使用 GAN-BERT 对法律制药文本进行分类。”载于第 13 届语言资源与评估会议期间举办的首届计算与社会责任研讨会论文集，第 52–57 页。2022。[12] Irissappane，Athirai A.、Hanfei Yu、Yankun Shen、Anubha Agrawal 和 Gray Stanton。“通过对抗训练，利用 GPT-2 在标注数据有限的情况下对垃圾评论进行分类。”arXiv 预印本，arXiv:2012.13400（2020）。[13] Goodfellow，Ian、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。“生成对抗网络。”《ACM 通讯》63 卷，第 11 期（2020）：139–144。[14] Salimans，Tim、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford 和 Xi Chen。“改进生成对抗网络的训练技术。”《神经信息处理系统进展》第 29 卷（2016）。[15] Devlin，Jacob、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。“BERT：用于语言理解的深度双向 Transformer 预训练。”arXiv 预印本，arXiv:1810.04805（2018）。[16] Clark，Kevin、Minh-Thang Luong、Quoc V. Le 和 Christopher D. Manning。“ELECTRA：将文本编码器预训练为判别器而非生成器。”arXiv 预印本，arXiv:2003.10555（2020）。[17] Bhattacharjee，Abhik、Tahmid Hasan、Wasi Ahmad Uddin、Kazi Mubasshir、Md Saiful Islam、Anindya Iqbal、M. Sohel Rahman 和 Rifat Shahriyar。“BanglaBERT：孟加拉语低资源语言理解评估的语言模型预训练与基准测试。”《北美计算语言学协会：NAACL 论文集》（2022）。[18] Lan，Zhenzhong、Mingda Chen、Sebastian Goodman、Kevin Gimpel、Piyush Sharma 和 Radu Soricut。“ALBERT：用于语言表征自监督学习的轻量级BERT。”arXiv 预印本，arXiv:1909.11942（2019）。[19] Hasan，Tahmid、Abhik Bhattacharjee、Kazi Samin、Masum Hasan、Madhusudan Basak、M. Sohel Rahman 和 Rifat Shahriyar。“不再是低资源：用于孟加拉语—英语机器翻译的对齐器集成、批次过滤与新数据集。”arXiv 预印本，arXiv:2009.09359（2020）。