arXiv:2601.00444v1 [cs.CL] 1 Jan 2026

轻量级 Transformer 模型比较效率分析：面向企业级自然语言处理部署的多领域实证基准测试

Muhammad Shahmeer Khan计
算与数学学院，阿尔斯特大学，曼彻斯特，英国
khan-MS15@ulster.ac.uk

摘要

在快速演变的企业级自然语言处理（自然语言处理）领域，能够处理多领域文本自动化任务的高效轻量级模型需求日益增长。本研究对三种知名轻量级 Transformer 模型——DistilBERT、MiniLM 和 ALBERT 进行了比较分析，涵盖三个不同领域：客户情感分类、新闻主题分类以及毒性/仇恨言论检测。我们使用 IMDB、AG News 和仇恨言论测量语料库的数据集，通过基于准确率的指标（准确率、精确率、召回率、F1 分数）和效率指标（模型大小、推理时间、吞吐量和内存使用量）评估模型表现。主要研究结果表明，没有任何单一模型能在所有性能维度上占据主导地位：ALBERT 在多个领域实现了最高的任务特定准确率，MiniLM 在推理速度和吞吐量方面表现出色，而 DistilBERT 在保持竞争力的效率的同时，在各项任务中展现出最稳定的准确率。所有结果均反映了在固定的面向企业的约束条件下进行的受控微调，而非穷尽式超参数优化。这些结果凸显了准确率与效率之间的权衡，并建议将 MiniLM 用于延迟敏感型企业应用，将 DistilBERT 用于均衡性能场景，将 ALBERT 用于资源受限环境。

关键词 Transformer 模型；轻量级自然语言处理；DistilBERT；MiniLM；ALBERT；文本分类；模型效率；推理延迟；内存占用；基准测试；IT 服务自动化；新闻分类；仇恨言论检测；企业自动化；中小企业部署。

1 引言

企业越来越依赖自然语言处理来自动化基于文本的流程，例如对客户反馈进行情感分析、对新闻源进行主题分类，以及为网络安全进行内容审核。然而，在资源受限的环境中部署 BERT 等大型 Transformer 模型，会面临计算成本、延迟和内存使用量方面的挑战，从而可能阻碍其在企业环境中的可扩展性 [1, 2], ；而在这些环境中，实时处理和成本效率至关重要。轻量级模型通过减少参数数量和推理时间，同时保留大部分性能来解决这些问题，因此非常适合部署在边缘设备或云资源受限环境中。尽管以往的研究已对 BERT 和 RoBERTa 等大型模型进行了广泛的基准测试，但针对多个与企业相关的领域对轻量级 Transformer 模型进行比较评估，仍缺乏充分研究。现有研究通常只在单一数据集或任务上进行基准测试，限制了其在现实场景中的适用性。这使得针对多领域企业应用场景、经过优化的轻量级替代方案成为研究空白；在这些场景中，模型必须能够泛化至多种文本类型，同时保持高效。

我们的贡献包括：
(1) 一个系统对 DistilBERT、MiniLM 和 ALBERT 在三个与企业相关的领域进行系统比较 s

（2）整合准确率和效率指标，以突出实际应用中的权衡；（3）根据实证结果，为企业部署提供建议。

本研究并未提出新的模型架构或训练算法。相反，其主要贡献在于针对广泛采用的轻量级 Transformer 模型，在一致的企业约束下开展面向部署的实证基准测试。本研究采用相同的微调协议，评估模型在多个文本分类领域中的准确率与效率权衡，为真实企业级自然语言处理系统中的模型选择提供实用决策指导；在这些系统中，延迟、吞吐量和资源效率往往比准确率的微小提升更受重视。

2 相关工作

2.1 企业级自然语言处理中的 Transformer 模型

Transformer 架构已成为自然语言处理的基础， [3], 支持从语义理解到文档分类等各类应用。然而，BERT 和 RoBERTa 等大型模型会带来显著的计算开销，限制了它们在对延迟敏感且资源受限的企业环境中的适用性[2]。

这推动了对轻量、易于部署的变体的需求，这些变体在保持强劲性能的同时，能够降低推理延迟、内存使用量和运营成本。

在企业级部署中——例如客户关系管理分析、风险评估系统和内容审核流水线——吞吐量和单次预测成本往往比准确率的细微提升更重要。这一趋势推动了参数高效型 Transformer 变体的开发与评估，使其针对实时生产工作负载进行优化。

2.2 轻量级 Transformer 架构

为解决全尺寸模型的计算限制，研究者提出了多种架构：

DistilBERT [1] 通过知识蒸馏将 BERT 的规模缩减 40%，在推理时间减半的同时保留约 97% 的性能。其在准确率与速度之间取得平衡，因此成为现实世界自然语言处理系统中应用最广泛的紧凑型模型之一。

MiniLM（Wang 等人，2020）提出了深度自注意力蒸馏，通过压缩注意力关系而非完整隐藏状态来缩小模型。该方法使模型规模显著缩小，同时在各项基准测试中保持有竞争力的准确率；尤其是由于隐藏维度减少，推理速度表现出色。

ALBERT （Lan 等人，2020）通过因式分解嵌入和跨层参数共享实现参数效率，大幅缩小模型大小（ALBERT-Base 最小可至 12M 个参数）。虽然这有利于降低内存占用，但共享参数会引入序列依赖，可能增加推理延迟。

总体而言，这些架构代表了当前专为可扩展部署而设计的轻量级 Transformer 格局。

2.3 多领域自然语言处理评估

基准测试研究通常只在少数任务或单一数据集上评估模型，例如情感分析或问答。虽然这种做法在学术研究中有用，但无法体现企业环境中的多领域差异性。近期研究强调，应在多种分类任务中评估模型，以测试其泛化能力。 [4, 5] 然而，这些研究要么：

- 未纳入效率指标（例如推理速度、吞吐量），
- 仅关注单一领域，
- 或对较大型的 Transformer 架构进行基准测试，而非轻量级架构。

因此，在与企业相关的约束下，对轻量级 Transformer 模型进行全面的跨领域评估，这一方向仍鲜有探索。

3 方法论

本节介绍了本研究中使用的轻量级 Transformer 模型、数据集、实验设置和评估协议。

3.1 模型

我们评估了三种广泛使用的轻量级 Transformer 架构，这些架构都是适用于企业级部署的实用候选方案：

- **DistilBERT**（distilbert-base-uncased）：BERT 的蒸馏版本，约有 6600 万个参数和 6 个编码器层。它通过从 BERT 进行知识蒸馏来训练，在降低模型深度和计算成本的同时，保留了大部分语言理解能力。
- **MiniLM**（microsoft/MiniLM-L12-H384-uncased）：一种紧凑型 Transformer 模型，采用深度自注意力蒸馏，将较大的模型压缩为更小的网络（约 3300 万个参数）。这种方法着重于保留注意力模式，同时降低隐藏维度，从而实现高效推理。
- **ALBERT**(albert-base-v2)：BERT 的参数高效变体，约有 1200 万个参数。ALBERT 采用因式分解嵌入和广泛的跨层参数共享，在保持模型深度的同时大幅减少可训练参数的数量。

对于每种架构，我们都使用 Hugging Face Transformers 库中公开提供的预训练检查点，并在每个下游任务上通过线性分类头对其进行微调。

3.2 数据集与任务

我们考察了 三个与企业相关的领域，并将每个领域都定义为一项有监督文本分类任务 ：

- **领域 1 – 客户情感分类（IMDB）**：我们使用了 IMDB 电影评论数据集，其中包含 50,000 条评论，每条评论均标记为正面或负面。采用标准的数据集划分，其中 25,000 个样本用于训练，25,000 个样本用于测试。
- **领域 2 – 新闻主题分类（AG News）**：AG News 语料库包含 120,000 个训练样本和 7,600 个测试样本，这些样本被标注为四个类别：世界、体育、商业和科学/技术。该任务旨在近似企业内容路由和基于主题的自动化流水线。
- **领域 3 – 毒性 / 仇恨言论检测（Measuring Hate Speech）**：我们使用仇恨言论测量语料库 [6], ，该语料库为每条评论提供连续的仇恨_言论_分数，以与实际的内容审核分类体系保持一致；我们将该分数离散化为三个类别：1. **中性**：分数 < 0.32. **冒犯性**：$0.3 \leq$分数 < 0.73. **仇恨言论**：分数 ≥ 0.7

由于仇恨言论测量数据集发布时仅包含一个训练集划分，我们使用固定随机种子 42，将其按 80/20 进行分层划分，得到训练集和验证子集。验证子集被用作该领域的评估集。

在所有领域中，任务均被视为标准的有监督文本分类问题，每个实例对应一个单一标签。

3.3 实验设置

所有实验均使用 Python，并基于 Hugging Face Transformers 和 Datasets 库实现。训练集和评估均在单个 NVIDIA T4 GPU 上使用 Google Colab 完成，以确保运行条件一致。因此，我们的研究结果应被理解为快速微调约束下的模型行为，而非完全收敛的性能。

对于每种模型–数据集组合，我们采用以下共享超参数：

- 优化器：AdamW（Transformers Trainer 的默认设置）

- 学习率： 2×10^{-5}
- 训练轮数：3
- 每设备训练批次大小：16
- 每设备评估批次大小：16
- 权重衰减：0.01
- 日志记录频率：每 200–500 步（取决于数据集）

使用各模型对应的分词器进行分词。对于所有数据集，输入文本字段均为 text。我们启用截断，将最大序列长度固定为 128 个词元。这一长度涵盖了所有数据集中大多数文本的长度，同时能显著提升面向部署场景的推理效率。批处理时使用 DataCollatorWithPadding 动态填充，以尽量减少对较短序列的不必要计算。所有模型均微调三个训练轮数，以确保不同架构下的收敛稳定，同时将训练时间控制在适合企业级部署场景的范围内。

对于 IMDB 和 AG News 数据集，我们使用官方测试划分进行评估。对于仇恨言论测量数据集，留出的 20% 验证子集用作评估集。
所有模型均在相同的超参数下独立微调，以确保对准确率和效率进行公平比较。

3.4 评估协议

我们使用两类指标评估各模型。

基于准确率的指标。 为评估分类质量，我们报告：

- 准确率，定义为正确分类实例所占的比例。
- 加权精确率、召回率和 F1 分数，其中各类别的指标按类别样本数加权。这对于存在类别不平衡的仇恨言论检测任务尤为重要。
- 针对每个模型和领域分别计算混淆矩阵，以分析误分类模式。受篇幅限制，这些矩阵见补充材料（附录 A–C）。

效率指标。 为评估在企业约束下的部署可行性，我们测量了：

- 模型大小（MB），计算方法为可训练参数数量乘以 4 字节（FP32 表示）。
- 单个样本的推理时间（毫秒），通过对固定的 200 个评估样本子集进行预测计时，并计算每个实例的平均实际运行时间得出。
- 吞吐量（每秒样本数），计算方式为已评估样本数除以总推理时间。
- RAM 使用量（MB），以评估后 Python 进程的常驻内存衡量。

效率测量详情。所有推理测量均在禁用梯度计算的情况下使用 torch.no_grad() 进行。推理延迟采用批量大小为 1 进行测量，以反映实时部署场景；测量前先进行预热阶段，以减少内核编译和缓存效应的影响。

吞吐量测量反映了在相同运行条件下的总体处理速率。RAM 使用量报告的是近似的进程级内存消耗，其中包含框架开销；因此，应将其视为粗略指标，而非仅针对模型的指标。所有效率测量均依赖硬件，应作为相对比较来解读，而非视为绝对的部署保证。为确保不同架构之间比较公平，结果均在相同运行条件下报告；结论依据一致的效率排名得出，而非原始延迟值。全程使用 FP32 推理，以尽可能提高可复现性和模型间的可比性，同时避免引入特定精度优化。

可复现性与报告。 所有数据集划分和训练运行均使用固定随机种子 42，以确保可复现性。由于本研究的主要目标是进行面向部署的基准测试，因此我们报告单次运行结果，不进行重复试验或统计显著性检验。在企业环境中，吞吐量、延迟和

表 1：性能与效率指标 — 领域 1（IMDB 情感分析）

模型	准确率指标				效率指标		
	Acc.	精确率	Rec.	F1	大小（MB）	推理时间（毫秒）	吞吐量
DistilBERT	0.932	0.929	0.936	0.932	255.42	14.58	68.57
MiniLM	0.937	0.934	0.942	0.938	127.26	10.11	98.95
ALBERT	**0.944**	**0.943**	**0.945**	**0.944**	**44.58**	32.18	31.08

表2：性能与效率指标——领域 2（AG News）

模型	准确率指标				效率指标		
	Acc.	精确率	Rec.	F1	大小（MB）	推理时间（毫秒）	吞吐量
DistilBERT	0.946	0.946	0.946	0.946	255.42	3.30	302.81
MiniLM	0.947	0.948	0.947	0.947	127.26	**2.14**	**466.88**
ALBERT	**0.949**	**0.949**	**0.949**	**0.949**	**44.58**	7.48	133.66

资源效率是　　通常优先于细微的性能差异，而确定性的实验设置　　p
确保相对比　　较在实践中仍有意义。
为提高可解释性，我们通过一张汇总表呈现各领域的结果，将基于准确率的指标与效率指标相结合，并辅以条形图展示关键效率指标。我们的分析并不侧重于微小的准确率差异，而是关注在相同训练条件下，不同模型和领域中稳定的相对排名及性能变化趋势。因此，对于准确率相近的情况，我们不会声称存在统计显著优势，而是分析与部署决策相关的稳定性和权衡。这样的呈现方式便于直接比较，并有助于为现实世界的企业级自然语言处理部署做出明智的模型选择。

4 结果

4.1 领域1：客户情感分类（IMDB）

在情感分类任务中，ALBERT 的准确率和 F1 分数最高，表明只要经过充分微调，参数效率高的架构也能在二元情感任务中取得有竞争力的表现。DistilBERT 的表现稳定；MiniLM 的准确率略低，但推理延迟更低、吞吐量更高。ALBERT 紧凑的体量在内存受限的企业部署中尤其有优势，而 MiniLM 则在效率与预测性能之间实现了良好平衡。
该领域的混淆矩阵见附录A。

4.2 领域2：新闻主题分类（AG News）

在四分类 AG News 新闻主题分类任务中，ALBERT 的分类准确率最高，DistilBERT 和 MiniLM 紧随其后，这表明主题分类主要依赖粗粒度语义线索，而非细粒度的上下文推理。MiniLM 的吞吐量最高，适用于新闻聚合和内容路由系统中的快速大规模文档分类；在这些系统中，延迟至关重要。

该领域的混淆矩阵见附录 B。

4.3 领域3：毒性与仇恨言论检测

毒性与仇恨言论检测任务是一个更具挑战性且带有主观性的分类问题。这些阈值是务实的操作性近似，常用于实际内容审核场景，

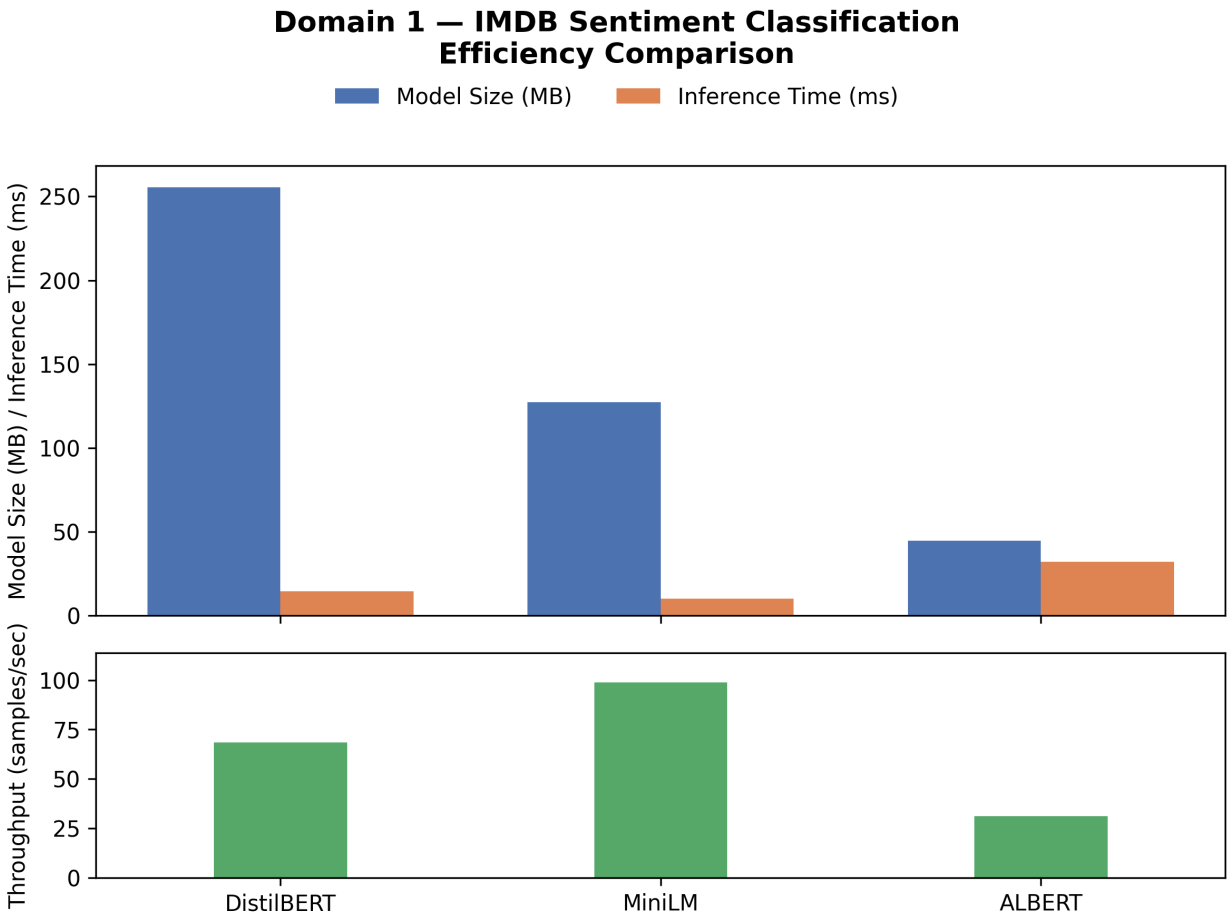


图 1：领域 1（IMDB 情感分析）各模型的推理延迟、吞吐量和模型大小比较。

表 3：性能与效率指标 — 领域 3（网络欺凌）

模型	准确率指标				效率指标		
	Acc.	精确率	Rec.	F1	大小（MB）	推理时间（毫秒）	吞吐量
DistilBERT	0.949	0.948	0.949	0.949	255.42	2.71	368.75
MiniLM	0.906	0.897	0.906	0.900	127.26	**1.85**	**540.17**
ALBERT	**0.951**	**0.950**	**0.951**	**0.950**	**44.58**	6.15	162.51

从而能够进行类别评估，同时承认仇恨言论存在于一个连续谱上。我们并不主张这些界限具有规范性，且采用其他阈值可能会影响类别分布和绝对性能指标。尽管架构紧凑，ALBERT 在标签歧义的情况下仍表现出较强的泛化能力，取得了最高的整体准确率和 F1 分数。MiniLM 推理速度最快、吞吐量最高，因此尤其适用于实时内容审核流水线。ALBERT 在保持竞争力的准确率的同时，模型占用空间最小，适合部署在低资源或边缘环境中。所有模型的 F1 分数均有所降低，反映了数据集中固有的标注噪声和类别重叠。

该领域的混淆矩阵见附录 C。

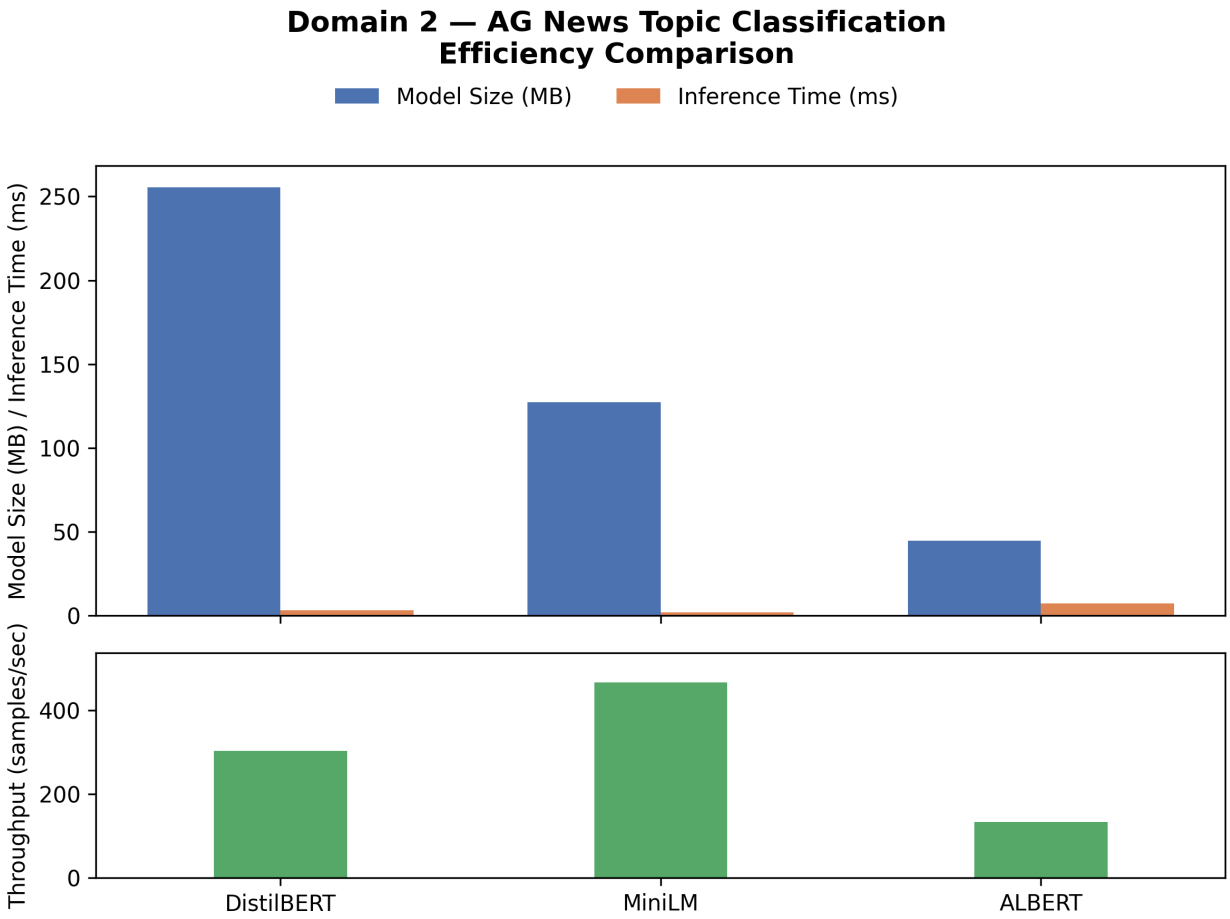


图 2：领域 2（AG News）中各模型的推理延迟、吞吐量和模型大小比较。

5 跨领域分析

在评估的三个领域——情感分类、新闻主题分类和毒性检测——中，评估的轻量级 Transformer 模型呈现出一致的性能模式。在这三个领域中，没有任何单一模型始终占据准确率优势。ALBERT 在多个领域中取得了最高的任务特定准确率，而 DistilBERT 在所有任务中均表现稳定且具有竞争力，这表明尽管模型深度有所降低，它仍具备强大的泛化能力。这种一致性表明，知识蒸馏能够有效保留与任务相关的语言特征，适应多样的文本分布。

MiniLM 始终能提供出色的推理速度和吞吐量，处理速率通常比竞争模型高出 1.5–2× 。这种性能优势在复杂度各异的领域中保持稳定，体现了深度自注意力蒸馏在降低计算开销的同时维持竞争力十足的准确率方面的有效性。这种稳定性凸显了 MiniLM 对高吞吐量、延迟敏感型企业工作流的适用性。

在所有任务中，ALBERT 的模型占用空间始终最小，这表明参数共享和嵌入分解能大幅降低内存需求。然而，其推理速度相对较慢，反映了跨层参数共享带来的架构权衡：这种设计会增加前向传播过程中的序列依赖 [7]。尽管存在这一局限，ALBERT 的准确率仍具竞争力，尤其是在由粗粒度语义线索主导的任务中。

表 4 总结了整体跨领域比较。

这些研究结果展现出一致的准确率与效率的权衡：DistilBERT 最适合对精确率要求极高的任务，MiniLM 适用于高吞吐量部署场景，而 ALBERT 则适合受内存或存储限制的环境。

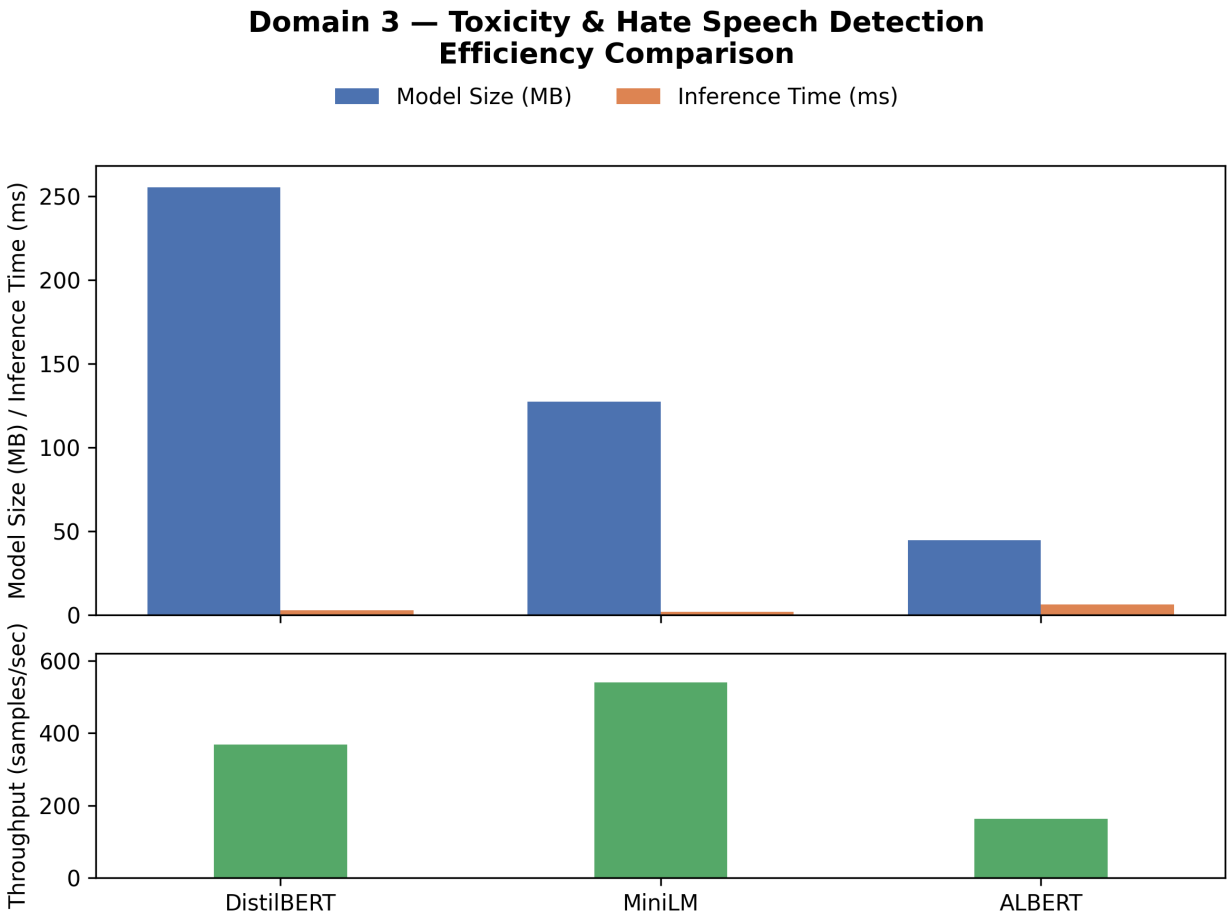


图 3：领域 3（网络欺凌）中各模型的推理延迟、吞吐量和模型大小比较。

表 4：总体跨领域比较摘要

类别	最佳模型	理由
准确率（特定任务）	ALBERT	在多个领域中准确率最高
一致性	DistilBERT	在所有任务中表现稳定
推理速度	MiniLM	延迟最低，推理速度最快
吞吐量	MiniLM	每秒处理的样本数最多
模型大小	ALBERT	最小的内存占用
资源效率	ALBERT	参数最少，同时保持稳定的准确率

6 讨论

虽然 IMDB 和 AG News 是标准的学术基准测试，但它们可以近似反映情感监测、文档路由和自动内容标注等企业级工作负载。这些数据集为客户反馈系统、内部知识库和内容审核流程中所遇到的现实文本流提供了受控且可复现的代理。这些数据集并不声称是专有的企业数据；相反，它们充当了企业级自然语言处理工作负载的标准化、可复现代理，例如情感监测、文档路由和内容审核。

观察到的性能差异可直接归因于架构设计选择。DistilBERT 的高准确率源于知识蒸馏：该方法将 BERT 中丰富的上下文表示迁移到更浅层的模型中，同时保持语义保真度 [1]。这使 DistilBERT 能够在异构领域中实现有效泛化，而不会带来过高的计算开销。

MiniLM 的效率优势源于深度自注意力蒸馏，该方法优先学习注意力关系，而不是完整的隐藏状态表示 [8]。这一设计选择显著减少了隐藏维度和计算成本，使模型仅以适度的准确率下降为代价，便能持续保持较高吞吐量。因此，MiniLM 为实时和大规模企业级自然语言处理系统提供了良好的平衡。

ALBERT 通过激进的参数共享和因式分解嵌入实现紧凑化，大幅缩小模型大小 [7]。虽然这种设计最大限度地降低了内存使用量，但也带来了顺序计算方面的限制，对推理延迟产生负面影响。尽管如此，ALBERT 具有竞争力的准确率表明，这类压缩技术仍适用于低资源场景或边缘部署，尤其是在内存限制比延迟更受关注的情况下。

与 LLaMA 或 Mistral 等较大型的当代模型相比，这些轻量级架构显著降低了推理时间、内存消耗和运营成本。例如，MiniLM 的吞吐量使企业级处理流水线每秒可处理的文档数量增加约 1.5–2× ，从而直接降低基础设施成本并提升可扩展性 [4]。

7 局限性

本研究存在若干局限性。首先，所有实验均基于英语数据集开展；在多语言或低资源语言环境中，性能可能有所不同。其次，评估仅在单一图形处理器配置下进行，若采用仅使用 CPU 的部署或分布式推理环境，结果可能有所差异。第三，超参数调优范围有意受到限制，以反映企业约束，因此可能未能探索潜在的性能提升。

此外，结果来自单次运行的评估，未提供置信区间或进行统计显著性检验。虽然这符合面向部署的基准测试做法——即优先考虑吞吐量和延迟，而非重复进行随机训练——但这限制了正式的统计比较。我们不声称准确率相近的模型之间存在统计显著优势。观察到的差异应被视为固定训练条件下的趋势，而非具有统计显著性的差距。在企业级部署场景中，这类趋势往往足以为架构选择提供参考。未来工作可通过多语言评估、更广泛的硬件性能分析以及受控方差分析来解决这些局限。

可复现性。 所有实验均使用公开数据集和预训练模型进行。完整源代码（包括训练脚本、评估流水线和实验配置）已公开发布于 https://github.com/shahmeer07/enterprise-nlp-lightweight-transformer-benchmark。

8 结论

本研究在快速微调约束下，对轻量级 Transformer 模型进行了面向部署的基准测试。通过在情感分析、新闻分类和毒性检测任务上的评估，我们证明没有任何单一模型能在所有性能维度上占据主导地位。DistilBERT 的准确率最稳定，MiniLM 的推理效率更高，而 ALBERT 的模型占用空间最小。

本研究并未推崇某个单一的最佳模型，而是为企业级自然语言处理部署提供了实用的决策框架，使从业者能够根据准确率要求、延迟限制和资源可用性选择模型。通过整合多个领域的准确率与效率指标，这项基准测试为在现实企业系统中部署可扩展的自然语言处理解决方案提供了切实可行的指导。

参考文献

[1] Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。《DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量》。*arXiv* 预印本 *arXiv:1910.01108*，2019。[2] Ananya Bhattacharjee 等。《Transformer 的现实应用：自然语言处理应用综述》。*Information*，14(4):242，2023。[3] Ashish Vaswani 等。《Attention Is All You Need》。载于神经信息处理系统进展，2017。[4] Shuo Zhang、Minghao Wang、Shuai Liu、Peng Chen 和 Jun Xiong。《一种用于多域分类的轻量级 Transformer 注意力多任务学习模型》。*Scientific Reports*，14:10650，2024。[5] Harsh Patel、Darshan Patel 和 Pankaj Patel。《Deltran15：一种用于 X 上灾害帖子多类别分类的深度轻量级 Transformer 框架》。*IEEE*计算社会系统汇刊，2024。

[6] Brendan Kennedy、Mohammad Atari、Aida Mostafazadeh Davani、Lillian Yeh、Afra Omrani、Youngwoo Kim 和 Morteza Dehghani。《仇恨言论测量语料库：运用拉施测量理论实现数据视角主义》。载于第一届自然语言处理视角主义方法研讨会论文集，线上，2022。计算语言学协会。

[7] Zhenzhong Lan、Mingda Chen、Sebastian Goodman、Kevin Gimpel、Piyush Sharma 和 Radu Soricut。Albert：一种用于自监督学习语言表示的轻量级 BERT。*arXiv* 预印本 *arXiv:1909.11942*，2019 年。

[8] Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou。MiniLM：用于预训练 Transformer 任务无关压缩的深度自注意力蒸馏。*arXiv* 预印本 *arXiv:2002.10957*，2020 年。

领域 1混淆矩阵

本附录展示了使用 DistilBERT、MiniLM 和 ALBERT 对客户情感分类任务（IMDB 数据集）进行分类所得的混淆矩阵。这些矩阵有助于了解类别级预测行为和误分类模式。

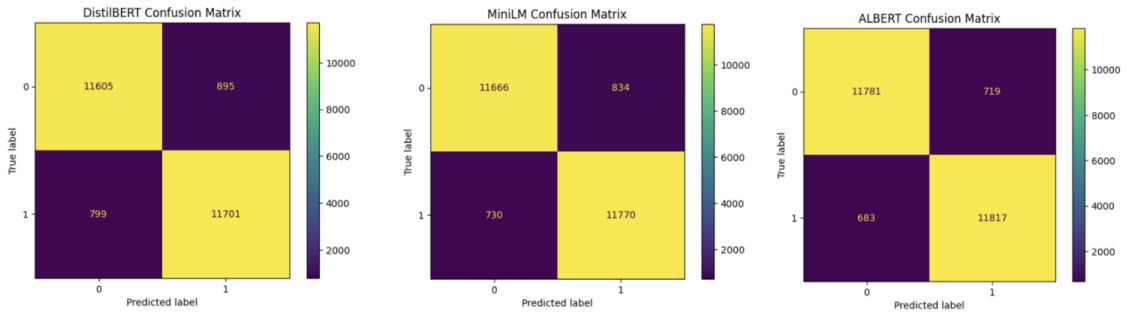


图 4：使用 DistilBERT、MiniLM 和 ALBERT 对领域 1（IMDB 情感分析分类）进行分类所得的混淆矩阵。

B 领域 2 混淆矩阵

本附录报告了使用 AG News 数据集进行四分类新闻主题分类任务的混淆矩阵。结果突显了世界、体育、商业和科学/技术类别之间的类别间混淆。

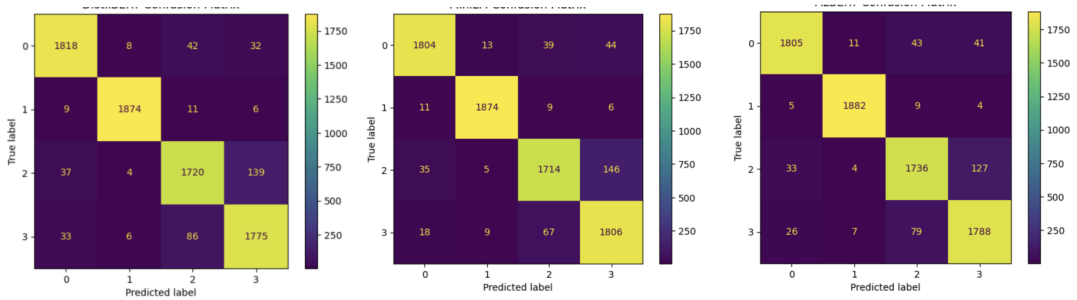


图 5：使用 DistilBERT、MiniLM 和 ALBERT 对领域 2（AG News 主题分类）进行分类的混淆矩阵。

C 领域 3 混淆矩阵

本附录展示了基于仇恨言论测量语料库的三分类毒性与仇恨言论检测任务的混淆矩阵。这些矩阵展示了标签歧义和类别重叠对真实世界内容审核数据集的影响。

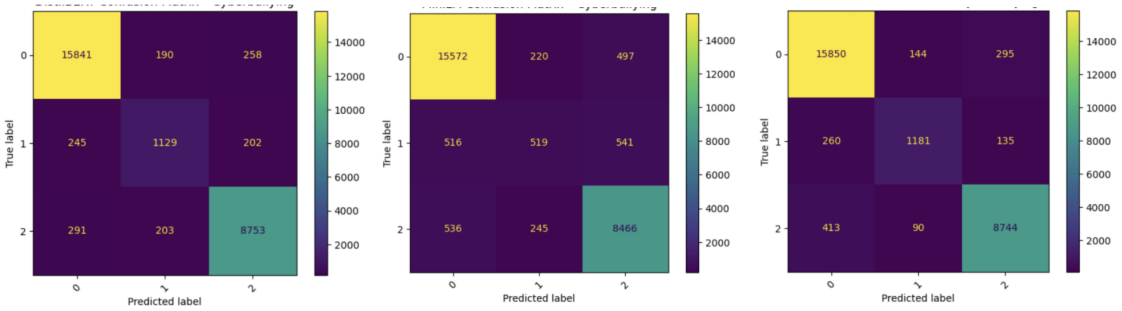


图 6：使用 DistilBERT、MiniLM 和 ALBERT 进行领域 3（毒性与仇恨言论检测）的混淆矩阵。