arXiv:2304.01238v3 [cs.CL] 7 May 2023

Spam-T5：少样本电子邮件垃圾邮件检测的大语言模型基准测试

Maxime Labonne* 和 Sean Moran

摩根大通{maxime.labonne,sean.j.moran}
@jpmchase.com

摘要。 本文通过比较三个不同系列的代表性模型——类 BERT、句子转换器和序列到序列模型，研究了大语言模型（LLM）在电子邮件垃圾邮件检测中的有效性。此外，我们还考察了成熟的垃圾邮件检测机器学习技术，例如朴素贝叶斯和 LightGBM，并将其作为基线方法。我们在四个公开数据集上评估了这些模型的表现，采用了不同数量的训练样本（完整训练集和少样本设置）。研究结果表明，在大多数情况下，大语言模型的表现优于常用的基线技术，尤其是在少样本场景中。这种适应性使大语言模型特别适用于垃圾邮件检测任务，因为这类任务的标注样本数量有限，且模型需要频繁更新。此外，我们还推出了 *Spam-T5*，这是一种经过专门调整和微调、用于检测电子邮件垃圾邮件的 Flan-T5 模型。结果显示，在大多数场景下，Spam-T5 的表现优于基线模型和其他大语言模型，尤其是在可用训练样本数量有限时。我们的代码已公开发布于 https://github.com/jpmorganchase/llm-email-spam-detection。

关键词： 垃圾邮件检测 · 大语言模型 · 少样本学习。

1 引言

电子邮件通信仍是我们日常生活中不可或缺的一部分，为个人和企业用户在全球范围内实现高效的异步通信。鉴于电子邮件的重要性，它也成为垃圾邮件、网络钓鱼攻击等欺诈和恶意活动的主要目标。垃圾邮件会引发诸多问题，从未经请求的通信给用户带来不便，到服务器计算资源过载，以及电子邮件中旨在危害个人和企业安全的欺诈链接与恶意软件附件导致的安全漏洞。据估计，2022年，通过互联网发送的电子邮件中，近49%是垃圾邮件[1]，，这凸显了该问题持续存在，也说明了探索更先进方法的必要性。

* 通讯作者[1]

[1] https://securelist.com/spam-phishing-scam-report-2022/108692/

要减少垃圾邮件数量，采用更先进的机器学习技术至关重要。自动检测并过滤垃圾邮件的技术对于确保个人和企业电子邮件服务的可用性至关重要，也持续吸引着研究人员的浓厚兴趣 [10]。

垃圾邮件的检测面临多项挑战，我们将其归纳为数据不平衡、数据分布偏移和对抗性漂移。尽管互联网中的电子邮件垃圾邮件十分普遍，但训练垃圾邮件检测模型的主要障碍之一，是欺诈邮件带标签的数据集十分稀缺，这使得难以获得具有代表性的样本来训练有效的机器学习模型。对于电子邮件垃圾过滤的私有、公司内部应用而言，由于任务本身的性质，这一问题会进一步加剧，因为可用的带标签数据可能更少（例如，为特定用户群体（如高净值人士）过滤欺诈和网络钓鱼攻击）。这会导致不平衡学习问题：模型可能没有足够的欺诈样本可供学习，从而导致较差的检测性能。

此外，电子邮件通信一直在不断发展变化，从而带来了额外的挑战。由于垃圾邮件发送者所采用的演变策略和技术，垃圾邮件的性质不断变化，可能导致数据集中的数据分布发生偏移 [20,38,41,44]。当训练数据和测试数据的分布不一致时，就会出现数据偏移问题，这种情况可能发生在现实世界的场景中。例如，在某个时间点，“赌场”一词可能是垃圾邮件的指示词，但随着时间推移，该词的相对重要性可能会发生变化（漂移）。这违背了监督学习的基本假设，并可能导致分类模型随着时间推移无法在部署环境中实现泛化。使用来自该领域的、近期且具有代表性的数据持续更新此类模型至关重要。

此外，这一环境具有高度对抗性，其特点是垃圾邮件发送者与电子邮件服务所有者之间不断升级的军备竞赛，而这场竞赛主要由成功实施欺诈和网络钓鱼所能带来的丰厚收益驱动 [20]。攻击者实际上已形成一个小型产业，不断设计新颖巧妙的方法来绕过垃圾邮件过滤器，导致出现了对抗性漂移。他们集中力量，试图胜过文本过滤器，方法是扰动从电子邮件正文和可读邮件头中提取的数据。常见策略包括采用混淆技术、伪装电子邮件内容，或操纵邮件头信息。

为应对这些重大挑战，我们认为一种有前景的方法是使用少样本学习 [48] 来训练分类器，使其能够利用有限的样本检测欺诈邮件。采用这种方法，我们可以减少对大型带标签数据集的需求，并构建能够更好地泛化到未见数据和不断演变数据的分类器。

本文比较了大语言模型（LLMs）与传统机器学习技术在少样本学习设置下进行序列分类的性能。少样本学习非常适用于垃圾邮件检测任务，因为异常类别（垃圾邮件）的出现频率远低于

正常类别（正常邮件）。我们的主要贡献是针对垃圾邮件检测领域最常用的四个数据集，为传统机器学习算法和大语言模型建立基准测试。我们在传统监督学习设置和少样本学习设置下评估了这些模型的性能。此外，我们提出了一种新模型 Spam-T5，这是专为电子邮件垃圾邮件检测设计、经过微调的 Flan-T5。研究结果表明，Spam-T5 在该基准测试中优于所有其他模型。为促进该领域的进一步研究，我们在 https://github.com/jpmorganchase/emailspamdetection 上公开提供了代码和数据。

2 相关工作

2.1 使用机器学习进行垃圾邮件检测

电子邮件垃圾邮件检测任务在相关文献中已得到充分研究 [10] ，并且通常被用于本科机器学习课程，作为学习和研究的入门案例。尽管电子邮件垃圾邮件检测任务广为人知，但由于数据分布漂移、任务所处环境的对抗性以及类别不平衡，它仍然给实践者带来了真正的挑战。针对这一任务，研究人员已经投入了大量研究工作；为了应对不断演变的数据和持续变化的垃圾邮件发送技术格局，还需要持续开展研究。

电子邮件垃圾邮件检测任务被定义为开发一种有效的自动化算法，用于区分垃圾邮件和正常邮件（非垃圾邮件）。垃圾邮件检测方法可根据基于规则的方法、协作方法和基于内容的方法进行分类 [25]。基于规则的方法包括：将收到的电子邮件地址与允许的电子邮件地址白名单、常用垃圾邮件来源的黑名单进行比对，以及使用人工编写的规则（例如查找空的 *To:* 字段或某种特定的关键词组合）。协作方法例如会对垃圾邮件样本的内容计算哈希函数，并将其共享给社区，再将该哈希函数与新电子邮件进行比对，以检测垃圾邮件 [5, 39, 50]。这些简单方法无法很好地泛化，且较为脆弱 [9], ，因此人们开始探索涉及机器学习的基于内容的方法 [4]。

对于基于内容的方法，早期研究探索了传统机器学习方法及其变体，包括朴素贝叶斯分类器 [35],K近邻 [12], 支持向量机（支持向量机） [8, 15, 43, 46], 以及提升树（例如 XGBoost） [6]。这项任务通常被定义为一个二分类问题：使用具有代表性的垃圾邮件和正常邮件（非垃圾邮件）电子邮件数据集训练分类器，并让其学会将数据点划分到这两个类别中。性能通过模型对新收到电子邮件数据的泛化能力来衡量，例如在标准基准测试数据集 [2, 26, 37] 上，或在面向工业应用场景的实时生产环境中进行评估。也有研究探讨了各种手工设计的特征表示对于输入机器学习分类器的有效性 [25]。其他研究则试图解决概念漂移问题，而这正是垃圾邮件检测任务的一个显著特征 [38, 41],

同时也致力于改进适用于该任务的传统分类技术 [3, 18]。一般而言，传统机器学习方法的训练和评估速度比深度神经网络更快，但捕捉数据内在复杂模式的建模能力较弱。

2.2 使用大语言模型进行垃圾邮件检测

随着大语言模型（LLMs）的出现，机器学习领域发生了变革。这是一系列强大的深度学习方法，在理解和学习自然语言数据集中固有的结构、关系和语义模式方面表现出卓越的能力 [24]。大语言模型擅长处理多种任务，能够取代大多数为自然语言任务量身定制的模型。例如，它们非常擅长文本生成、情感检测、问答和文本摘要。ChatGPT 以及近期 GPT4 [28]向公众发布后，在全球引起了轰动；由于模型的回复自然且近似人类，这种影响至今仍在持续 [2]。尽管电子邮件垃圾邮件检测广受关注且十分重要，但此前很少有研究探讨大语言模型在这项任务中的优势。本文将填补这一空白。

鉴于该领域广受关注，关于大语言模型的文献数量众多、内容丰富，新进展几乎每天都在迅速涌现。我们聚焦于学术论文和工业运营系统中常用作基线模型的一系列基准模型，包括 RoBERTa [24],SetFit [45],和 Flan-T5 [49]。这些基准模型的优势在于其代码和权重已公开发布，因此能够针对新任务进行比较和评估。该领域近期几乎所有创新的基础，都是 Google Research 在 2017 [47] 引入的开创性 Transformer 架构。Transformer 架构摒弃了循环层和卷积层，主张主要利用注意力机制进行学习；事实证明，这种做法在处理大数据时具有更强的可扩展性。这项技术被应用于多代生成式预训练 Transformer（GPT）模型（GPT-n） [30], ，使其在众多自然语言理解任务中取得了显著成效 [28]。除较早的 GPT-2 模型 [31], 外，近期一系列 GPT 模型均为闭源，权重被隐藏在应用程序接口之后，因此无法进行比较。

开源且广泛可用的模型包括 RoBERTa、SetFit 和 Flan-T5。BERT（来自 Transformer 的双向编码器表示） [14]解决了现有模型只能从文本中先前的词元学习这一问题，创建了更强大的（双向）Transformer表示，能够从更广泛的上下文中获取信息。此后，BERT 已发展成为适用于许多任务的高性能开箱即用模型。BERT 后续一项广受欢迎的演进体现在 RoBERTa 模型中 [24] ，该模型通过改进训练方法（包括使用更大的小批量和更高的学习率）提升了 BERT。SetFit（句子Transformer微调） [45] 是一种

[2] https://openai.com/blog/chatgpt, https://openai.com/research/gpt-4

近期提出的学习框架，可高效地（使用少几个数量级的参数）微调句子转换器模型 [34]。最后，与 SetFit 理念相近，Flan-T5 [7] 是一种改进的模型微调方法，通过利用由更多任务构成、以指令形式表达的数据集，并将其应用于 T5 大型语言模型（LLM） [32]。Flan 微调方法在标准自然语言理解任务上带来了显著提升。据我们所知，本文首次探索将 RoBERTa、SetFit 和 Flan-T5 用于电子邮件垃圾邮件检测任务，并将它们分别与传统模型进行比较。

在垃圾邮件检测方面，一些作者已将早期深度模型应用于这一任务，包括长短期记忆网络 [17] 和 BERT 架构 [11], ，不过目前相关文献仍非常有限。我们通过在少样本学习范式下探索较新的大语言模型架构，为这一领域作出贡献。

3 方法

3.1 问题表述

设 $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}^n$ 为一个电子邮件数据集，其中每封电子邮件 $\mathbf{x}_i$ 都表示为 $\mathbb{R}^d$ 中的一个特征向量，并被标记为垃圾邮件（$y_i = 1$）或非垃圾邮件（$y_i = 0$），后者也称为正常邮件。垃圾邮件检测分类问题可以表述为学习一个函数 $f: \mathbb{R}^d \to \{0, 1\}$，该函数将电子邮件特征向量映射为二元标签，使得 $f(\mathbf{x}_i)$ 对数据集中的所有电子邮件样本 i 都能近似于 y_i 。

这个问题可以通过监督学习技术来解决：使用训练集 $\mathcal{D}_{train}$ 学习函数 f。具体而言，目标是为分类器 f_θ 找到最优参数 θ，使其最小化适当的损失函数 $L(\theta)$，例如交叉熵损失：

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(f_\theta(\mathbf{x}_i)) + (1 - y_i)\log(1 - f_\theta(\mathbf{x}_i))) \tag{1}$$

其中， n 是训练集中电子邮件的数量。通过最小化损失函数，可以找到最优参数θ^*：

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, L(\theta) \tag{2}$$

找到最优参数后，学习得到的分类器 f_{θ^*} 即可用于预测训练期间未见过的新电子邮件是否为垃圾邮件。

然而，在某些场景下，我们可能无法获取像 $\mathcal{D}_{train}$这样的大型标注数据集。这时，少样本学习便派上了用场。少样本学习是监督学习的一种变体，旨在从少量标注数据 [48]中学习。在少样本学习设置下，我们会获得一个由少量标注样例组成的支持集 $\mathcal{S}$，其中包含 $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}^k$，且 $k \ll n$；我们需要学习一个函数 f，使其能够泛化到训练期间未见过的新样例。

这种设置在垃圾邮件检测场景中尤为重要，因为标注样本稀缺，且需要频繁更新以应对数据偏移和对抗性漂移。

3.2 文本数据的数据处理流程

传统的机器学习文本分类需要对文本数据进行预处理并提取有用的特征。由此得到的数值特征可用于训练机器学习模型 [21, 23]。

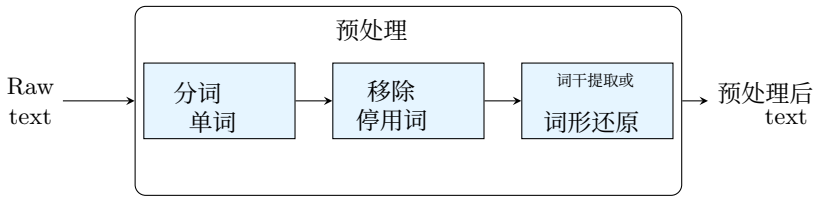

图 1： 用于文本分类的传统预处理步骤流程图。

预处理。预处理的目标是将原始文本转换为更简洁、结构更清晰的表示形式，以便应用特征提取算法（见图 1）。

第一步是分词，即将输入文本拆分为单独的单词、短语或其他具有意义的单位。第二步是移除停用词，停用词是“the”“and”或“in”等常见词语，几乎不承载语义信息。删除这些词不会影响文本含义，因此可以安全地将其舍弃。第三步是应用词干提取 或词形还原，这些技术用于将单词还原为其基本形式或词根。词干提取通过移除单词的后缀和前缀来实现，而词形还原则是将单词映射为其规范形式或词典形式。

我们的预处理流程包括单词分词和词干提取，并使用Porter词干提取算法[3] [29]。

特征提取。 下一步是从预处理后的文本中提取相关特征。可以考虑使用多种算法完成此任务，例如词袋 [16] 和Word2Vec [27]。本研究将重点介绍广泛使用的词频-逆文档频率（tf-idf）编码。

TF-IDF编码是一种常见方法，可将文本文件表示为数值向量，并将其输入机器学习模型。这种编码的核心思想是：对于在特定文档中频繁出现（词频较高），但在其他文档中较少出现（逆文档频率较高）的单词，赋予更高的权重。这有助于捕捉文档的独特特征，并将其与语料库中的其他文档区分开来。

形式上，文档 d 中词项 t 的TF-IDF编码可定义为：

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) \tag{3}$$

[3] https://www.nltk.org/_模块/nltk/stem/porter.html

其中，tf(t, d) 是文档 d 中词项 t 的词频，而 idf(t) 是词项 t 的逆文档频率，计算方式如下：

$$\mathrm{idf}(t) = \log \frac{N}{n_t} \tag{4}$$

其中，N 是语料库中的文档总数，n_t 是包含词项 t的文档数。对数函数会减弱 n_t值极低的罕见词项所产生的影响。

4 实验设置

4.1 数据集

我们采用了四个广受认可的垃圾邮件数据集：Ling-Spam 数据集、SMS 垃圾邮件集合、SpamAssassin 公共语料库和安然电子邮件数据集。选择这些数据集，是因为它们在垃圾邮件检测领域广受欢迎 [1,36,42]，且涵盖了多种通信渠道，包括短信、邮件列表和其他来源。

Ling-Spam 数据集。 Ling-Spam 数据集（2003年）是一个消息集合，曾用于与垃圾邮件检测相关的实验 [37]。该数据集由垃圾邮件和通过 Linguist 邮件列表发送的正常邮件组成。Linguist 邮件列表是一个经审核的语言学主题邮件列表。语料库共包含 2893 条消息，其中 2412 条为正常邮件，这些邮件是从该列表的档案中随机下载邮件摘要并移除服务器添加的文本后获得的。其余 481 条为垃圾邮件，均由其中一位作者收到，垃圾邮件比例约为 16%。Linguist 邮件列表中的消息涉及多个主题，包括招聘信息和软件可用性公告。

SMS 垃圾邮件集合。SMS 垃圾邮件集合[4] （2011年）是一个包含 5,574 条英语短信的数据集，这些短信被标记为正常邮件或垃圾邮件 [2]。该数据集收集自互联网上各种免费或可用于研究的来源。这些来源包括一个英国论坛（手机用户在此举报短信垃圾邮件）、新加坡国立大学为研究收集的合法短信数据集、一篇博士论文中的正常邮件列表，以及 SMS Spam Corpus v.0.1 Big。该数据集类别分布不均衡，垃圾邮件比例约为 13%。

SpamAssassin 公共语料库。 SpamAssassin 数据集[5] 是一个公开提供的电子邮件集合，适用于测试垃圾邮件过滤系统。该数据集包含 6,047 封邮件，垃圾邮件比例约为

[4] https://archive.ics.uci.edu/ml/datasets/sms+垃圾邮件+collection[5] https://spamassassin.apache.org/old/publiccorpus/

31%。这些邮件发布于公共论坛、发送给创建者，或源自公共新闻网站的新闻简报。该语料库分为五个部分：垃圾邮件、易识别的正常邮件、难识别的正常邮件、易识别的正常邮件 2 和垃圾邮件 2；区分垃圾邮件与非垃圾邮件的难度各不相同。本研究使用了这五个部分的最新版本，分别来自 2003年 和 2005年。

安然电子邮件数据集。安然电子邮件数据集[6]，也称为安然语料库，于2002年在对安然公司的破产进行调查期间收集。该数据集由158名员工生成，包含60多万封电子邮件。我们使用的是该数据集的一个较小版本，其中包含33,716封电子邮件，并标注为垃圾邮件或正常邮件。安然电子邮件数据集是平衡的，包含17,171封垃圾邮件和16,545封正常邮件（垃圾邮件比例约为49%）。

所有这些数据集均经过预处理，删除了重复项、NaN值和空消息。图2展示了每个预处理后数据集中垃圾邮件和正常邮件的比例。

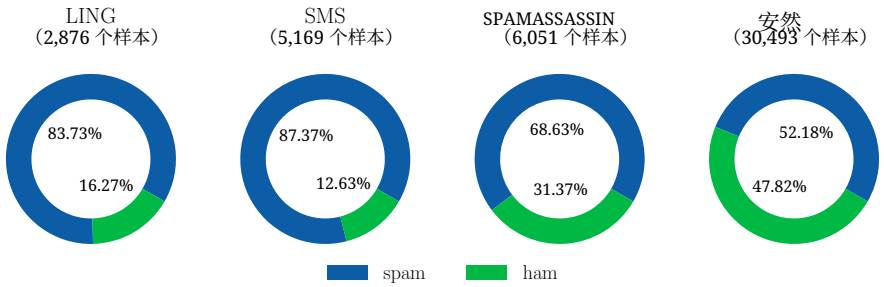

图2：各数据集中垃圾邮件和正常邮件的分布。值得注意的是，LING、短信和 SpamAssassin 数据集呈现出不平衡学习的情形，垃圾邮件数量多于正常邮件。相比之下，安然数据集中的垃圾邮件和正常邮件分布均衡。

4.2 大语言模型

在本实验中，我们评估了三个大语言模型（RoBERTa、SetFit 和 Flan-T5），它们分别属于三种不同的架构系列（类 BERT、句子转换器、序列到序列）。

RoBERTa。 RoBERTa（2019年）是一种预训练自然语言处理模型 [24]，它建立在 BERT [14] 的成功基础之上。RoBERTa 通过调整关键超参数来改进 BERT，例如移除下一句

[6] https://www.cs.cmu.edu/~安然/

预训练目标，并使用更大的小批量和更高的学习率进行训练。此外，RoBERTa 还探索了使用比 BERT 更多的数据进行训练，并延长训练时间。

我们使用了 HuggingFace 对 RoBERTa 模型（roberta-base）的实现，并采用基于字节级别的字节对编码（BPE）分词器 [31]。

SetFit。 SetFit（2022年）是一种高效且无需提示的框架，用于对句子转换器（ST）进行少样本微调 [34, 45]。通过以对比孪生方式对预训练的 ST 使用少量文本对进行微调，SetFit 能生成丰富的文本嵌入，用于训练分类头。这种方法无需提示或 verbalizer，且仅需远少于现有技术的参数即可实现较高的准确率。

我们结合 setfit库[7],中的 SetFit 实现与 sentence-transformers 中的 MPNet 实现（sentence-transformers/all-mpnet-base-v2） [40],，并使用 WordPiece分词器 [51]。我们的实现会生成 20 个训练文本对。通过余弦相似度测量所得嵌入之间的距离。

Flan-T5. Flan-T5（2022年）是一系列基于 T5（2019年）的模型， [33], 采用在多种语言任务上训练的编码器-解码器Transformer架构。Flan-T5 模型经过了 1,800 多项语言任务的指令微调，推理能力和提示适应性因此显著提升。不过，值得注意的是，Flan-T5 模型并未经过垃圾邮件检测任务的训练。

我们使用了 HuggingFace 对 Flan-T5 模型（google/ Flan-t5-base）的实现以及 SentencePiece分词器 [22]。我们的实验包括 Flan-T5 模型的小型版本（8,000 万个参数），但该版本展现出的泛化能力有限，因此未纳入本研究。

Flan-T5 模型是一种序列到序列模型，能够生成文本输出，而不是二元标签或概率。为了利用该模型进行垃圾邮件检测，我们将其微调为一项新任务，并为每个样本添加专用前缀“classify as ham or spam:”。因此，模型经过训练后能够根据输入文本正确输出“正常邮件”或“垃圾邮件”。为了获得分类指标的数值，我们在后处理步骤中将文本标签映射为 0 和 1。我们将这个针对垃圾邮件检测任务微调后的改进模型称为 $Spam\text{-}T5$ 。

微调详情。 我们发现，批量大小、学习率和训练轮数都是重要的超参数，会影响模型对新数据的泛化能力以及训练过程中的收敛速度。表1列出了用于微调这些模型的具体取值。

[7] https://github.com/huggingface/setfit

表1： 表格标题应置于表格上方。

模型	训练批次大小	评估批次大小	LR	训练轮数
RoBERTa	16	8	5e-5	10
SetFit	16	16	2e-5	3
Spam-T5	8	8	5e-5	5

4.3 基线技术

我们选取了6个在垃圾邮件检测中表现良好的基线模型：朴素贝叶斯 (NB)、逻辑回归 (LR)、K近邻 (KNN)、支持向量机、XGBoost 和 LightGBM。

朴素贝叶斯是一种概率模型，假设特征之间相互独立，因此在大型数据集上速度快、效率高。逻辑回归是一种线性模型，使用 Sigmoid函数来预测二元结果。K近邻是一种非参数模型，根据数据与其邻居的接近程度进行分类。在我们的实现中，我们选择了一个邻居（$K = 1$）。支持向量机是一种线性模型，通过寻找最优超平面将数据分为不同类别。我们使用 Sigmoid 核函数实现该模型，并将 gamma 设为 1。0。XGBoost 是一种高性能的梯度提升实现，它采用按层策略并行构建决策树，并优化训练集中的分裂质量。我们将其学习率设为 0。01，并设置了 150 个估计器。LightGBM 是另一种梯度提升框架，具备 XGBoost 的许多优点，但其决策树采用按叶构建方式，并使用高度优化的基于直方图的算法。同样，我们将其学习率设为 0。01，并设置了 20 个叶子。

微调详情。 我们发现，tf-idf 特征的最佳数量取决于模型。为确定这一数量，我们采用了分层五折交叉验证方法。具体而言，我们使用不同数量的 tf-idf 特征训练模型，并选取性能最高时对应的特征数量。表2展示了各模型的最佳特征数量。

表2： tf-idf 算法为每个模型生成的特征数量。这些数值是通过在验证集上进行网格搜索得出的。

模型	NB	LR	KNN	SVM	XGBoost	LightGBM
# 个特征	1000	500	150	3000	2000	3000

表 3： 各数据集上六种基线模型和三个大语言模型的测试集 F1 分数、精确率和召回率表现。

	Ling			SMS			SA			安然		
模型	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
NB	**1.00**	**1.00**	**1.00**	0.89	0.82	**0.98**	0.87	0.83	0.91	0.96	0.96	0.96
LR	0.98	0.96	**1.00**	0.87	0.78	**0.98**	0.92	0.89	0.96	0.97	0.98	0.96
KNN	0.93	0.96	0.90	0.81	0.74	0.89	0.92	0.88	0.95	0.91	0.94	0.89
SVM	**1.00**	**1.00**	**1.00**	0.90	0.83	**0.98**	0.94	0.92	0.97	0.98	**0.99**	0.97
XGBoost	0.92	0.94	0.90	0.78	0.65	**0.98**	0.94	0.92	0.96	0.91	0.98	0.85
LightGBM	0.95	0.96	0.94	0.87	0.82	0.93	**0.98**	**0.98**	**0.98**	0.98	**0.99**	0.96
RoBERTa	0.97	0.98	0.96	0.95	**0.97**	0.92	0.97	0.98	0.95	**0.99**	**0.99**	0.99
SetFit	0.99	0.98	**1.00**	**0.96**	**0.97**	0.95	0.95	0.96	0.94	—	—	—
Spam-T5	0.99	0.98	**1.00**	0.95	**0.97**	0.94	0.96	0.96	0.96	**0.99**	**0.99**	**1.00**

† 我们排除了 SetFit 模型在完整安然训练集上的结果，因为经过 104 小时的训练后，该模型仍未取得有意义的结果。

5 结果

5.1 完整训练集

我们评估了机器学习模型和大语言模型在能够完整访问训练集时的性能。此处的完整训练集指整个数据集的 80%。为了评估各模型的性能，我们采用了三项指标：F1 分数（F1）、精确率（P）和召回率（R）。评估结果见表 3。

我们发现，大语言模型在短信数据集和安然数据集上的表现优于基线模型。然而，我们也观察到，LLM 在 LING 和 SpamAssassin 数据集上的表现并不优于基线模型。在所有数据集中，就 F1 分数而言，短信数据集上表现最佳的基线模型与表现最差的大语言模型之间差异最大，相差 0.05 分。

我们的实验表明，Spam-T5 模型的总体表现最佳，平均 F1 分数为 0.9742。RoBERTa 和 SetFit 模型也以相同的 0.9670 分超过了基线模型。在基线模型中，支持向量机方法表现最佳，平均 F1 分数达到 0.9560。相反，XGBoost 模型效果最差，平均 F1 分数为 0.8842。这些结果表明，在大多数垃圾邮件检测场景中，大语言模型优于传统机器学习算法。

5.2 少样本学习

在少样本学习设置下，我们评估了每个模型在使用 $k \in \{4, 8, 16, 32, 64, 128, 256,$ Full$\}$ 样本训练后的性能。我们的分析结果见表 4。

表 4： 使用不同训练样本数时各模型的测试 F1 分数（四个数据集的宏平均）。“Full”列对应完整训练集（原始数据集的 80%）。

	训练样本数							
模型	**4**	**8**	**16**	**32**	**64**	**128**	**256**	**Full**
NB	0.145	0.210	0.211	0.243	0.361	0.505	0.663	0.930
LR	0.153	0.195	0.210	0.248	0.353	0.420	0.599	0.927
KNN	0.516	0.523	0.596	0.591	0.603	0.688	0.733	0.887
SVM	0.155	0.267	0.288	0.334	0.531	0.732	0.858	0.952
XGBoost	0.000	0.079	0.351	0.431	0.600	0.666	0.767	0.877
LightGBM	0.000	0.000	0.000	0.000	0.455	0.608	0.703	0.948
RoBERTa	0.241	0.174	0.575	0.738	0.459	0.915	0.929	0.970
SetFit	0.215	0.339	0.557	**0.855**	**0.887**	**0.929**	**0.941**	0.967
Spam-T5	**0.544**	**0.534**	**0.619**	0.726	0.806	0.864	0.933	**0.974**

NB、逻辑回归和支持向量机的性能特征相似，在较大的数据集上均有明显提升。相比之下，K近邻在较小的训练集上取得了相对更高的 F1 分数，当样本数为 4 和 8 时，分数超过 0.5。然而，随着样本数增加，其性能趋于停滞，在完整数据集上的最高分数为 0.887。梯度提升树模型（如 XGBoost 和 LightGBM）在最小的数据集规模（4、8、16 和 32）上表现不佳。随着训练集规模增加，其性能迅速提升，最终在完整数据集上分别达到 0.877 和 0.948。

RoBERTa 的表现随着训练集规模变化而有些不稳定：规模为 4 时为 0.241，规模为 8 时降至 0.174，而在完整数据集上则升至 0.970。相比之下，随着样本数量增加，SetFit 的表现持续提升，在完整数据集上取得了 0.967 的 F1 分数。该模型在数据集规模为 32、64、128 和 256 时表现最佳，这表明 SetFit 更适合这种特定的“中等”少样本学习。另一方面，Spam-T5 是极少样本场景下表现最佳的模型，即，当可用于训练的样本只有 4–16 个时。随着样本增多，其表现稳步提升，并在完整数据集上取得最高的 F1 分数。

图3展示了在所有训练样本规模下，大语言模型的 F1 分数始终优于基线模型。此外，表5列出了每个模型的平均 F1 分数，并显示 Spam-T5 的整体表现最佳，F1 分数为 0.7498。这些结果可归因于 Spam-T5 在极少样本设置下的高准确率，以及其在所有样本规模下始终稳健的表现。

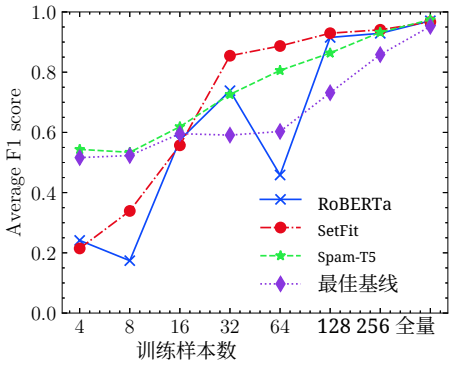

图 3： 不同训练样本数量下，大语言模型与最佳基线模型的测试集 F1 分数对比（四个数据集的宏平均）。

模型	F1 分数
NB	$0.4085_{.2734}$
LR	$0.3880_{.2621}$
KNN	$0.6421_{.1234}$
SVM	$0.5146_{.3005}$
XGBoost	$0.4716_{.3159}$
LightGBM	$0.3392_{.3871}$
RoBERTa	$0.6253_{.3139}$
SetFit	$0.7112_{.2990}$
Spam-T5	$\mathbf{0.7498}_{.1718}$

表5： 在所有训练样本数量下的平均测试集 F1 分数和标准差（四个数据集的宏平均）。

6 讨论

我们的实验结果揭示了大语言模型在垃圾邮件检测方面的优势与局限。具体而言，我们的发现表明，RoBERTa 和 MPNet（SetFit 使用的模型）等大语言模型总体表现良好，但在极少样本设置下不及 Spam-T5。这种性能差异可归因于参数数量：Spam-T5 拥有 2.5 亿个参数，而 RoBERTa 和 MPNet 分别拥有 1.25 亿和 1.1 亿个参数。

此外，我们的结果表明，使用大语言模型和基线技术进行垃圾邮件检测时，准确率与运行时间之间存在明显的权衡，如图 4 所示。虽然大语言模型更稳健，且在大多数情况下表现更好，但其训练和推理耗时较长。相比之下，基线技术速度更快，但准确率无法达到同等水平。这表明，大语言模型以牺牲计算效率为代价，提升了样本效率。这种权衡凸显了考虑任务具体需求的必要性，例如可用的计算资源以及期望的准确率水平。

图 4 还显示，随着训练样本数增加（即，测试样本数减少），基线模型的推理时间反而意外增加。这一趋势有悖直觉，因为我们预计其趋势会与大语言模型相似：由于需要处理的样本更少，推理时间应当缩短。这是由支持向量机模型使用的 sigmoid 核函数导致的。随着训练样本数增加，sigmoid 核函数在推理阶段需要更多计算，从而导致推理时间如观察到的那样增加。

大语言模型的实际应用受到其训练和部署所需大量计算资源的制约，这要求使用包括图形处理器和张量处理器在内的专用硬件。针对这一局限的研究

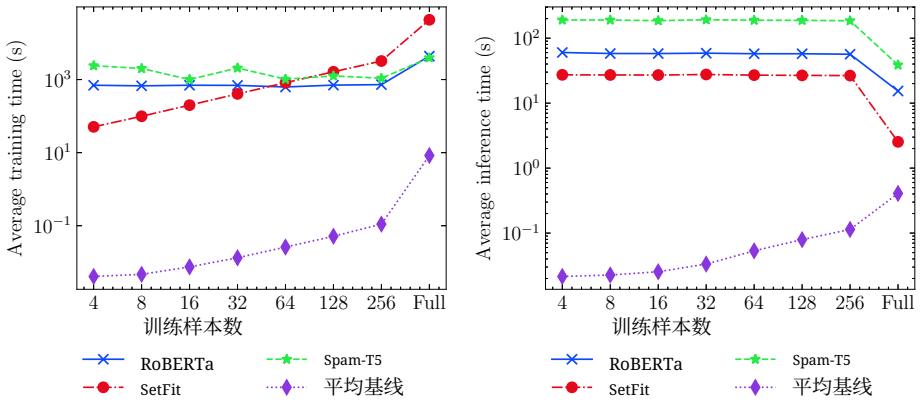

图 4： 三种大语言模型的训练和推理时间，以及基线技术的平均时间（基于四个数据集计算的宏平均）。此处，“推理时间”指处理整个测试集所需的执行时间。由于 SetFit 模型在完整的安然数据集上训练 104 小时后仍未取得有意义的得分，因此我们不计入其训练和推理时间。

正不断推进，已有许多技术可以减少大语言模型所需的内存占用和计算资源。例如，一些方法探索使用 8 位浮点格式来降低内存需求，如 [13] 所示。其他方法，例如 LoRA [19]，则旨在减少训练和部署大语言模型所需的计算资源。

7 结论

我们的研究展示了大语言模型在电子邮件垃圾邮件检测中的有效性，尤其是在少样本场景中。实验表明，在所有训练样本规模下，大语言模型的 F1 分数均优于成熟的基线技术。此外，我们的解决方案 Spam-T5 整体表现最佳，平均 F1 分数达到 0.7498。

这些发现表明，大语言模型，尤其是 Spam-T5，能够成为应对垃圾邮件检测持续挑战的有力工具。只需加入有限数量的欺诈样本，我们便能更新模型并提升其性能，而无需投入大量精力进行数据标注。这种方法简化了构建稳健模型的过程，使其能够应对动态数据分布，从而为现实世界的问题提供实用且有效的解决方案。

为了将大语言模型部署到现实应用中，未来的工作需要着重降低这些模型的计算需求。实现这一目标的一种方法是开发相关技术，尽量减少大语言模型所需的内存占用和计算资源，例如近期研究中探索的方法。

8 伦理声明

垃圾邮件检测是电子邮件系统和其他通信平台的重要组成部分，在这一领域使用有效的语言模型可以实现更准确的过滤并改善用户体验。然而，我们的研究引发了关于此类语言模型可能被滥用于审查的伦理担忧。将消息分类为垃圾邮件或非垃圾邮件的能力可能被用来压制或过滤不符合某些政治或社会议程的内容。我们认识到，防范此类滥用并推动在公共领域负责任地使用机器学习工具十分重要。

此外，大语言模型的开发与部署会对环境产生显著影响，消耗大量能源并增加碳排放。我们承认本研究可能带来的生态后果，并呼吁更加关注机器学习模型及其应用的环境可持续性。

参考文献

1. Agboola, O.：使用机器学习和深度学习进行垃圾邮件检测。路易斯安那州立大学博士论文（2022年） 2. Almeida, T.A., Hidalgo, J.M.G., Yamakami, A.：对短信垃圾邮件过滤的研究贡献：新数据集及研究结果。载于：第11届ACM文档工程研讨会论文集。第259–262页。DocEng '11，计算机协会，纽约，美国纽约州（2011年）。https://doi.org/10.1145/2034691.2034742, https://doi.org/10.1145/2034691.2034742 3. Alspector, J.：采用特定内容误分类代价的支持向量机垃圾邮件过滤。载于：文本挖掘研讨会论文集（2001年） 4. Androutsopoulos, I., Paliouras, G., Michelakis, E., Michelakis, E.：学习过滤未经请求的商业电子邮件（2006年） 5. Attenberg, J., Weinberger, K., Dasgupta, A., Smola, A., Zinkevich, M.：利用哈希技巧进行协作式电子邮件垃圾邮件过滤。载于：第六届电子邮件与反垃圾邮件会议（CEAS 2009）论文集（2009年1月） 6. Carreras, X., Salgado, J.：用于反垃圾邮件过滤的提升树。arXivcs.CL/0109015（2001年10月） 7. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.：扩展指令微调语言模型（2022年） 8. Cortes, C., Vapnik, V.：支持向量网络。《机器学习》**20**(3)，273–297（1995年） 9. Cranor, L.F., LaMacchia, B.A.：垃圾邮件！《ACM通讯》**41**(8)，74–83（1998年8月）。https://doi.org/10.1145/280324.280336, https://doi.org/10.1145/ 280324.280336

10. Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O., Ajibuwa, O.E.：电子邮件垃圾过滤的机器学习：综述、方法及尚待解决的研究问题。《Heliyon》**5**(6)，e01802（2019年）。https://doi.org/https://doi.org/10.1016/j.heliyon.2019.e01802, https://www.sciencedirect.com/science/article/pii/S2405844018353404 11. Debnath, K., Kar, N.：采用深度学习方法进行电子邮件垃圾邮件检测。载于：2022年机器学习、大数据、云计算与并行计算国际会议（COM-IT-CON）。第1卷，第37–41页（2022年）。https://doi.org/10.1109/COM- IT-CON54601.2022.9850588 12. Delany, S., Cunningham, P., Coyle, L.：基于案例的推理在垃圾邮件过滤中的评估。《人工智能评论》**24**，359–378（2005年11月）。https://doi.org/10 6 13 1007 10462 005 9006 . /s - - - . Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.：Llm.int8()：面向大规模Transformer的8位矩阵乘法（2022年） 14. 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（2019年） 15. Drucker, H., Wu, D., Vapnik, V.：用于垃圾邮件分类的支持向量机。《IEEE神经网络汇刊》**10**(5)，1048–1054（1999年）。https://doi.org/10 788645 16 1109 72. / . . Harris, Z.S.：分布结构。《WORD》**10**(2-3)，146–162（1954年）。https://doi.org/10.1080/00437956.1954.11659520, https://doi.org/10.1080/00437956.1954.11659520 17. Hochreiter, S., Schmidhuber, J.：长短期记忆。《神经计算》**9**(8)，1735–1780（1997年） 18. Hovold, J.：利用基于词位置的属性进行朴素贝叶斯垃圾邮件过滤。载于：第二届电子邮件与反垃圾邮件会议论文集，http://www. ceas. cc（2005年） 19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.：LoRA：大语言模型的低秩适配（2021年） 20. J´a˜nez-Martino, F., Alaiz, R., Gonz´alez-Castro, V., Fidalgo, E., Alegre, E.：垃圾邮件检测综述：垃圾邮件发送者策略与数据集偏移问题分析。《人工智能评论》**56**（2022年5月）。https://doi.org/10.1007/s10462- 022-10195-4 21. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.：文本分类算法：综述。《Information》**10**(4)（2019年）。https://doi.org/10.3390/info10040150, https://www.mdpi.com/2078-2489/10/4/ 150 22. Kudo, T., Richardson, J.：SentencePiece：一种简单且与语言无关的子词分词器和反分词器，用于神经文本处理（2018年） 23. Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L.：文本分类综述：从传统方法到深度学习。《ACM智能系统与技术汇刊》**13**3495162 24 2 2022 10 1145 ( )（2022年4月），https://doi.org/ . / . Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.：RoBERTa：一种稳健优化的BERT预训练方法（ 2019 1907 11692, 1907 11692 ），25 ´ ´ http://arxiv.org/abs/ . cite arxiv: . . Mendez, J.R., Fdez-Riverola, F., Dıaz, F., Iglesias, E.L., Corchado, J.M.：反垃圾邮件过滤领域特征选择方法的比较性能研究。载于：Perner, P.（编）《数据挖掘进展》。医学应用、Web挖掘、营销、图像与信号挖掘。第106–120页。Springer Berlin Heidelberg，德国柏林（2006年）

26. Metsis, V., Androutsopoulos, I., Paliouras, G.：使用朴素贝叶斯进行垃圾邮件过滤——哪种朴素贝叶斯？（2006年），http://citeseer.ist.psu.edu/757874.html 27. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.：单词和短语的分布式表示及其组合性（2013年）28. OpenAI：GPT-4 技术报告（2023年）29. Porter, M.：一种用于后缀剥离的算法。《Program》**14**（3），130–137（1980年）。https://doi.org/10.1108/eb046814, http://www.emeraldinsight.com/doi/abs/10.1108/eb046814 30. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.：通过生成式预训练提升语言理解（2018年）31. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.：语言模型是无监督多任务学习器（2018年），https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf 32. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.：探索统一文本到文本转换器的迁移学习极限。《机器学习研究杂志》**21**（140），1–67（2020年），http://jmlr.org/papers/v21/20-074.html 33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.：探索统一文本到文本转换器的迁移学习极限（2020年）34. Reimers, N., Gurevych, I.：Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。载于：Inui, K., Jiang, J., Ng, V., Wan, X.（编），EMNLP/IJCNLP（1），第3980–3990页。计算语言学协会（2019年），http://dblp.uni-trier.de/db/conf/emnlp/emnlp2019 19 35 1- .html#ReimersG。Saha mi, M., Dumais, S., Heckerman, D., Horvitz, E.：一种过滤垃圾邮件的贝叶斯方法。载于：AAAI-98 文本分类学习研讨会，第55–62页（1998年）36. Sahmoud, T., Mikki, D.M.：使用 BERT 进行垃圾邮件检测（2022年）37. Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., Stamatopoulos, P.：一种用于邮件列表反垃圾邮件过滤的基于记忆的方法。《信息检索》**6**（1），49–73（2003年1月）。https://doi.org/10.1023/A:1022948414856，https://doi.org/10 1022948414856 38 1023. / A:。Scholz, M., Klinkenberg, R.：一种用于处理概念漂移的集成分类器。载于：第二届数据流知识发现国际研讨会论文集（IWKDDS'05）（2005年）39. Shi, W., Xie, M., Huang, Y.：一种基于 MIME 指纹的协同垃圾邮件过滤技术。载于：2011年第9届世界智能控制与自动化大会，第225 5970733 40 230 2011 10 1109 2011 –（）。https://doi.org/ . /WCICA. . . Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.：MPNet：用于语言理解的掩码与置换预训练（2020年）41. Syed, N.A., Liu, H., Sung, K.K.：使用支持向量机处理增量学习中的概念漂移。载于：第五届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第317–321页。KDD ' 99，计算机协会，美国纽约州纽约市（1999年）。https://doi.org/10.1145/312129.312267, https://doi.org/10.1145/312129.312267 42. Tida, V.S., Hsu, S.：使用 BERT 模型迁移学习进行通用垃圾邮件检测（2022年）43. Torabi, Z., Nadimi-Shahraki, M.H., Nabiollahi, A.：用于垃圾邮件检测的高效支持向量机：一篇综述。《国际计算机科学与信息安全杂志》，第13卷，第1期，2015年1月**13** （2015年1月）

44. Tsymbal, A.：概念漂移问题：定义及相关研究。技术报告 TCD-CS-2004-15，都柏林大学三一学院计算机科学系，爱尔兰都柏林（2004年）45. Tunstall, L., Reimers, N., Jo, U.E.S., Bates, L., Korat, D., Wasserblat, M., Pereg, O.：无需提示的高效少样本学习（2022年）46. Vapnik, V.N.：《统计学习理论的本质》。Springer-Verlag New York, Inc.（1995年）47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.：Attention is all you need。载于：Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R.（编），《神经信息处理系统进展》，第30卷。Curran Associates, Inc.（2017年），https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf 48. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.：从少量示例中进行泛化：少样本学习综述。《ACM 计算调查》**53**（3）（2020年6月）。https://doi.org/10 3386252 49 1145 3386252，10 1145 . / https://doi.org/ . Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.：微调语言模型是零样本学习器。载于：国际学习表征会议（2021年）50. Wittel, G., Wu, S.：论对统计垃圾邮件过滤器的攻击。载于：电子邮件与反垃圾邮件会议论文集（CEAS）。美国加利福尼亚州山景城（2004年7月），http://www.ceas.cc/papers-2004/index.html 51. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., ˙Lukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.：谷歌神经机器翻译系统：弥合人工翻译与机器翻译之间的差距（2016年）