CTDGSI：全面利用实例选择方法进行自动文本分类

VII Concurso de Teses, Dissertações e Trabalhos de Graduação em SI
XXI Simpósio Brasileiro de Sistemas de Informação

Washington Cunha[1], Leonardo Rocha[2] (Co-advisor), Marcos Gonçalves[1] (Advisor)

[1]Department of Computer Science – Federal University of Minas Gerais – Brazil

[2]Department of Computer Science – Federal University of São João del Rei – Brazil

{ washingtoncunha,mgoncalv } @dcc.ufmg.br, lcrocha@ufsj.edu.br

Abstract. 自然语言处理（NLP）的进步一直由“更多”的规则主导：更多的数据，更多的计算能力和更复杂的模型，最佳的例子就是大型语言模型。然而，为特定应用训练（或微调）大型稠密模型通常需要大量的计算资源。本文博士论文聚焦于一个未被充分研究的 NLP 数据工程技术，其在当前情况下具有巨大潜力，即实例选择（IS）。IS 的目标是通过去除噪声或冗余实例来减少训练集的大小，同时保持已训练模型的有效性并降低训练过程的成本。我们对应用于一个重要的 NLP 任务——自动文本分类（ATC）的 IS 方法进行了全面且科学合理的比较，考虑了若干分类解决方案和多种数据集。我们的研究结果揭示了 IS 解决方案中尚未开发的巨大潜力。我们还提出了两个新颖的 IS 解决方案，这些方案注重于噪声和冗余，特别设计用于大型数据集和变压器架构。我们的最终解决方案在所有数据集中保持相同的有效性水平的同时，实现了训练集平均减少 41 %。更重要的是，我们的解决方案展示了 1.67 倍（最高 2.46 倍）的加速改进，使其能够扩展到包含数十万份文档的数据集。

学生等级：博士，预计完成日期为 2024 年 8 月 26 日

董事会成员：Franco Maria Nardini（ISTI-CNR）；Thierson Couto Rosa（UFG）；Rodrygo Luis Teodoro Santos（DCC-UFMG）；Anisio Mendes Lacerda（DCC-UFMG）

论文可在以下网址获取：http://hdl.handle.net/1843/76441

出版物可在以下网址查阅：http://bit.ly/3WvX1T5

1. 引言

数据可用性的指数级增长使得有效组织和检索内容变得具有挑战性。在这种背景下，自动文本分类（ATC）通过将文本文档（例如网页、电子邮件、评论、推文、社交媒体消息）映射到预定义的兴趣类别中，提供了解决方案。实际上，ATC 模型对于虚假新闻检测、仇恨言论识别、相关反馈、情感分析、产品评论分析、选举投票推断以及评估公共机构满意度等新兴应用至关重要。作为一项监督任务，ATC 受益于生成大量标注数据的应用程序，例如像 Twitter 这样的社交网络。此外，众包和软标签进一步降低了获取标注数据集的成本。

基于 Transformer 的架构，包括 RoBERTa、GPT-4 和 LLama3 等小型和大型语言模型（SLMs 和 LLMs），代表了 ATC 领域的最新技术水平（SOTA），通过在大型数据集上进行预训练和在特定领域任务上进行微调，取得了卓越的成果。虽然零样本能力是可行的，但微调仍然是实现最佳性能的关键，通过利用预训练的权重进行任务特定的调整 [de Andrade et al. 2023] 。

根据 Andrew Ng 的说法，这些（语言）模型的成功归因于在大规模数据集（例如，GPT-3 的 45TB）上进行的大量预训练，以及通过微调实现的预训练模型的适应性。与从头开始相比，这种方法能够实现更快的特定任务训练 [Uppaal et al. 2023] 。然而，尽管比全训练更快，微调仍然是资源密集型的，需要显著的计算能力和时间。例如，在我们的论文中考虑的其中一个数据集 XLNet 在 MEDLINE 数据集上的微调耗费了 80 个 GPU 小时。因此，实际限制常常会延迟再训练，影响在需要持续更新的场景中的模型性能，例如欺诈检测和推荐系统。

确实，公司和研究小组的资源限制限制了深度学习模型的实验，这些模型需要广泛的计算资源。例如，在这篇论文中，我们进行了 4000 个实验，耗时约 5600 小时。降低经济、计算和环境成本至关重要，因为大型模型的显著能耗和碳排放已经显示出来。鉴于数据量不断增加、重新训练的要求以及环境问题，提出可扩展和成本效益高的策略是必要的。这些策略包括创建高效的深度学习算法、使用先进的硬件或改进数据预处理技术。这篇论文侧重于后者，旨在提高模型性能同时降低成本。

在 [Cunha et al. 2020] 中，我们提出利用一组预处理技术，在数据转换流程中建立成本效益高的模型。这个解决方案在更低的成本下（例如构建 ATC 模型所用的时间更短）提高了有效性（例如模型具有更高的准确性）。这项工作主要贡献之一是在流程中明确引入了实例选择（IS）阶段，这是一组有前途的技术，正在成长的研究领域，有助于处理许多上述问题。

具体而言，与传统的特征选择方法 [Lu et al. 2017] 相比，其主要目标是选择最具信息量的术语（单词），实例选择方法则专注于为训练集选择最具代表性的实例（文档）[Garcia et al. 2012] 。这类算法的直觉是从原始训练集中删除可能产生噪声或冗余的实例，并在保持或甚至提高有效性的同时，在总训练时间方面提高性能。更具体地说，IS 方法有三个主要目标：（i）通过选择最具代表性的实例来减少实例数量；（ii）通过去除噪声和冗余来保持（或甚至提高）有效性；以及（iii）减少应用端到端模型的总时间（包括从传统预处理步骤到模型训练步骤）。通过选择最具代表性的实例，IS 方法还可以潜在地去除错误注释中的噪声。根据这些目标，IS 方法必须遵循三个基本约束，即在不损失有效性的情况下减少训练量并提高效率。IS 方法寻求同时优化这三项约束。

尽管它们有潜力，IS 方法在 ATC 中的研究仍不足，特别是在深度学习中 [Barigou 2018, Tsai et al. 2014] 。为其他领域开发的传统 IS 方法包括 CNN [Hart 1968] 、ENN [Wilson 1972] 和 Drop3 [Wilson and Martinez 2000] 。最近的方法，如 LSSm、LSBo [Leyva et al. 2015] 和 PSDSP [Carbonera and Abel 2018] ，主要是在小型表格式数据集上使用弱分类器如 KNN 进行测试。相比之下，文本分类数据集是非结构化的、更大并且更复杂，具有高维性和偏斜性。考虑到深度学习变换模型在处理大型训练数据集时的高计算成本，展示了应用 IS 技术的理想场景。

2. 假设和研究问题

这篇博士论文的主要假设 (H1) 是:

> H1: It is possible to simultaneously reduce data, maintain model quality, and improve time for fine-tuning ATC models through IS methods.

为了验证这个假设，我们为博士论文提出了三个研究问题。总而言之，RQ1 的目标是在传统 IS 方法的背景下评估所提出的 IS 方法在 ATC 任务中的三脚架约束：简化、有效性和效率，该部分将在第 ?? 节中呈现。接下来，考虑到文献中的空白，RQ2 的目标是证明提出一种新的 IS 框架的可行性，并展示其如何适应不同

场景中提出的不同要求，主要是与大数据相关的那些。最后，RQ3 旨在研究每个 IS 方法处理和有效去除噪声实例的能力。这个问题的答案也促使我们展示提出一种新的扩展 IS 框架的可行性，该框架能够同时从训练集中移除冗余和噪声实例。接下来，我们介绍这篇博士论文中考虑的每个 RQ。

RQ1. 在 ATC 环境中应用传统 IS 方法对于所提出约束的影响是什么？RQ1 旨在评估 ATC 任务背景下的传统 IS 方法，重点关注所提出的 IS 方法的三脚架约束：减少、有效性和效率。

RQ2. 一种专注于冗余消除的新型实例选择方法能否克服现有 IS 方法的局限性，以实现 ATC 场景中的三脚架限制？该 RQ 旨在证明提出新型 IS 框架的可行性，并表明它如何适应不同场景所提出的不同要求，主要是与大数据相关的场景。

RQ3. 是否有可能扩展之前的提议，不仅去除冗余，还去除噪音，从而在考虑所有三脚架标准时提高质量水平？此 RQ 的目的是证明提出一种新型扩展 IS 框架的可行性，该框架能够同时从训练集中移除冗余和噪声实例。

3. 出版物

我们在实例选择领域的工作已在过去四年间被验证并发表在主要的信息系统会议和期刊上，其中包括在《信息处理与管理》(IP & M)（h5 指数：96，影响因子：7.4，A1）中发表的两篇论文，这是信息检索领域的全球领先期刊 [Cunha et al. 2021, Cunha et al. 2020] 。我们还有一篇发表在《ACM 计算机调查》（CSUR）（h5 指数：103，IF：23.8，A1）中的文章，该期刊在计算机科学理论和方法领域处于领先地位，涵盖了传统和最新方法在 IS 领域的综合文献综述以及这些方法之间的广泛实验对比；国际会议 SIGIR [Cunha et al. 2023a] ，介绍了我们首次提出的新颖的冗余导向 IS 框架，针对大型数据集，特别关注变压器。最后，这篇论文还导致在 ACM 信息系统事务 (TOIS) [Cunha et al. 2024] 上发表了一篇文章（h5 指数：48，IF：5.6，A1），这是信息系统领域的全球领先期刊，包括针对 ATC 的扩展噪声导向和冗余感知 IS 框架的提议。

1. Cunha, Washington 等人。“一种面向噪声和冗余感知的实例选择框架。”ACM 信息系统交易（ACM TOIS）（2024）– h5 指数：48.0
2. Cunha, Washington 等人. “一种有效、高效且可扩展的基于置信度的实例选择框架，用于基于 transformer 的文本分类。”ACM SIGIR 2023 ——h5-index: 103.0。
3. Cunha, Washington 等人。“应用于非神经和基于 Transformer 的文本分类的实例选择方法的比较研究。”ACM Computing Surveys (2023) – h5-idx: 157.0
4. 库尼亚, 华盛顿, 等人. “关于神经和非神经方法及表示用于文本分类的成本效益：一项全面的比较研究。”IP & M (2021) - h5-index: 114.0
5. Cunha, Washington 等. “文本分类的扩展预处理流程: 关于元特征表征、稀疏化和选择性采样的作用.” IP & M (2020) – h5-index: 114.0

我们的论文在攻读博士学位期间直接为其他期刊做出了贡献：

1. Gonçalves, M., Cunha, W 等. (2024) “以少胜多——基于数据工程和先进人工智能的智能、可持续的自然语言处理。”2025-2035 巴西计算机重大挑战研讨会 - 巴西计算机学会。
2. França，C.，Cunha，W 等人。(2024)。关于基于表示学习的方法有效、高效和可扩展的代码检索。神经计算。– h5-指数：136.0
3. Andrade, C., Cunha, W (2023) 关于上下文嵌入表示的类别可分性——或者“当（文本）表示如此优秀时，分类器并不重要!”。IP & M . – h5-index: 114.0
4. Zanotto, B. S., Cunha, W 等人 (2021)。通过机器学习对电子健康记录进行自动分类用于价值导向计划的 Pcv50。《价值在健康》。– h5-index: 57.0
5. Zanotto, B. S., Cunha, W 等人 (2021)。通过电子病历进行中风结果测量：关于神经和非神经分类器有效性的横断面研究。JMIR Med – h5-idx: 52.0
6. Viegas, F. 等人, Cunha, W (2024)。在层次主题建模中利用上下文嵌入并探讨当前评估指标的限制。计算语言学。– h5-idx:38

7. Felix, L. 等人，Cunha, W (2024)。你为什么旅行？——从在线评论和领域知识中推断旅行概况。在线社交网络与媒体。——h5 指数：28.0
8. Viegas, F., Cunha, W 等。(2024)。用于短文档情感分析的语义扩展与噪声过滤管道处理——clusent 方法。《互动系统期刊》。——h5-index: 9.0
9. Cunha, W 等人 (2021) 文本分类的扩展预处理管道：关于元特征、稀疏化和选择性采样的作用。CTDBD - SBBD'21 – h5-index: 7.0
10. Gomes, C., Cunha, W 等 (2019). CluWords: 探索词之间的语义簇以增强主题建模. CTIC - 计算机科学研究入门. – h5-index: 1.0

我们的论文直接为博士期间的其他会议论文做出了贡献：

在本节中，我们对实例选择（IS）领域中最传统和/或最近（即最新技术）的提案进行了批判性分析（又称为 快速（基于系统的）文献综述 ）。这次审查旨在全面评估与不同场景中应用的 IS 策略相关的最重要的工作。特别是，我们关注实验导向的研究，即那些有强大实验和实证成分来支持其发现的研究。为了实现我们的目标，我们收集了包含最多引用文献的 100 篇关于 IS 相关文章。我们假设被大量引用的文章可能具有影响力，因为它们受到了很多关注。在这些文献中，我们选择了最受欢迎的方法来包含在我们对应用于 ATC 的 IS 方法的实验评估中。我们还选择了一些最近提出的方法（被认为是最先进的），以完成我们的实验比较。在这个文献综述过程结束时，详细如下，我们最终得到一个由传统和最先进方法混合组成的集合，包括 13 种将在下一节 ATC 场景中评估的 IS 策略。一简化版快速综述过程如图 1 所示。首先，我们从提交给 Google Scholar 的四个查询返回的文章中收集文献。其次，我们使用每篇文章的唯一 URL 去除重复项（去重阶段）。此过程产生了 1740 篇独特文章。第三，我们按照引用次数对剩余文章进行排名。第四，我们按照期望标准对文章进行排名，并进行分类：(i) 与 IS 相关，以及 (ii) 进行方法之间的实验比较。第五，我们过滤掉不符合标准的文章，直到达到 100 篇相关文章。到这一步为止，我们已按排名顺序检查了 702 篇文章。从这些文章中，我们选择了最受欢迎和最新的方法进行我们的实验评估。

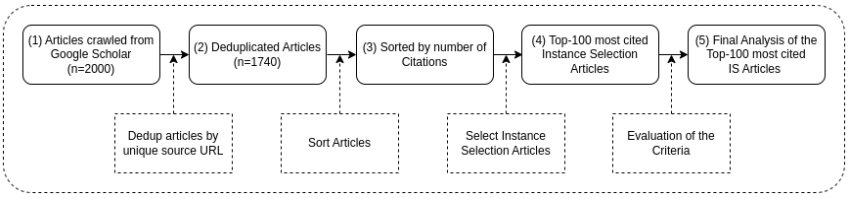

Figura 1. **研究选择的工作流程和实例选择方法（快速）系统文献综述的分析。**

我们的搜索和分析结果加强了我们的认知，即 IS 方法几乎被专门应用于表格结构的数据，而在自然语言处理任务中的应用却很少，这显得有些奇怪，因为这是可以从这种方法中收益最大的领域之一。总之，在考虑的 100 篇 IS 论文中，有 92 篇只考虑了表格数据。我们建议研究将 IS 方法与 ATC 模型结合使用，ATC 模型在诸如假新闻和仇恨言论检测、情感分析、产品特征修订、推断意见以及评估产品和服务的满意度等各种应用中都非常流行。我们还建议通过扩展一份 10 年前在 [Garcia et al. 2012] 提出的分类法，来创新性地构建一个新的 IS 策略分类法。先前的 IS 分类法是在一个不同的背景下提出的，其中考虑了不同类型的数据集（小表格型）和学习方法。例如，深度学习神经网络方法在当时提出的分类法的设想情境中并未被考虑。如前所述，自原始分类法提出以来，该领域已经显著发展。这些进展已被我们新扩展的分类法所仔细记录，并在我们的新扩展分类法中有所反映，其中包括三个新类别——在图 ?? 中以绿色表示——和八个新近提

出的方法——在图 ?? 中以黑色字母书写。新类别指的是自 2015 年以来提出的基于密度、空间超平面和聚类的研究进路。最后，在我们的博士论文中，我们也提供了对我们在批判性分析中证实的十三种 IS 方法的详细信息（由于篇幅限制，此处未展示），这些方法我们将在 ATC 环境下的实验评估中考虑。因此，在下一节中，我们提供一个完整的实验评估，比较上述详细的 IS 方法，并将它们与 SOTA ATC 策略相结合。我们进行这一组实验的目的是回答第一个研究问题（RQ1）。。

在本节中，我们提议评估这 13 种最具代表性的传统 IS 方法（第 ?? 节）在 ATC 任务中的简化、效率和效果之间的权衡，这些方法使用大型和多样化的数据集。需要注意的是，前面节中选择的 IS 方法仅在小型结构化表格数据集（如来自 UCI 库的数据集）中进行了测试。关于 ATC 方法，我们考虑了 ATC 领域当前的 SOTA：基于变压器的架构，如 BERT、XLNet、RoBERTa 等。这些方法在计算资源方面成本极高，特别是在处理大型标记训练数据时。因此，它们构成了 IS 技术应用的理想场景。最后但同样重要的是，我们应强调，据我们所知，我们的工作是首次将 IS 作为预处理步骤应用于 ATC 环境中的基于变压器的架构之前。这个贡献是通过一个实验研究实现的，其严谨性和规模（7 种变压器方法和 13 种 IS 方法）尚未在 IS 文献中报道过。。

接下来，我们展示了比较场景和实验设置，旨在评估这 13 种最具代表性的传统信息系统方法应用于 ATC 时，在简化、效率和有效性之间的权衡。

Task	Dataset	Size	Dim.	# Classes	Density	Skewness
Topic	DBLP	38,128	28,131	10	141	Imbalanced
	Books	33,594	46,382	8	269	Imbalanced
	ACM	24,897	48,867	11	65	Imbalanced
	20NG	18,846	97,401	20	96	Balanced
	OHSUMED	18,302	31,951	23	154	Imbalanced
	Reuters90	13,327	27,302	90	171	Extremely Imbalanced
	WOS-11967	11,967	25,567	33	195	Balanced
	WebKB	8,199	23,047	7	209	Imbalanced
	Twitter	6,997	8,135	6	28	Imbalanced
	TREC	5,952	3,032	6	10	Imbalanced
	WOS-5736	5,736	18,031	11	201	Balanced
Sentiment	SST1	11,855	9,015	5	19	Balanced
	pang_movie	10,662	17,290	2	21	Balanced
	Movie Review	10,662	9,070	2	21	Balanced
	vader_movie	10,568	16,827	2	19	Balanced
	MPQA	10,606	2,643	2	3	Imbalanced
	Subj	10,000	10,151	2	24	Balanced
	SST2	9,613	7,866	2	19	Balanced
	yelp_reviews	5,000	23,631	2	132	Balanced
Large	AGNews	127,600	39,837	4	37	Balanced
	Yelp_2013	335,018	62,964	6	152	Imbalanced
	MEDLINE	860,424	125,981	7	77	Extremely Imbalanced

Tabela 1. **数据集统计，包括：规模、维度、类别、文档密度和类别分布。**

数据集、数据表示与预处理 我们采用了 22 个真实世界的数据集，这些数据集来自两个广泛的 ATC 任务中的各种来源 [Li et al. 2022]：（i）主题分类；和（ii）情感分析。数据集涵盖了多个领域，大小、维数、类别、文档密度和类别分布的多样性。TFIDF 表示作为所有 IS 方法的输入。其他文本表示可以作为 IS 方法的输入，例如静态嵌入（例如，FastText [Joulin et al. 2017]）或由变压器架构构建的上下文嵌入（无论是通过微调模型传送文档还是使用零射击方法）。然而，正如之前在 [Andrade et al. 2024] 中所示：（i）静态嵌入可以显著降低分类方法的速度；以及（ii）直接使用上下文嵌入作为 IS 输入是低效且效果不佳的，可能是因为它牺牲了稀疏性。在这个背景下，考虑到寻找经济高效解决方案的目标，TF-IDF

表示仍然是实例选择方法输入的最优选择。

我们移除了停用词，并保留至少出现在两个文档中的特征。我们使用 L2 范数对 TF-IDF 的结果进行了标准化。实际上，如图 2 所示，我们首先使用分层 K 折交叉验证方法对数据集进行划分——对于较小的数据集，我们使用 k=10 折分割，而对于较大的数据集，由于过程的成本，我们采用 5 折分割——然后我们为 IS 阶段构建文档的 TFIDF 矩阵表示，然后使用选定的原始文档作为 Transformer 分类器的输入。

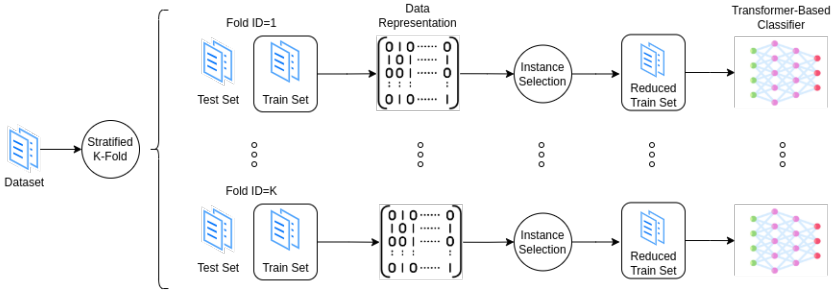

Fig. 1.4: **数据表示和预处理过程**

正如所提到的，我们的目标是研究和比较我们提出的方法与 SOTA（最先进的）IS 技术在 Transformers 架构背景下的表现——尤其是在多个领域中达到 SOTA 的分类。我们将比较以下几种最新版本的 Transformers（RoBERTa、BERT、DistilBERT、BART、AlBERT 和 XLNet）在所有测试数据集上的有效性。鉴于需要调优的大量超参数，不可能对所有参数进行网格搜索和交叉验证。因此，为了确定最佳超参数，我们应用了来自 [Cunha et al. 2021] 的方法。我们将初始学习率固定为 5×10^{-5} ，最大训练轮数设为 20 ，并将 5 轮作为耐心值。最后，我们对 max_len（150 和 256）和 batch_size（16 和 32 ）进行了网格搜索，因为这些指定的值直接影响效率和效果。

实例选择方法 在本章中，我们考虑了一组在第 ?? 节中描述的 13 种 IS 方法，具体为：凝聚最近邻（CNN）；编辑最近邻（ENN）；迭代案例过滤（ICF）；基于实例的 3（IB3）；递减缩减优化过程（Drop3）；基于局部集合的平滑器（LSSm）；局部集合边界选择器（LSBo）；基于局部密度的实例选择（LDIS）；基于中心密度的实例选择（CDIS）；扩展的基于局部密度的实例选择（XLDIS）；基于密集空间划分的原型选择（PSDSP）；增强的全局密度实例选择（EGDIS）；以及好奇实例选择（CIS）。所有 IS 方法的参数都通过网格搜索来定义，使用交叉验证在初步的实验轮次中进行训练集上的验证。

指标和实验协议 所有实验均在一台 Intel Core i7-5820K（6 核心，12 线程，主频 3.30GHz）计算机，64Gb RAM，GeForce GTX TITAN X（12GB）显卡，以及 Ubuntu 19.04 操作系统上执行。由于数据集的不对称性，我们使用宏平均 F1（MacroF1）来评估分类效果。我们采用 95 % 置信水平的配对 t 检验来比较我们交叉验证实验的平均结果。最后，我们应用了 Bonferroni 校正 [Hochberg 1988] 以处理多重检验问题。我们定义缩减平均为 $\overline{R} = \frac{\sum_{i=0}^{k} \frac{|T_i|-|S_i|}{|T_i|}}{k}$ ，其中 T 是原始训练集，S 是由被评估 IS 方法选择的实例组成的解决方案集。最后，为了分析成本效益权衡，我们还评估了每种方法在构建模型所需的总时间方面的成本。加速比计算为 $S = \frac{T_{wo}}{T_w}$ ，其中 T_w 是使用 IS 方法进行模型构建所花费的总时间，T_{wo} 是未使用 IS 阶段执行所花费的总时间。实验结果：我们的研究发现显示了 IS 的显著潜力。在某些情况下，这些方法可以在保持有效性的同时减少高达 90 % 的训练集。我们研

究的 IS 方法实现了 15.6 % (LSSm) 到 91.1 % (XLDIS) 的平均减少。然而，尽管有去除噪声的动机，选择方法也未能提高文本分类模型的有效性。这凸显了进一步研究这一问题的必要性，这可能对数据科学和机器学习领域产生深远影响。我们展示了三种传统的 IS 方法（LSSm, CNN, LSBo）能够在保持有效性的前提下，减少 19 个数据集中的 12 个数据集上文本分类模型的构建时间——加速比从 1.04x (CNN - Books) 到 5.69x (LSBo - Reuters90) 不等。在其他数据集中，我们观察到引入 IS 方法在生成模型的总时间上（运行 IS 方法 + 模型构建）产生了额外的开销，从计算成本的角度来看，使整个过程更加昂贵。综合考虑这三个制约因素和所有数据集，最佳的传统 IS 方法是 CNN。在回答 RQ1 时，我们对制约因素（减少 - 效率 - 有效性）的评估表明，在某些数据集中，特定的选择方法可以在不损失有效性的情况下减少训练集，并提高效率。然而，我们的实验表明，没有一种 IS 方法可以在所有情况下都满足所有限制。同时，在某些情况下，应用 IS 方法可能会在构建模型的时间上产生额外的开销。我们的结果对关于进行微调需要大规模训练集的提出问题给出了部分回答。在某些情况下，可以使用 IS 方法在不失去有效性和提高效率的情况下减少训练集——这意味着我们并不总是需要大量数据。因此，这些发现既不完全支持也不完全否定我们提出的传统 IS 方法能够同时满足所有提议限制的假设。

CNN	(P) : Best overall method, achieving the best trade-off on the tripod Reduction-Effectiveness-Efficiency (C) : Limited effectiveness when applied to the largest datasets (R) : Medium-to-small topic-related data or sentiment analysis tasks
LSSm	(P) : Achieved the best overall Effectiveness (C) : Lowest overall Efficiency when considering the three best IS approaches (R) : General tasks that cannot deal with effectiveness losses
LSBo	(P) : Considering the three best IS methods in terms of effectiveness, achieved the best reduction rate (C) : Limited results considering effectiveness when applied to the topic-related tasks (R) : Sentiment analysis tasks that need performance and scalability
EGDIS	(P) : Considering the IS methods in this list, achieved the best reduction rate and speed up trade-off (C) : Limited results considering effectiveness when applied to the medium-to-large topic-related task (R) : Tasks that need performance and scalability and may afford some (up to 5 %) effectiveness losses
CIS	(P) : Good effectiveness results in sentiment analysis tasks (C) : Lowest overall Efficiency (R) : Not recommended for large NLP tasks due to high associated computational cost
IB3	(P) : 4^{th} best IS method considering effectiveness (C) : Limited results considering effectiveness when applied to large topic-related tasks (R) : Medium-to-small topic-related data or sentiment tasks that can deal with a limited effectiveness loss

Tabela 2. **实例选择的优点** (P)、**缺点** (C) **和建议** (R)。

在表 2 中，我们提供了六种最佳 IS 方法的优点（P）、缺点（C）和建议（R）。其余七种方法在 ATC 环境中效果不佳，在 133 个（7 种方法 x 19 个数据集）中，有 116 个结果在统计上更差。我们的关注点尤其在于自然语言处理（NLP）任务，特别是 ATC 任务。我们的结果激励了在 ATC 环境中进一步探索 IS 方法，特别是与新型 transformer 有关的方法。我们关于 RQ1 的研究也为设计新的、更有效、有效和可扩展的 IS 方法在当前 ATC 及一般大数据场景中打开了空间。为帮助弥合这一差距，在下一章中我们引入了 E2SC 框架，这是一个新的两步框架，可以满足三脚架的所有约束，并且可以在实际情况下使用，即使数据集中包含数千个实例，尤其关注基于 transformer 的体系结构。

最终备注：第 4 和第 5 部分在发表一篇有影响力的论文于 ACM Computing Surveys (CSUR) [Cunha et al. 2023b] 中起到了至关重要的作用。

4. 一种有效、高效且可扩展的基于置信度的实例选择框架，用于基于 Transformer 的文本分类

在上一节中，我们发现没有一种方法能够在所有情况下都满足所有限制。此外，在某些情况下，使用这些方法增加了构建模型所需的时间，这对应于文献中关于能够同时遵守所提出限制的方法的空白。

为了帮助缩小这一差距，本节的主要贡献是提出了 E2SC ——有效、高效和可扩展的基于置信度的样本选择——一个新颖的两步框架，特别针对基于变换器的架构的大型数据集的首个冗余导向的实例选择解决方案。E2SC 是一种满足支架约束并适用于包括成千上万实例的数据集等现实场景的技术。E2SC 的总体结构见图 ?? 。E2SC 的第一步——图 ?? (a)——旨在为每个实例从训练集中被移除分配一个概率（α 参数）。我们采用一个精确的 KNN 模型解决方案 [1] 来估计实例被移除的概率，因为它被认为是一个经过校准的 [2] 模型，并且计算成本低（速度快）[Cardoso et al. 2017] 。我们的第一个假设（H1）是高置信度（如果模型在训练中已校准为正确的类别）与冗余在构建分类模型方面有正相关。因此，我们将难以分类的实例（可能位于决策边界区域），根据置信度加权，保留用于下一步，其中我们仅部分去除简单的实例。作为我们方法的第二步——见图 ?? (b)——我们建议通过使用验证集和一个弱但快速的分类器来估计一个近似最优的减少率（β 参数），以避免降低深度模型的效果。我们的第二个假设（H2）是，我们可以通过在一个较弱的模型中分析和变化选择率来估计一个稳健模型（深度学习）的效果行为。同样，我们在这方面探索了 KNN。更具体地，我们引入了一种迭代方法，利用验证集，统计比较 KNN 模型在没有数据减少的情况下的效果，与具有迭代数据减少率的模型的效果。这样，我们可以估计一个不会影响 KNN 模型效果的减少率。最后，将这两个步骤的结果结合起来考虑（见图 ?? (c)），随机采样 $\beta\%$ 实例，按 α 分布加权，从训练集中去除。

实验结果：我们将 E2SC 方案与六个先进的稳健实例选择基线方法进行比较，这些方法在一个包含 19 个数据集的大型基准中被认为是七个最佳深度学习文本分类方法的输入。我们的实验评估表明，E2SC 成功地显著减少了训练集（平均减少 27 %；减少幅度在 10 % 到 60 % 之间），同时在 19 个数据集中 18 个数据集维持了相同的效果水平。此外，我们发现 E2SC 能够减少所有（19 个）考虑数据集的总 ATC 模型构建时间，同时保持有效性，平均加速 1.25 倍，速度从 1.02 倍（Books）到 2.04 倍（yelp_reviews）不等。

总的来说，综合考虑三个三脚架约束和所有数据集，目前最好的 IS 方法是我们首次提出的框架。最后，为了展示我们框架应对大型数据集的灵活性，我们提出了两个修改。我们增强的解决方案在保持所有数据集中相同的效果水平的同时，成功地将训练集的减少率提高到平均 29 %，平均加速 1.37 倍。该框架扩展到大型数据集，将其减少到最多 40 %，同时在统计上保持相同的效率，平均加速 1.70 倍。尽管在有效性、效率和减少方面取得了显著成果并体现了创新，E2SC 框架仅仅关注冗余，未触及一些可能进一步减少训练的其他方面。其中一个方面是噪声，这里定义为在数据集 [Martins et al. 2021] 中由人类错误标记的实例，以及不利于（甚至阻碍）模型学习（调整）的（可能的）离群值。事实上，根据 [Martins et al. 2021] ，无论是普通个人还是专家，在标记复杂实例时都会有一个合理数量的错误——比

[1] 我们从一个前提出发，即精确的 KNN 模型解决方案是冗余性的合理代理。在我们的论文中，我们对这一前提进行了测试和验证。

[2] 一个经过校准的分类器是其概率类别预测与其准确性高度相关的分类器，例如，对于那些以 80 % 的置信度预测的实例，该分类器在大约 80 % 的情况下是正确的。

例在 56 % 至 64 % 之间。其他少数噪声实例（即，离群值）可能被正确标记，但它们与同一类别的其他实例存在显著差异，以至于在学习（调整）方面无用，甚至可能对过程有害。在这些环境中，噪声实例可能占据可用数据的显著比例。噪声（训练）实例不仅会通过在模型中引入误导性模式来降低模型的有效性，还可能因需要处理这些模式以将其融入模型而对性能造成损害。如果噪声量很大，肯定会对有效性和效率产生负面影响。然而，在设计用于评估 IS 基本方法及我们之前解决方案是否能够去除噪声的模拟场景中，没有任何 IS 解决方案令人满意地完成了任务。

最后说明：本节在 ACM SIGIR 信息检索研究与发展会议 [Cunha et al. 2023a] 中的会议论文中进行了总结。

5. 一种面向噪声且具备冗余感知的变压器自动文本分类实例选择扩展框架

本节的主要贡献是提出了一个扩展的双目标实例选择（biO-IS）框架，建立在我们第一个框架的基础上，旨在同时去除冗余和噪声实例。

本节的主要贡献是提出了一个扩展的双目标实例选择（biO-IS）框架，旨在同时去除冗余和噪声实例。正如图 ?? 所示，我们的扩展框架包含三个主要组件：一个弱分类器、一个基于冗余的方法和一个基于熵的方法。在图 ?? 中（蓝色），我们从上一节提出的原始解决方案出发，考虑使用逻辑回归作为校准的弱分类器，而不是原始工作中的 KNN。包括决策树（DT）、逻辑回归（LR）、XGBoost、LightGBM（LGBM）和线性 SVM 在内的几种弱分类器可能性的深入比较分析（详见论文），表明在效率-校准-成本权衡方面 LR 是最佳选择。为了解决去除噪声的第二个目标（图 ?? 的下半部分 - 紫色），我们提出了一个基于熵的步骤和一种新颖的迭代过程来估计近似最优的减少率。考虑弱分类器错误预测的实例，主要目标是为每个实例分配一个概率，以基于实例是噪声的概率将其从训练集中移除。为此，我们建议使用熵函数作为代理来确定这些实例在训练 ATC 模型时引起的减少行为。这一步背后的直觉是，当经过校准的弱分类器提供的预测不正确时，后验概率的熵与分类器的置信度呈负相关。这意味着，当分类器以绝对确定性将一个实例分配给错误的类别时，会出现低熵；而当分类器在多个类别之间不确定时，会出现高熵。因此，我们考虑噪声实例被移除的机会与预测的熵的倒数成正比。因此，提出的 biO-IS 框架提供了一种综合解决方案，可以同时解决冗余和噪声移除问题。实验结果：

我们将 biO-IS 与七种稳健的最先进的 IS 基线方法进行比较，包括我们在文本分类领域的首次提议 E2SC，考虑相同的 22 个数据集的基准。我们的实验评估表明，在一个设计用于评估 IS 基线方法及我们先前解决方案去除噪声能力的模拟场景中，没有一个 IS 解决方案能令人满意地完成任务。另一方面，biO-IS 成功去除了高达 66.6 个% 的人为加入的噪声。此外，biO-IS 在所有考虑的数据集中保持相同的有效性水平的同时，成功显著减少了训练集（平均减少 40.1 个%；减少范围在 29 % 到 60 % 之间）。同时，biO-IS 在平均情况下提供了 1.67 倍的加速（一共最多加速 2.46 倍）。考虑所有三个标准，甚至我们先前的最先进解决方案也没有能达到这种质量水平的基线。实际上，唯一能在所有数据集上保持有效性的其他方法是 E2SC；biO-IS 在减少率方面比 E2SC 提高了 41 个%，在加速方面从平均 1.42 倍提高到 1.67 倍，在实例选择领域达到了最先进的水平。

最后备注：本节已总结在信息系统领域顶级期刊《ACM 信息系统事务》[Cunha et al. 2024] 的一篇期刊论文中。

6. 结论与未来工作

本论文调查了经典和最新的 IS 方法，揭示了显著的进展但应用范围有限。大多数传统方法针对小型表格数据集，而在自然语言处理中的应用则很少，尽管这具有潜在的好处。

为了填补这一空白，我们对经典和 SOTA 的 IS 方法在 ATC 中的应用进行了严格的比较研究。该研究评估了缩减、效果和成本之间的权衡，这主要是由于深度学习嵌入和数据量增加导致新 ATC 解决方案的成本上升。从超过 5000 次实验中得到的主要发现包括：(i) IS 方法很少能提高 ATC 模型性能，少数例外如 XLNet。(ii) 在 13 种 IS 方法中，LSSm，CNN 和 LSBo 表现最好，在 12 个数据集上实现了显著缩减（平均 46.6 %），同时保持了有效性。(iii) 微调 transformers 对于效果至关重要，在很多情况下 IS 方法有助于减少训练成本（190 个中的 161 个）。与普遍看法相反，神经网络往往需要具有代表性而非庞大的数据才能在 ATC 中表现良好。总体而言，IS 技术能够有效减少训练集的大小而不影响效果，特别是在微调 transformer 模型时。然而，传统方法往往无法满足所有权衡，凸显了在大数据场景下更高效、可扩展的 IS 技术的必要性。

为了解决这些挑战，我们提出了 E2SC，这是一种面向冗余的两步 IS 框架，专为大型数据集和基于 transformer 的架构设计，主要引入：(i) 校准的弱分类器，用于在 transformer 训练期间估计数据的有效性。(ii) 确定最佳缩减率的迭代过程和启发式方法。E2SC 在 22 个数据集上保持有效性的同时，将训练集平均减少了 30 %，达到了最高 70 % 的加速效果。

为了克服 E2SC 在有效处理噪声方面的局限性，我们开发了 biO-IS，这是一种扩展框架，可以同时去除冗余和噪声实例。实验结果表明，biO-IS 平均减少了 41 % 的训练集（最多达到 60 %），去除了 66.6 % 的人为插入噪声，并在不同数据集上保持了有效性。它提供了一致的加速效果（平均 1.67 倍，最高达 2.46 倍），在缩减和效率方面优于 E2SC。

这些发现证实了论文的假设：小型和大型语言模型可以在数据较少的情况下进行训练，而不会牺牲效果。这不仅能够节省大量成本和能源，还有助于减少碳排放。如此有希望的结果为自然语言处理领域的未来带来了希望，在该领域中，信息系统技术的进步可以带来巨大的环境和经济利益。

通过提出和评估导致提出的 biO-IS 框架的增强（修改和扩展），我们的工作开启了几个新的机会。在未来的工作中，我们打算在其他模型上评估我们提出的框架，例如 T5 和 DeBERTa，以评估其泛化能力。我们还计划将 IS 概念应用于 ATC 之外的其他领域，比如搜索、排序、推荐，以及其他 NLP 任务，如问答和监督主题建模。到目前为止，我们已经测量了整个文档在训练（调优）过程中的重要性，但在未来，我们可以在更细粒度的层次上进行，例如针对评论的方面 [Luiz et al. 2018] 和文档的段落。我们还计划在未标记的深度学习预训练阶段使用我们的框架，例如，更有效地从头开始构建大型语言模型。

我们还希望在 AutoML 解决方案的背景下研究引入实例选择（IS）技术，作为 [Cunha et al. 2020] 中那样的一个转换流程中的步骤。此外，考虑到文本数据集的维度庞大，评估特征选择方法如何与实例选择相互作用可能是非常有趣的。这主要涉及与表示文本数据的文档-特征矩阵的行（文档）和列（特征）进行协作。最后但同样重要的是，我们打算对大语言模型（LLM）的成本效益进行分析（效率对比成本），以便可能通过实例选择进行 LLM 的微调，同时利用新的可扩展计算范式，如量子计算 [Ferrari Dacrema et al. 2024] 。

这项工作得到了 CNPq、CAPES、INCT-TILD-IAR、FAPEMIG、AWS、Google、NVIDIA、CIIA-Saúde 和 FAPESP 的部分资助。

Referências

Andrade, C., Cunha, W., Fonseca, G., Pagano, A., Santos, L., Pagano, A., Rocha, L., and Gonçalves, M. (2024). Explaining the hardest errors of contextual embedding based classifiers. In Proceedings of the 28th Conference on Computational Natural Language Learning.

Barigou, F. (2018). Impact of instance selection on knn-based text categorization. Journal of Information Processing Systems, 14(2).

Carbonera, J. L. and Abel, M. (2018). Efficient instance selection based on spatial abstraction. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI).

Cardoso, T. N., Silva, R. M., Canuto, S., Moro, M. M., and Gonçalves, M. A. (2017). Ranked batch-mode active learning. Information Sciences, 379:313–337.

Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M. A., and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. IP&M, 57(4):102263.

Cunha, W., França, C., Fonseca, G., Rocha, L., and Gonçalves, M. A. (2023a). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., and Gonçalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. Information Processing & Management.

Cunha, W., Moreo, A., Esuli, A., Sebastiani, F., Rocha, L., and Gonçalves, M. (2024). A noise-oriented and redundancy-aware instance selection framework. ACM Transactions on Information Systems.

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., and Gonçalves, M. A. (2023b). A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. ACM Comput. Surv.

de Andrade, C. M., Belém, F. M., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations –or "the classifier does not matter when the (text) representation is so good!" . IP&M, 60(4):103336.

Ferrari Dacrema, M., Pasin, A., Cremonesi, P., and Ferro, N. (2024). Quantum computing for information retrieval and recommender systems. In European Conference on Information Retrieval, pages 358–362.

Garcia, S., Derrac, J., Cano, J., and Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hart, P. (1968). The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3):515–516.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. Biometrika, 75(4).

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In Proceedings of the Conference European Chapter Association Computational Linguistics (EACL).

Leyva, E., González, A., and Pérez, R. (2015). Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective. Pattern Recognition.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST), 13(2):1–41.

Lu, X., Duan, X., Mao, X., Li, Y., and Zhang, X. (2017). Feature extraction and fusion using deep convolutional neural networks for face detection. Mathematical Problems in Engineering, 2017.

Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. (2018). A feature-oriented sentiment rating for mobile app reviews. In Proceedings of theWebConf'18.

Martins, K., Vaz de Melo, P., and Santos, R. (2021). Why do document-level polarity classifiers fail? In Proceedings of the 2021 Conference of the NAACL: Human Language Technologies.

Tsai, C.-F., Chen, Z.-Y., and Ke, S.-W. (2014). Evolutionary instance selection for text classification. J. Syst. Softw., 90(C):104–113.

Uppaal, R., Hu, J., and Li, Y. (2023). Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. arXiv preprint arXiv:2305.13282.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, pages 408–421.

Wilson, D. R. and Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. Machine learning, 38(3):257–286.