PedagoSense：一种以教育学为基础、用于学习对话中教学策略检测与情境化响应生成的大语言模型系统

Shahem Sultan[1] Shahem Fadi[2] Yousef Melhim[2] Ibrahim Alsarraj[3] Besher Hassan[3]

[1] 安达卢斯大学[2] 阿治曼大学[3] 穆罕默德·本·扎耶德人工智能大学

arXiv:2602.01169v1 [cs.CL] 1 Feb 2026

摘要

本文针对如何通过检测并推荐辅导教师采用的有效教学策略，提升基于对话的学习环境中的交互质量这一关键问题展开研究。我们提出了一种新颖方法，将自然语言处理技术应用于师生对话中教学策略分类的建模。该方法采用两层分类器：首先对教学策略进行二元分类，随后进行细粒度分类，以确定所使用的具体策略。我们的实验使用了经过全面标注教学策略标签的对话数据集。初步结果表明，系统在两类分类任务中均表现出良好的准确率。此外，我们还利用了一些最新模型，根据对话情境推荐最合适的教学策略。本研究旨在弥合教学理论与实践应用之间的差距，推动开发更有效、个性化的教育技术。

GitHub代码库

1 引言

1.1 概述

本项目旨在创建 Tutor Copilot 系统，使其能够检测辅导教师在基于对话的学习情境中所使用的教学策略，并提出策略建议。随着智能辅导系统在教育领域日益普及，有必要模拟辅导教师与学生之间的互动，以便为学生提供更好的个性化教学指导。发现并提出这些策略是本项目的共同工作，将有助于开发面向学生的温和而智能的辅导系统，从而提升其体验。这些洞见将极大地促进更先进、更有效的教育技术在自然语言处理领域的发展。

1.2 相关工作

Wang等人（Wang和Liu，2023）探讨了如何借助大语言模型，在教育对话中利用专家决策提升新手辅导教师的表现。他们展示了自然语言处理系统通过认知任务分析缩小专家与新手辅导教师之间知识差距的可行性。

Demszky等人（Demszky和Wang，2022）介绍了量身定制的教育技术相关工作。他们探讨了如何向教师及其他使用自然语言处理系统评估课堂互动的人员提供反馈，并得出结论：采用适当的教学方法能够提升学习者对课程的关注度。

Li 等人（Li和Liu，2020）考虑开发一种对话式人工智能，用于评估并分类教育环境中的导师与学生互动。学术讨论中一直有人呼吁对导师对话进行更细致的编码，以便更好地理解互动。

Hosseini 等人（Hosseini 和 Zhang，2021）提出了一种采用深度学习模型来揭示电子学习环境中所用教学策略的技术。准确率更侧重于特定策略的固有特征，并为利用教育自然语言处理策略开展策略识别的进一步方向奠定了基础。

在借鉴这些主题和先前研究的基础上，本项目将致力于在统一的自然语言处理模型中纳入教学策略检测。具体而言，我们希望构建一个整合基础模型与基于深度学习的语言模型的系统，以便在双向交流过程中对教学策略进行分类，

并实现实时分类。在此情况下，我们致力于增强智能辅导系统，使其能够实时提供更易理解、更个性化且更能结合语境的教学策略建议。

2 项目构想

2.1 拟议方法

我们开发的系统利用学生与导师对话中的现有数据来识别并运用有效的教学策略，从而实现自适应辅导。教学策略可以包括：探询、提供提示、提出类似问题或解释概念等。例如，当学生与导师交流时，表示自己在某个问题上遇到困难或感到困惑，经过微调的 BERT 生成模型会查看对话内容，并根据具体情境推荐一种似乎适合该场景的教学策略。随后，这一推荐策略会传递给一个 LLM（通过 API 调用 GPT-4o），由其生成合适的回复，供导师向学生解释。LLM 生成回复后，会由另一个基于 BERT Base 的二元分类模型进行审核，以检查回复中是否包含教学策略。在此情境下，“0”表示未使用任何策略，“1”表示已采用教学策略。如果回复中包含策略，则会将其交给经过微调的 BERT 分类器，以识别回复中所采用的教学策略类型。我们会将分类结果与生成模型提出的策略进行比较；如果两者一致，就能确认所提出的教学策略，并验证我们的系统是恰当且有用的。您可以通过以下链接观看演示我们系统 Tutor Copilot 的视频：视频

3 数据集

Dr. Kaushal Kumar Maurya 提供了两个数据集，其中包含导师与学生之间的对话样本。此外，还使用了来自 Hugging Face DailyDialog 数据集的 697 个样本来扩充二元分类中的类别 0，该数据集包含一般性对话。

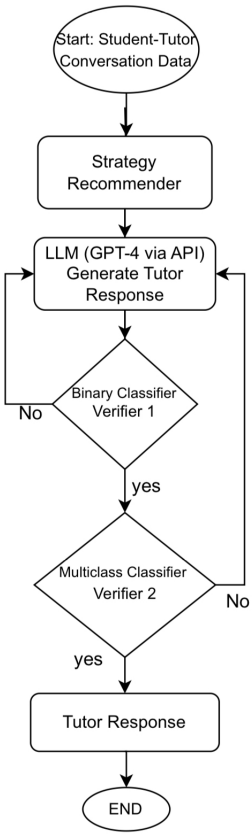

图 1：系统流程图

3.1 二元分类特征

教学策略标签：在二元分类任务中，标签“0”表示未采用任何策略，而“1”表示采用了某种策略。

规模：1.4K 对

3.2 多类别分类与策略推荐特征

对话历史：该数据列中的每一行都包含一段语义相对完整的导师与学生之间的对话，这部分内容用作策略预测模型的输入。

专家人工导师回复：人类导师在相应对话历史之后给出的回复，也是多分类模型的输入。

教学策略标签：导师所采用的具体策略详情，例如探询、解释概念、提供提示或无。预处理后，数据集中包含 8 种不同的教学策略。

4 实验与评估

4.1 二元教学策略检测

4.1.1 数据准备与预处理

在这一阶段，我使用了专家_人工_导师特征，舍弃无关特征并移除重复项，以避免数据污染或偏差。预处理包括转换为小写、移除标点符号、移除停用词以及词形还原，从而规范化数据以提升泛化能力。去重后，数据集出现类别不平衡：类别 0 有 303 个样本，类别 1 有 279 个样本。因此，为了平衡数据，首先采用了 SMOTE。该方法通过识别特征空间中每个少数类样本的最近邻，然后在原始样本与其邻居之间进行插值，生成新样本。这一过程能够生成真实且多样的样本，在实现平衡的同时避免简单地重复数据。因此，如图2所示，每个类别的样本数都增加到了 303 个。虽然 SMOTE 是一种简单快速的数据平衡方案，但数据增强等技术能够生成质量更高的数据，从而提升模型性能。因此，第二种方案采用 GPT-4o API 进行数据增强。对于类别 1，我使用 GPT-4o API 生成了 721 个额外样本，使样本总数增至 1,000 个。这些样本主要包含教学策略相关示例。对于类别 0，我从 Hugging Face 的 DailyDialog（Li 等人，2017）数据集中纳入了 697 个样本，该数据集包含不涉及教学策略的通用对话，使样本总数增至 1,000 个。最终的训练数据集实现了平衡，共有 2,000 个样本，两个类别各占一半。测试集和验证数据集则保持不变，分别包含 223 个测试集样本和 73 个验证集样本。

虽然 SMOTE 有效地平衡了数据集，但它在高维数据中引入了噪声，可能导致过拟合。这凸显了对更稳健方法（如数据增强）的需求，以增强泛化能力。

4.1.2 模型选择与设置

我选择了 BERT Base，用于二元分类任务。我使用 Hugging Face 库，基于分词后的文本数据对模型进行微调，并将最大 token 长度设为 128，以确保数据集中的一致性。关键训练参数包括学习率 5×10^{-5}、批量大小 16，以及

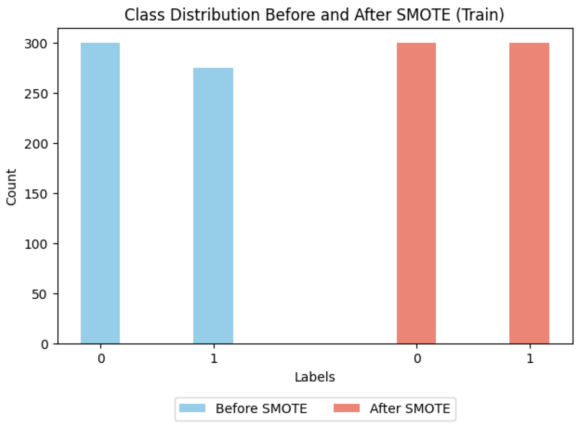

图 2：SMOTE 前后的类别分布

一种用于缓解过拟合的早停机制。早停通过在验证损失不再改善时停止训练，优化了训练性能和泛化能力。

4.1.3 基线模型与评估指标

基线模型采用逻辑回归，并使用 TF-IDF 特征，在验证集和测试集上的 F1 分数分别达到 95.83% 和 94.90%。基线模型和 BERT 模型均采用准确率、精确率、召回率、F1 分数和损失作为评估指标；BERT 的具体性能详情见“实验结果”部分。

4.1.4 实验结果

采用 GPT-4o 进行数据增强的 BERT 模型在第 6 个训练轮次取得了 98.85% 的最佳验证集 F1 分数，在 10 个训练轮次内的验证损失为 0.0636。在测试集上，该模型的 F1 分数达到 98.5%，显著优于基线逻辑回归模型。表 1 比较了各模型在二元教学策略检测任务中的 F1 分数，并纳入了 SMOTE 和数据增强方法。SMOTE 提升了朴素贝叶斯和支持向量机模型的性能。数据增强则提高了大多数模型的 F1 分数，尤其是微调后的 BERT、多层感知机和 LightGBM。随机森林和决策树在数据增强后的性能有所下降，这可能是由于它们对噪声或过拟合较为敏感。BERT Base 始终取得最高分，证实了其在该用例的自然语言处理任务中的有效性。

4.1.5 大语言模型评估

图 3展示了四种大型语言模型（大语言模型）在 50 个样本上的 F1 分数，其中 25

模型	F1 分数 (SMOTE) F1 分数 (数据增强)	
朴素贝叶斯	97.73 98.0	
BERT Base 微调后	97.34	**98.5**
SVM	97.34	96.0
MLP	96.55	98.0
逻辑回归	95.31	95.0
随机森林	93.02	87.0
决策树	88.16	80.0
XGBoost	86.61	92.0
AdaBoost	86.07	85.0
LightGBM	84.13	90.0
LDA	79.38	95.0
微调前的 BERT	77.0	77.0
KNN	22.82	52.0

表 1：F1 分数对比：SMOTE 与数据增强在测试集上的表现

个样本来自类别 0，另外 25 个来自类别 1，用于检验它们识别教学策略的能力。LLaMA-2 得分最高，达到 94.12%，在分析和识别教学策略方面表现出色。GPT-4o 达到 89.80%，表现良好，但略逊于 LLaMA-2。Gemini 和 Claude 的得分则明显较低，显示出识别教学策略方面的挑战。

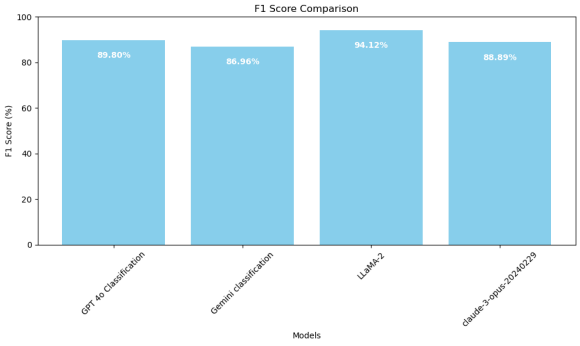

图 3：词语对分类的贡献

4.1.6 分析

为解释 BERT 模型的预测结果，我对五个分类错误的样本应用了局部可解释模型无关解释（LIME），突出显示了对每项预测至关重要的词语。如图 4所示，LIME 使我们能够观察某些词语如何对模型决策产生正面或负面影响。例如，“great”等词语具有较强的正面影响，从而导致分类错误并增加错误类别预测的可能性；而“say”和“got”等词语则产生负面影响，使预测偏离实际目标类别。这些观察表明，模型可能过度依赖少数显著的关键词，而没有考虑更广泛的语境倾向，因此不利于进行细致入微的预测。此外，SMOTE 和数据增强对模型性能的影响呈现出不同的模式。例如，数据增强通过引入多样化的训练样本，提高了 BERT 和多层感知机等模型的性能，使这些模型能够更好地泛化。然而，它也为随机森林等基于树的模型引入了噪声，可能削弱了这些模型区分有意义模式的能力，这凸显了根据不同模型类型的特点量身定制预处理技术的重要性。

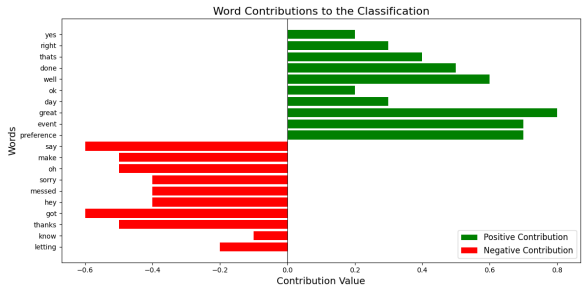

图 4：词语对分类的贡献

4.2 细粒度分类

4.2.1 数据准备与预处理

由于数据集规模较小，数据质量至关重要。我移除了出现次数少于 20 次的所有标签，因为模型无法从中有效学习，而且这些标签会成为噪声，影响模型性能。我使用的两个主要列是 (Ex-pert_Human_Tutor)，即导师在对话结束后对学生的回复；以及 (Peda_Strategies)，即目标教学策略。输入文本使用 BERT分词器进行分词，并转换为 ID，序列最大长度为 128 个词元。如果输入超过词元数限制，就会被截断。另一方面，较短的序列会补入填充词元。此外，还使用注意力掩码来区分实际词元和填充词元。注意力掩码和填充后的输入序列被转换为张量，作为模型的输入。目标教学策略使用标签编码器进行编码，将字符串标签转换为适合训练的数值。

4.2.2 数据增强

鉴于可用数据集规模较小，且教学类别分布不平衡。

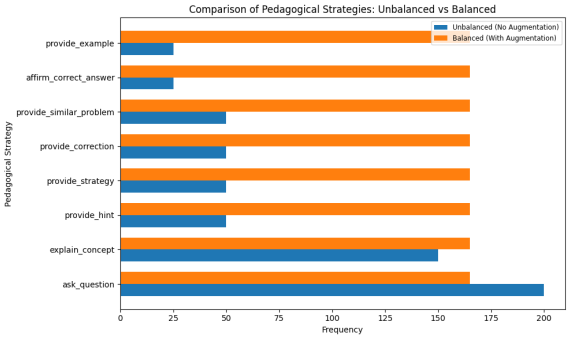

图 5：教学策略对比：不平衡与平衡。

GPT-4o被用于生成并增强专家_人类_导师数据，并采用目标教学策略，以获得更具泛化能力的模型。图5 展示了数据增强前后每种教学策略的频率。

4.2.3 模型选择与设置

我选择了预训练的 BERT Large Uncased 语言模型进行微调。使用不区分大小写的版本，是因为它不会区分大写和小写字母，因此具有较强的鲁棒性。训练期间使用带权重衰减的 AdamW 优化器更新模型权重。我们采用了 Hugging Face Transformers 库中专为分类任务设计的变体 BertForSequenceClassification，其输出经过 SoftMax 层后，将得到涵盖八种不同教学策略的概率分布。表 2 展示了测试后选出的最佳超参数。

超参数	值
学习率	2×10^{-5}
批量大小	16
训练轮数	20（启用早停）
优化器	AdamW（启用权重衰减）
最大序列长度	128 个词元

表2：细粒度分类模型超参数。

4.2.4 基线与评估指标

我的基线模型将采用较小版本的 BERT Uncased，即 BERT Base Uncased 模型。

4.2.5 实验结果与分析

表 3 展示了增强平衡数据集上两个不同 BERT 变体之间的验证准确率、测试宏平均 F1 分数和测试准确率。

指标	BERTLarge Uncased (%)	BERTBase Uncased(%)
验证准确率	60.27	57.53
测试准确率	56.65	53.76
宏平均F1分数	45.95	44.87

表3：模型性能指标（增强的平衡数据集

指标	BERTLarge Uncased（%）	BERTBase Uncased（%）
验证准确率	55.15	57.53
测试准确率	49.13	49.13
宏平均 F1 分数	42.17	32.67

表 4：模型性能指标（不平衡数据集 | 未进行数据增强

另一方面，表4 展示了在通过数据增强处理的不平衡数据集上，两个不同 BERT 变体之间的验证准确率、测试宏平均 F1 分数和测试准确率。

表 5 展示了 BERT Large Uncased 在增强平衡数据集上各教学策略类别的平均类别准确率。不同模型的更多实验结果请参见附录 A.1.1

类别	平均准确率（%）
肯定正确答案_ _	66.67
提问_问题	88.17
解释_概念	67.86
提供_纠正	28.94
提供_示例	25.25
提供_提示	26.67
提供_相似_问题	30.00
提供_策略	26.60

表 5：教学类别平均准确率。

根据实验结果，可以得出 4 点观察结果：

1）两种预训练的 BERT 变体在验证准确率、测试准确率和宏平均F1分数方面均表现出性能提升。例如，BERT Large Uncased 的宏平均F1分数从 42.17% 提高到 45.95%，证明了平衡不均衡教学策略的有效性。

2）BERT Large Uncased 的表现始终高于 BERT Base Uncased。例如，BERT Large 的宏平均F1分数达到 45.95%，而 BERT Base 为 44.87%，展示了更大模型检测并捕捉细微模式的能力，尽管差距并不显著。

3）提供_示例和提供_提示等教学策略是模型面临的主要问题，使其难以加以区分。

与询问_问题和解释_概念等高平均准确率类别相比，这表明它们的区别更为细微或模糊，因而需要更多数据增强和稳健的表征。

4）高准确率的主导教学类别对整体宏平均F1分数贡献很大。

4.3 教学策略推荐

4.3.1 数据集清理

与第 4.2.1 节相同，仅保留初始数据集中出现频率大于 20 的标签。在此任务中，唯一使用的特征数据列为对话历史。

4.3.2 基线模型

混合传统投票。 该方法利用三种传统机器学习模型——支持向量机（SVM）、朴素贝叶斯和提升法——的互补优势，通过简单的多数投票机制预测教学策略。每个模型均独立地在同一数据集上训练，该数据集经过停用词去除、词干提取、标签编码，以及将文本特征转换为 TF-IDF 向量的预处理，以确保特征表示一致。支持向量机擅长利用分离超平面对高维特征空间建模，在精确率方面表现出色。朴素贝叶斯是一种假设特征独立性的概率分类器，召回率表现出色；而提升法通过重点处理较难分类的样本，逐步改进分类效果，整体表现均衡。混合传统投票方法通过多数投票汇总各个模型的预测结果，有效结合了它们的优势，从而提高了教学策略的推荐准确率。该基线模型的指标见表6。

基于标签概率分布的推荐（LPD）。假设当前数据集反映了教学策略在现实世界中的最佳应用情况，那么数据集中不同策略标签的出现频率，必然蕴含着人类教师使用这些策略的相关信息。例如，如果“提问”策略的出现频率是其他策略的三倍，那么在大多数情况下推荐该策略是合理的。因此，在基线推荐模型中，我将标签频率分布用作概率来推荐教学策略。该基线模型的指标见表6。

结合 BM25 和嵌入相似度（BES）的推荐方法。我根据当前对话历史的 BM25 得分，选取了排名靠前的k（$k = 5$）条最相似的对话历史。随后，我计算了当前对话历史与这些 k 条相似对话之间的嵌入相似度。我引入了一个超参数 α 来平衡这两种方法，最终得分按以下公式计算：（$\alpha \times$bm25_得分$+$（$1-\alpha$）$\times$嵌入_得分）$\times$策略_概率，其中策略_概率表示标签的先验概率。当 $\alpha = 0.2$时，该模型取得了最佳结果。此方法会推荐综合得分最高的教学策略。该基线模型的指标见表6。

BERT-Base。此方法使用预训练的 BERT 模型，具体而言是`bert-base-uncased`，不进行任何数据增强，用于预测教学策略。模型使用批量大小为 8、学习率为 2×10^{-5}的 AdamW 优化器，在清理后的数据集上进行微调，训练 20 个训练轮次。BERT-Base 利用预训练的语言表示，根据输入的对话历史直接对策略进行分类。然而，该方法表现不佳，甚至略逊于混合传统投票方法，这凸显了数据集规模有限所带来的挑战。该基线模型的指标汇总于表6。

4.3.3 数据增强

我审阅了数据集中的全部677条记录，并发现了两个关键结论。1）数据集规模明显过小，不足以让模型学习到足够的知识以进行有效的策略推荐。2）不同标签之间存在显著的重叠与混淆。例如，被标注为需要“提供策略”方法的对话历史，与需要“提供提示”方法的对话历史并无明显区别；本质上，这两种标签都可以合理地应用于这些对话历史。

为丰富数据集的信息，我使用 GPT-4o-2024-08-06 模型进行数据增强。具体而言，我生成了：1）教学策略的解释，说明当前对话历史为何应采用相应的教学策略；2）具体关键词或短语列表，作为帮助识别类似情境的“线索”，以判断当前策略是否有效。

这些内容作为两个新列添加到了数据集中。

值得注意的是，为了确保 API 返回包含格式化字符串和列表的结构化输出，我采用了 OpenAI 最新的“结构化输出”技术，这需要在代码中预先定义返回类。

4.3.4 拟议模型

增强型 BERT。 我使用标签编码器 将策略标签从字符串转换为数值，用于训练。为进行数据增强，我使用OpenAI GPT API 为每个数据实例生成策略解释和线索列表。随后，将这些增强文本与对话历史拼接，形成丰富的输入数据。我从transformers库中加载bert-base-uncased模型及其对应的分词器，并配置用于分类的输出层，使其与数据集中唯一策略标签的数量一致。TeachingStrategiesDataset继承自torch.utils.data.Dataset，对增强后的输入文本和标签进行分词与编码，并返回用于输入_ID、注意力_掩码和标签。我使用批量大小为 8 的设置训练模型 10 个训练轮次，并使用 AdamW 优化器更新模型参数（学习率=2e-5）。我使用`torch.argmax`从逻辑值中选取预测标签。该增强模型的指标见表6。

混合 BERT 投票方法。 混合 BERT 投票方法整合了四种来源的预测：BM25、嵌入相似度、标签概率和数据增强 BERT 模型。每种来源都会针对给定输入独立预测教学策略。通过实验观察，我发现这些方法往往能在不同样本上做出正确预测。受此启发，我尝试结合这些方法，以充分发挥数据集的潜力。最终推荐结果通过简单的多数投票机制确定，汇总了这些不同方法的预测。通过结合传统信息检索技术和数据增强 BERT 模型的优势，该方法旨在提高教学策略推荐的稳健性和准确率。该方法的指标见表 6。

混合 BERT 概率投票方法。与混合BERT投票方法类似，混合 BERT 概率投票方法整合了来自四个来源的预测。不过，该方法并未依赖简单的多数投票机制来选出单一策略，而是采用了概率方法。我通过加权和结合所有来源的概率，计算出每种教学策略的可能性。组合权重如下：BERT (0.5), LPD (0.2)，以及 BES (0.3)。该方法根据最终计算出的概率生成推荐排序列表，从而细致呈现模型对各项策略的置信度。与混合BERT投票方法不同，这种概率方法为需要多项建议或策略优先级排序的应用提供了更灵活的框架。该方法的指标见表6。

指标	精确率	召回率	F1 分数
混合传统方法	0.2543	0.3212	0.2306
LPD	0.2118	0.2424	0.2247
BES	0.1562	0.3182	0.2026
BERT-Base	0.1909	0.3109	0.2000
增强型 BERT	0.3717	0.3729	0.4663
混合BERT	0.5470	0.5389	**0.4752**
混合BERT概率	0.5546	0.5337	**0.4815**

表 6：测试集上的性能指标。

根据实验结果，可以得出以下5项观察：

1) LPD 方法在 F1 分数上优于 BERT-Base 方法，证实了策略标签的概率分布本身蕴含着人类教师使用这些标签的知识，正如4.3.2中所讨论的。

2) 当 α=0 时，BES 方法取得了最佳结果，表明嵌入相似度方法发挥了主导作用，并且比 BM25 算法更有效。

3) 混合 BERT 概率投票方法在权重设置如下时取得了最佳结果：BERT（0.5）、LPD（0.2）和 BES（0.3）。这一配置表明，经过数据增强的 BERT 模型在组合预测中发挥了主导作用，能够提供源自丰富输入数据的高效概率。BES 组件的适度贡献突出了通过嵌入相似度和 BM25 分数利用上下文相关性的重要性，而 LPD 较小的权重则反映了其作为补充先验的作用，确保与标签频率分布保持一致。

4）数据增强后观察到性能显著提升，凸显了

数据量在机器学习任务中的关键作用。由于模型的知识完全来源于其学习的数据，数据可用性往往成为提升模型性能的主要瓶颈。

5）我们的观察发现，不同算法在不同样本上的正确预测表现各有优势，凸显了它们互补的长处。为充分发挥数据集的潜力，我们采用简单多数投票和概率投票策略，整合所有算法的输出。这种组合方法有效利用了各算法的优势，从而提升了推荐效果，并构建出更稳健的策略预测系统。

5 结论与未来工作

总之，我们创建了 Tutor Copilot，这是一个用于在基于对话的学习中分类和推荐教学策略的系统。它采用了一种新颖的方法，结合了两个分类层：用于检测是否存在教学策略的二元分类，以及用于判断教学策略类型的细粒度分类；此外，还包括策略推荐，即根据对话语境推荐教学策略。实验结果表明，该系统在分类和推荐的各项指标上均表现出色。未来的工作将着重采用更多样、规模更大的数据集，其中包含更多教学策略和场景，以进一步提升模型的泛化能力。此外，还将探索面部表情和手势识别等多模态数据输入，从而获得更深入的洞察，并显著提升 Tutor Copilot 模拟人类导师的能力。

A 补充实验

A.1 细粒度分类

A.1.1 补充实验

表7 比较了不同模型在增强后的平衡数据集上的表现。

1) BERT Large Uncased 在所有不同模型中取得了最高分，验证准确率为 60.27%，宏平均F1分数为 45.95%，测试准确率为 56.65%。

2) 决策树和 k近邻等传统模型表现相对较差，这反映出它们处理细微差别和文本关系的能力有限。

参考文献

Dorottya Demszky 和 Lingyi Wang。2022。利用自然语言处理为教师自动生成反馈。载于*2022*年自然语言处理实证方法会议（*EMNLP*）论文集，第 2456–2468 页。

Reza Hosseini 和 Yu Zhang。2021。利用深度学习检测在线学习中的教学策略。《教育中的人工智能》期刊，31：152–169。

Jianyu Li 和 Zhenhui Liu。2020。面向教育场景的对话式人工智能。载于*2020*年计算语言学会议（*COLING*）论文集，第 342–351 页。

Yanran Li、Hui Su、Xiaoyu Shen、Wenjie Li、Ziqiang Cao 和 Shuzi Niu。2017。DailyDialog：一个人工标注的多轮对话数据集。载于第八届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第 986–995 页。

Xi Wang 和 Ming Liu。2023。通过决策模型弥合新手与专家之间的差距：关于纠正数学错误的案例研究。《教育自然语言处理》期刊。

模型	验证准确率（%）	测试宏平均 F1 分数（%）	测试准确率（%）
BERT Large Uncased	60.27	45.95	56.65
BERT Base Uncased	57.53	44.87	53.76
多层感知机	51.22	44.70	50.41
朴素贝叶斯	52.03	40.43	46.34
支持向量机 (SVM)	56.10	41.17	45.53
随机森林	47.15	36.97	45.53
逻辑回归	56.10	40.29	44.72
RoBERTa Base	49.45	40.37	44.51
k近邻	47.15	35.50	41.46
决策树	40.65	32.99	40.65
感知机	53.66	32.94	38.21

表 7：不同模型的性能。