

---

# 用于在线课程讨论论坛管理的多维分类帖子

---

① **Antonio Leandro Martins Candido**

Federal Institute of Education, Science, and Technology of Ceará (IFCE) - Caucaia, CE, Brazil  
State University of Ceará – CCT-PPGCC-UECE – 60740-903 – Fortaleza, CE – Brazil  
leandro.candido@aluno.uece.br

② **José Everardo Bessa Maia**

State University of Ceará – CCT-PPGCC-UECE – 60740-903 – Fortaleza, CE – Brazil  
jose.maia@uece.br

## ABSTRACT

在线课程中讨论论坛的自动整理需要不断更新，这使得大型语言模型（LLMs）的频繁再培训成为一项资源密集型过程。为绕过昂贵的微调需求，本文提出并评估一种贝叶斯融合策略。该方法结合了预训练的通用 LLM 的多维分类分数与在本地数据上训练的分类器的分数。性能比较表明，所提出的融合方法在单个分类器的基础上改进了结果表现，并且与 LLM 的微调方法相比具有竞争力。

**Keywords** online course forum curation · multidimensional classification · Bayesian fusion · Large Language Model

## 1 引言

多维（或多输出）分类 [Jia and Zhang, 2024] 的特点是除了考虑一维分类中特征与类别之间的相关性之外，还考虑类别（标签）之间的相关性。为每个类别构建一个独立的分类器是一个可行的选择 [Candido and Maia, 2023]，但它没有利用类别之间的相关性，而这种相关性往往会带来重要的影响。实际上，可以考虑这些信息来改进在模糊区域的一维分类器的决策边界。这在自然语言文本分类中尤为重要，因为语言歧义反映在特征的歧义性上，并在辨别文本类别时起到核心作用。

### 1.1 背景：在线教学论坛的整理

在线教学论坛的策划指的是在虚拟学习环境中选择、组织和将内容与互动进行背景化的过程，目的是促进有意义的讨论、参与者的参与和协作的知识构建。这个过程对于将数字空间转变为有效的学习社区至关重要，尤其是在远程教育和大规模在线课程（MOOCs）的背景下。

在在线教育环境中，讨论论坛在促进学生和教师之间的互动方面起着至关重要的作用。然而，如果没有适当的管理，这些论坛可能会变得无序，并降低对促进学习的效果。有效的管理包括积极的监督、促进相关讨论以及整合相关的教育资源。例如，MUIR 框架提出在大型开放式网络课程中交叉链接资源，以丰富讨论论坛，提高参与者的学习体验 [An et al., 2020]。

此外，在线教育论坛的管理还涉及创建一个支持探究社区的环境，在这个环境中，参与者进行批判性和反思性讨论。通过在线讨论论坛支持探究社区的指南强调了促进学生批判性思维和合作的策略的重要性 [Mudau and Van den Berg, 2023]。

图 1 展示了我们提出的在线讨论论坛精选框架。一个被插入为论坛成员的精选引擎，通过读取帖子和回应论坛状态与论坛互动。引擎通过两种信息源的结合作出资源或行动推荐，甚至通过对话直接回应：其一是特定于课程及其环境的本地知识库，其二是预先训练的大型语言模型（LLM）的支持。算法 1 显示了该框架操作的宏观步骤。从这个算法可以看出，如果引擎的信心没有超出阈值  $th$ ，它会将当前的情况状态转发给人类导师进行解决。设计目标是将不足  $\%_2$  的干预情况提交给人类。

**Algorithm 1** 策展引擎宏步骤**Require:** posts, context,  $LKB_i$ .**Ensure:** response to forum OR refer to human.

- 1: Read new posts and retrieve context and current state.
- 2: Create and submit analysis prompt to LLM OR Multi-label classification of new posts.
- 3: Update state.
- 4: State update confidence  $> Th$ ? If not, refer to human.
- 5: Set post-label-status priority.
- 6: Create and submit a response generation prompt to LLM OR Retrieve a response from a local knowledge base.
- 7: Retrieve a response complement from a specific local knowledge base.
- 8: Compose an intervention response and submit it to the forum.

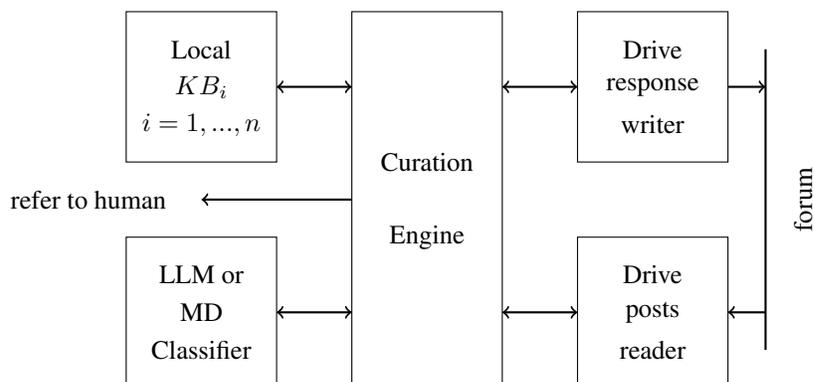


Figure 1: 在线课程论坛管理代理软件框架的功能结构。

在大规模语言数据上预训练的大型语言模型（LLMs）已知在开放领域中表现良好，但在使用专业语言的特定领域中表现较差 [Gururangan et al., 2020]。在各种领域的课程背景下，后者的情况需要对 LLM 模型进行微调以取得更好的表现。这使得大型语言模型（LLMs）的频繁再训练成为一个资源密集型过程。为了规避高昂的微调成本，本文提出并评估了贝叶斯融合的使用。该方法结合了预训练通用 LLM 的多维分类得分和在本地数据上训练的分类器的得分。性能比较表明，所提出的融合方法改善了与单个分类器相比的结果，无论分类器是仅在该领域训练的还是预训练的全局 LLM，并且在与 LLM 微调方法相比时具有竞争力。

数字整理也被认为是媒体和数字素养教育中的一项重要实践。在数字和媒体素养教育中探索整理作为核心能力，强调了教授学生整理技能的必要性，以便他们能够导航和批判性地评估在线可用的大量信息 [Mihailidis and Cohen, 2013]。

总之，在线教学论坛中的策划是一种多方面的实践，涉及积极管理内容和互动，以促进有意义的协作学习。这需要结合教育学、技术和管理策略，以创建有效且引人入胜的在线学习环境。

## 2 方法

### 2.1 多维分类

多维分类 (MDC) 是一种监督学习范式，其中每个输入实例与多个分类变量相关联，每个变量代表一个独特的语义维度。与传统分类不同的是，传统分类为每个实例关联一个单一标签，而 MDC 试图预测一个类别向量，以捕捉数据的语义复杂性。例如，当在教育论坛中对帖子进行分类时，一个帖子可能同时表达观点、包含问题并表现出紧迫性，每个都对对应一个不同的维度 [Jia and Zhang, 2024]。

MDC 和多标签分类的主要区别在于输出的结构。多标签分类为每个实例与来自单一类别空间的一组二元标签相关联，而 MDC 则涉及多个异质的类别空间，每个空间有其自己的类别。这一区别非常重要，因为在 MDC 中，各个维度在语义上是不同的且不可直接比较，需要特定的方法来捕捉它们之间的依赖关系 [Jia and Zhang, 2024, Huang et al., 2015]。

建模不同维度之间的依赖关系是多维分类中的主要挑战之一。诸如分类器链 [Jia and Zhang, 2021] 之类的算法通过显式地建模类别变量之间的依赖关系来应对这一挑战。其他方法，如使用深度学习技术 [Arya et al., 2023]，也已被探索以更有效地捕捉这些复杂关系。

MDC 在多个领域有实用的应用，例如在线论坛策展 [Zhang et al., 2016]，在这些领域，帖子需要跨多个维度进行分类以优先处理响应和干预。同时，它也适用于推荐系统 [Anagnostopoulos et al., 2020]、情感分析 [Garg, 2021] 和医学诊断 [Chen et al., 2018]，在这些领域中需要同时考虑多个特征或症状。

在评估方面，MDC 使用指标来考虑每个维度的准确性，以及评价模型在所有维度上的整体性能的汇总指标。这些包括微观和宏观准确性的测量，它们提供了模型在不同粒度级别上的性能洞察 [Grandini et al., 2020, Opitz, 2024]。

## 2.2 分类器融合

分类器融合可以根据每个分类器生成的输出类型分为三个层次。每个层次激发不同类型的组合规则 [Xu et al., 1992]。

- 摘要层次：每个分类器提供一个单一的类别标签，挑战在于将这些标签组合成一个最终决策。这里常见的策略包括多数表决规则。
- 排序等级：生成的分类器是从候选类别的排序列表中选出的，目标是合并这些列表以生成一个新的单一分类。博尔达计数法是适用于此等级的一种方法的例子。
- 测量水平：每个分类器为每个类别分配一个置信评分。目标是结合这些建议以获得整体置信度量。通常使用的聚合运算符包括求和、乘积和最大值。本文的方法属于后一类。

## 2.3

### 贝叶斯融合

一种常见的测量级组合规则的理论框架是贝叶斯方法。在此框架中，假设每个分类器  $C$  估计一个模式属于类别  $\omega_j$  的后验概率  $P(\omega_j|x)$ 。

通过组合  $L$  个分类器，并假设它们在条件上统计独立，可以推导出以下表达式 [Kittler et al., 1998]：

$$P(\omega_j|x_1, \dots, x_L) = \frac{P(\omega_j) \prod_l P(x_l|\omega_j)}{P(x_1, \dots, x_L)} \propto \prod_l C_{j,l}(x_l) \quad (1)$$

这个公式直接引出了乘积规则。在关于分类器分布的额外假设下，其他规则例如和、最大值、最小值和中位数也可以在这个框架内得到合理的解释。

与传统的集成方法如 bagging 或 boosting 不同，这些方法通常将模型视为黑箱，不明量化不确定性，贝叶斯融合允许从不同模型中结合概率预测，并根据它们的可靠性和方差对其加权。这产生的后验估计反映了每个来源的共识和置信程度。例如，通过结合不同相关水平的分类器，贝叶斯层次模型可以捕捉到它们之间的显式依赖关系，正如 [Trick and Rothkopf, 2022] 所示。

贝叶斯融合的主要优势之一是其处理异质信息源的能力，即使它们表现出不同级别的噪声、粒度或格式。Sander 和 Beyerer (2012) 指出，通过将这些信息源转换为一个共同的概率表示，可以在复杂性或不确定性较高的情境中以稳健的方式进行融合。此外，贝叶斯方法提供了一个统一的框架，用于整合观测数据和专家知识，促进知情决策 [Sander and Beyerer, 2012]。

在实际应用中，例如推荐系统、传感器融合或在线教育论坛的组织，贝叶斯融合可以用来结合不同的分类器或启发式策略，从而得到更精确和自适应的分类。比如，在多个模型评估论坛帖子的重要性或紧迫性的情况下，贝叶斯融合能够整合这些评估，考虑到每个模型的可靠性，并提供一个综合的估计和相关的 uncertainty 度量。

贝叶斯框架的适用性受到理论和实践限制的制约。值得注意的是，假设分类器之间的条件独立性构成了一个重要的简化，其有效性是值得怀疑的。有证据表明，与这种假设相反，在特定情况下，分类器之间的相关性可能对整体性能有利 [Terrades et al., 2009]。

## 3 相关工作

几项研究分析了经典分类系统（如朴素贝叶斯和支持向量机 (SVM)）在不同场景和预处理方法中的性能。结果表明，这些模型存在严重的局限性 [Lima and Maia, 2018, Candido et al., 2019, Candido and Maia, 2025a,b]。

多维分类 (MDC) 在机器学习社区中变得越来越重要，因为它能够处理与多个类别变量相关的实例，每个类别变量代表一个独特的语义维度。Jia 和 Zhang (2024) 对 MDC 范式进行了全面的回顾，讨论了代表性算法以及建模类别之间依赖关系所面临的挑战 [Jia and Zhang, 2024]。

在在线教学论坛的背景下，有效的策划对于促进有意义的讨论和参与者的参与至关重要。An 等人 (2020) 提出了 MUIR 框架，该框架旨在通过资源的互联来丰富 MOOC 中的讨论论坛，提高用户 [An et al., 2020] 的学习体验。

此外，数字策展被认为是媒体教育和数字素养中的一项核心能力。Mihailidis 和 Cohen (2013) 探讨了教授学生策展技能的重要性，使他们能够导航和批判性地评估在线可用的大量信息 [Mihailidis and Cohen, 2013]。这些工作突出了高级分类技术与有效教学实践的交汇点，强调了在在线教育论坛的分析和策展中采用多维方法的重要性。

贝叶斯融合在提高文本分类精度方面得到了广泛探索，尤其是在具有噪声或稀疏数据的环境中。Zhang (2006) 提出了一个基于遗传编程的框架，用于结合来自多个来源的证据，显示出在有限或噪声文本内容的文档分类中具有显著改进的效果 [Zhang, 2006]。

王等人 (2021) 介绍了类别指定主题模型 (CSTM)，该模型假设类别之间存在共享的潜在主题以及每个类别的特定主题。通过贝叶斯推断，该模型在文本分类和摘要任务中表现出优越的性能，尤其是在半监督场景中 [Wang et al., 2021]。

Linghu 等人 (2022) 提出了一种用于少样本分类的贝叶斯证据学习方法，特别是在明确建模不确定性和提高数据 [Linghu et al., 2022] 有限的情况下的泛化能力方面。

随着大型语言模型 (LLMs) 的进步，已经提出了几种方法将这些技术应用于教育文本的分类。Liu 等人 (2024) 开发了 AGKA (基于标注指南的知识增强) 方法，该方法使用 GPT-4 从标注指南中提取知识，从而改善了教育文本中参与度的分类，而无需进行大量的微调 [Liu et al., 2024]。

Koufakou (2023) 进行了一项比较研究，使用 BERT、RoBERTa 和 XLNet 等模型对课程评价进行情感分析和主题分类。结果表明，在情感分析中，RoBERTa 达到了 95.5 % 的准确率，而 SVM 在主题分类中更为有效 [Koufakou, 2023]。

Toba 等人 (2024) 介绍了 BE-Sent 方法，该方法通过将情感分析与 Bloom 分类法结合在课程讨论论坛中执行层次分类，从而允许对教育互动进行更深入的评估 [Toba et al., 2024]。

Wang 等人 (2024) 提出了一种使用 LLMs 进行文本分类的适应性和可靠性范式，证明了在几个分类任务中，少量示例和微调策略可以超越传统方法 [Wang et al., 2024]。

## 4 结果与分析

本文首先对英语进行了测试，并将在未来的工作中开发用于葡萄牙语 (巴西)。

### 4.1 数据集

我们使用了斯坦福 MOOC 论坛帖子数据集<sup>1</sup>，该数据集包含了从 11 门不同在线课程中收集的 29,604 个匿名帖子。学术研究人员可根据请求访问该数据集。这些课程涉及三个主要领域：教育、人文/科学和医学。数据集中的每一条目由三位独立注释者手动标记，以建立一个黄金标准参考 [Agrawal et al., 2015]。

每个帖子从六个维度进行评估：观点、问题、回答、情感、困惑和紧迫性。其中，观点、问题和回答类别被分配二进制标签，而情感、困惑和紧迫性则使用 1 到 7 的量表进行评分，以便于探索以下研究问题：

- 这篇帖子是一个问题吗？(是/否)
- 这篇文章是一个答案吗？(是/否)
- 这篇文章是观点吗？(是/否)
- 这篇帖子表达了什么情感？(1-7)
- 这则帖子表达了多少困惑？(1-7)
- 此帖子需要多紧急的干预？(1-7)

为澄清紧急程度的尺度：得分为 1 的帖子表示没有紧急性，可以安全忽略，而得分为 7 则表示需要教师立即关注的高度紧急消息。为了进行二元分类，我们将 1-7 的紧急程度转换为两个类别，将紧急得分低于 4 的帖子标记为 0 (非紧急)，得分为 4 或更高的标记为 1 (紧急)。此外，我们排除了仅包含数字字符的帖子，最终得到 29,604 个帖子的数据集。其中，23,186 个被归类为非紧急，6,418 个为紧急。关于帖子的分布的详细信息可以在表 1 中找到。

<sup>1</sup><http://datastage.stanford.edu/StanfordMocPosts>

areas / sets	no / yes	opinion	question	answer	confusion	sentiment	urgency
DS1	no	938	9168	9690	6715	1690	9418
	yes	8941	711	189	3164	8189	461
DS2	no	2782	3574	3179	481	1277	3314
	yes	2402	1610	2005	4703	3907	1870
DS3	no	2471	2232	2244	490	196	2053
	yes	559	798	786	2540	2834	977
Education	no	938	9168	9690	6715	1690	9418
	yes	8941	711	189	3164	8189	461
Humanities / Science	no	6016	7671	8002	1358	846	7247
	yes	3707	2052	1721	8365	8877	2476
Medicine	no	6181	6806	5857	1581	1851	6521
	yes	3821	3196	4145	8421	8151	3481

Table 1: 样本在其各自领域中的分布。来自 StanfordMOOC 数据集的帖子被分成几组。第一组，DS1（教育-EDUC115N: 如何学习数学），DS2（医学-SciWrite, 2013 秋季），和 DS3（人文/科学-统计学习, 2014 冬季）。第二组，按学科领域分为教育（EDU）、医学（MED）和人文/科学（H & S）。

## 4.2 结果

Configuration	MD classif.			LLM GPT			Bayesian Fusion			LLM fine-tuning		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Intracourse	0.81	0.80	0.78	0.80	0.78	0.77	0.81	0.80	0.78	0.82	0.82	0.80
Intradomain	0.79	0.79	0.77	0.84	0.78	0.78	0.84	0.79	0.78	0.85	0.81	0.80
Crossdomain	0.73	0.68	0.67	-	-	-	0.73	0.68	0.67	0.77	0.70	0.69

Table 2: 多维分类的表现。所有类别的测试集的平均值。

表 2 展示了在三种不同实验设置下的多维度分类（MD Classif.）、通用大型语言模型（GPT）、贝叶斯融合和大型语言模型微调方法的性能结果：课程内、领域内和跨领域。这些场景旨在模拟训练和测试数据之间不同程度的熟悉性，从而对模型的泛化能力和鲁棒性进行可靠评估。模型在斯坦福 MOOC 数据集的一个子集上进行了评估，该子集包括来自教育、人文/科学和医学领域课程的论坛帖子。

在课程内部场景中，训练和测试在同一课程内进行，模型表现最佳。贝叶斯融合实现了 0.78 的 F1 得分，与 MD 分类器（F1 得分也是 0.78）持平。微调的 LLM 实现了最高的 F1 得分，为 0.80，而通用 GPT 达到 0.77。这些结果表明，当源数据和目标数据相同时，像 MD 分类器这种利用课程特定标签之间相关性的专门方法非常有效。

在域内设置中，测试数据属于与训练数据相同的域（例如，医学）但来自不同的课程时，观察到了性能的可喜变化。微调后的 LLM 再次获得了最高的 F1 分数（0.80），以及最高的召回率（R）（0.81）。然而，贝叶斯融合和通用 GPT 在精确率（P）方面表现出色，都达到了 0.84，表明较低的假阳性率。两者的 F1 分数为 0.78，优于 MD 分类器（0.77）。这表明，当特定课程背景发生变化时，预训练 LLM 的泛化能力开始成为一个重要的区别因素。

跨域场景是最具挑战性的，使用来自与训练域完全不同的域的测试数据。正如预期的那样，在这种情况下，所有模型的性能都大幅下降。例如，MD 分类器的 F1 得分下降到 0.67，而微调的 LLM 则下降到 0.69。在此设置中未测试通用 GPT。贝叶斯融合在此情况下取得了 0.67 的 F1 得分，并在所有模型中获得了最高的准确率（0.77），被证明是应对域间不熟悉性问题的更为稳健的方法。

比较分析表明，在所有场景中没有一个绝对的赢家。然而，贝叶斯融合方法始终表现出色，能够与其他方法相匹敌或胜过其他方法，尤其是在最具挑战性的场景中。结果进一步支持了论文的核心命题：融合是一种可行的替代方案，它结合了不同模型的优势以提高整体性能，而无需承担全面微调的高计算成本。最终的分析表明，尽管大语言模型的微调得分很高，但在需要快速适应的情况下，与所获得的收益相比，可能不足以证明这种成本的合理性。

## 4.3 分析

结果的比较分析揭示了模型的专业化和其泛化能力之间的明显动态。用本地数据训练的 MD 分类器在课程内场景中以 0.78 的 F1 分数脱颖而出。在这种情况下，它能够高效地探索课程特定标签之间的相关性，证明了其高性能。然而，随着上下文逐渐偏离原始训练环境，其性能逐步下降，在域内场景中下降到 0.77 的 F1 分

数，在跨域场景中急剧下降到 0.67。这突出了它对特定上下文数据的强烈依赖性，限制了其在更广泛环境中的稳健性。

另一方面，通用的 LLM (GPT) 展现了显著的泛化能力，特别是在域内场景中，其取得了 0.78 的 F1 分数和 0.84 的准确率。其高准确率表明较低的误报率，这在一个策展系统中非常有价值，可以避免错误的干预。然而，LLM 微调方法尽管在课程内和域内场景中实现了 0.80 的 F1 分数，但并不能证明其高计算成本的合理性。该研究本身对其在快速适应环境中的有效性提出了质疑，因为性能的提升可能无法弥补训练的投入。

在此背景下，贝叶斯融合成为最平衡和稳健的解决方案。通过将通用大型语言模型 (LLM) 的置信度分数与本地训练分类器的置信度分数相结合，该融合在所有场景中实现了具有竞争力和稳定的性能，F1 分数为 0.78 (课程内)、0.78 (领域内) 和 0.67 (跨领域)。在最具挑战性的场景 (跨领域)，贝叶斯融合的性能与 MD 分类器相匹配，但具有更高且更平衡的精准率和召回率。这个结果很重要，因为它表明融合可以继承 LLM 的泛化能力，同时受益于本地分类器的专门化，从而创建一个巩固和可靠的估计。该方法与项目的动机完美契合：通过提供一个有效和高性能的替代方案，结合不同模型的优势，以避免对 LLM 进行频繁且代价高昂的再训练。

整体而言：参考表 7，预计使用领域特定的知识库进行微调会对 LLM 性能产生更大的影响。查看 F1 分数列，我们可以看到，对于该数据集，影响很小。我们打算研究为什么会发生这种情况以及如何构建对 LLM 性能影响更大的本地知识库。

## 5 结论

本文提出并评估了使用贝叶斯融合方法，将通用 LLM 的多维分类与在本地数据上训练的分类器相结合，旨在整理在线课程中的讨论论坛。其主要动机是寻找一种替代 LLM 微调过程的方法，尽管微调强大，但在资源上消耗巨大，对于需要频繁再训练的教育论坛等背景而言并不实际。任务的复杂性由于多维分类的性质而更加突出，其中单个帖子可以同时表达意见、包含问题并显示紧迫性，需要进行复杂的分析。

结果表明，贝叶斯融合方法不仅可行，而且极具竞争力。该融合相比于每个单独作用的分类器表现有所提升，并被证明是一个强有力的替代方案，其性能与微调大语言模型 (LLMs) 相媲美。在高特异性 (课程内) 和泛化 (领域内和跨领域) 的场景中，贝叶斯融合保持了一致的高性能和平衡性，成为所分析方法中最为稳健的方法。它即便在最具挑战性的跨领域场景中也能保持出色的表现，这突显了其在异质数据环境中的有效性。

结论是，贝叶斯融合代表了一种有效且高效的教育论坛管理策略。该技术结合了两者的优势：利用了预训练大模型广泛的知识和本地模型的专业知识，提供了一种准确且可适应的解决方案。通过这种方式，研究展示了一条在教育环境中实施人工智能工具的实用且可持续的途径，绕过了与大语言模型完全重训练相关的高计算成本的障碍，同时不牺牲分类质量。

## References

- B. Jia and Min-Ling Zhang. Multi-dimensional classification: paradigm, algorithms and beyond. *Journal of Artificial Intelligence Research*, 1(3):1–25, 2024.
- Antonio Leandro Martins Candido and José Everardo Bessa Maia. Detecting urgent instructor intervention need in learning forums with a domain adaptation. In *Intelligent Systems Design and Applications*, pages 502–512. Springer Nature, 2023. ISBN 978-3-031-35507-3.
- Y. An, M. K. Chandrasekaran, M. Kan, and Y. Fu. The muir framework: Cross-linking mooc resources to enhance discussion forums, 2020.
- P. K. Mudau and G. Van den Berg. Guidelines for supporting a community of inquiry through graded online discussion forums in higher education. *Education Sciences*, 13(9):963, 2023.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics, 2020.
- P. Mihailidis and J. N. Cohen. Exploring curation as a core competency in digital and media literacy education. *Journal of Interactive Media in Education*, 2013(1):2, 2013.
- S. J. Huang, S. Chen, and Z. H. Zhou. Multi-label active learning: query type matters. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 946–952, Buenos Aires, Argentina, 2015.
- Bin-Bin Jia and Min-Ling Zhang. Decomposition-based classifier chains for multi-dimensional classification. *IEEE Transactions on Artificial Intelligence*, 3(2):176–191, 2021.

- Shivvrat Arya, Yu Xiang, and Vibhav Gogate. Deep dependency networks for multilabel classification. *arXiv preprint arXiv:2302.00633*, 2023.
- Shaodian Zhang, Edouard Grave, Elizabeth Sklar, and Noémie Elhadad. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. In *arXiv preprint arXiv:1603.08458*, 2016.
- Ilias Anagnostopoulos, Sotiris Kotsiantis, et al. Multidimensional group recommendations in the health domain. *Algorithms*, 13(3):54, 2020.
- Satvik Garg. Drug recommendation system based on sentiment analysis of drug reviews using machine learning. *arXiv preprint arXiv:2104.01113*, 2021.
- Jianguo Chen, Kenli Li, Huigui Rong, Kashif Bilal, Nan Yang, and Keqin Li. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *arXiv preprint arXiv:1810.07762*, 2018.
- Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multiclass classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.
- Juri Opitz. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 2024.
- L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435, May/June 1992.
- J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.
- S. Trick and C. A. Rothkopf. Bayesian classifier fusion with an explicit model of correlation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 2282–2310. PMLR, 2022.
- J. Sander and J. Beyerer. Bayesian fusion: Modeling and application. Technical report, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), 2012.
- Oriol Ramos Terrades, Ernest Valveny, and Salvatore Tabbone. Optimal classifier fusion in a non-bayesian probabilistic framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1630–1644, Sept 2009. doi:10.1109/TPAMI.2008.224.
- João Marcos Carvalho Lima and José Everardo Bessa Maia. A topical word embeddings for text classification. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 25–35. SBC, 2018.
- Antonio Leandro Martins Candido, Corneli Gomes Furtado Júnior, and Francisco Aislan da Silva Freitas. Moderação inteligente de mensagens em ambientes virtuais de aprendizagem para alunos privados de liberdade. In *Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019)*, VIII Congresso Brasileiro de Informática na Educação (CBIE 2019), pages 1838–1847, 2019. doi:10.5753/cbie.sbie.2019.1838.
- Antonio Leandro Martins Candido and José Everardo Bessa Maia. Domain adaptation with diet-rasa and xlnet in urgent post detection. In *Hybrid Intelligent Systems*, pages 41–51. Springer Nature Switzerland, 2025a. ISBN 978-3-031-78925-0.
- Antonio Leandro Martins Candido and José Everardo Bessa Maia. Short text classification of social groups focused on courses. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. SBC, 2025b.
- B. Zhang. *Intelligent Fusion of Evidence from Multiple Sources for Text Classification*. PhD thesis, Virginia Tech, 2006.
- F. Wang, J. L. Zhang, Y. Li, K. Deng, and J. S. Liu. Bayesian text classification and summarization via a class-specified topic model. *Journal of Machine Learning Research*, 22(89):1–48, 2021.
- X. Linghu, Y. Bai, Y. Lou, S. Wu, J. Li, J. He, and T. Bai. Bayesian evidential learning for few-shot classification, 2022.
- S. Liu, S. Liu, L. Sha, Z. Zeng, D. Gasevic, and Z. Liu. Annotation guidelines-based knowledge augmentation: Towards enhancing large language models for educational text classification, 2024.
- A. Koufakou. Deep learning for opinion mining and topic classification of course reviews, 2023.
- H. Toba, Y. T. Hernita, M. Ayub, and M. C. Wijanto. Bloom-epistemic and sentiment analysis hierarchical classification in course discussion forums, 2024.
- Z. Wang, Y. Pang, Y. Lin, and X. Zhu. Adaptable and reliable text classification using large language models, 2024.
- Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. Youedu: Addressing confusion in mood discussion forums by recommending instructional video clips. In *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015*, pages 297–304, 2015.