研究论文自动分类 面向可持续发展目标：基于布尔查询的计算框架

Sahil Dewani[1] 和 Kiran Sharma[1,2*]

[1]工程与技术学院，*BML Munjal*大学，古尔冈，印度[2]高级数据与计算科学中心，*BML Munjal*大学，古尔冈，印度 通讯作者*：*kiran.sharma@bmu.edu.in*

摘要

各学科领域的学术出版物迅速增长，使得系统评估研究对联合国可持续发展目标（可持续发展目标）的贡献变得愈发困难。由研究专家手动对研究论文进行领域分类极不现实，因为出版物数量庞大、耗时且成本高，而且人工分类的结果可能不一致。为应对这些挑战，本文提出一种基于专家整理的布尔查询映射、按照可持续发展目标对研究论文进行分类的自动化规则型计算模型。该系统包含用于输入数据和显示结果的基于Web的界面、用于高吞吐量处理的后端应用程序编程接口，以及一个基于Python 的分类引擎，后者使用结构化布尔表达式处理书目元数据（标题、摘要和关键词）。该框架支持单篇论文分类和批量分类，并提供清晰易懂的输出，明确展示触发各项可持续发展目标分配的查询部分。对大型书目数据集进行的实验测试表明，该系统每小时可处理数千条研究记录，并得到可复现且一致的结果。对于希望系统分析研究与可持续发展目标契合度的机构、研究人员和政策制定者而言，该方法提供了一种可行的解决方案，无需使用输入和输出难以理解的机器学习模型。

关键词： 可持续发展目标；自动分类；布尔查询匹配；研究分析；文本挖掘。

1. 引言

联合国2030年议程是一项全球计划，以17项可持续发展目标为基础，旨在帮助解决各国面临的社会、经济和环境问题，例如消除贫困、改善民众健康、应对气候变化以及实现长期环境可持续性 [1]。科学研究与创新是实现这些目标的关键，因为它们能够提供证据依据，助力制定明智的政策、推动技术进步并改造社会。与此同时，可持续发展目标背景下的学术研究日益受到大学、资助机构、政策制定者和国际组织的重视。

尽管人们日益关注以可持续性为基础的研究，但系统综述研究论文对可持续发展目标的适用性已成为一项重大挑战。各领域的学术文献数量急剧增长，由相关领域专家进行人工筛选已变得极不切实际，既耗时又昂贵，还容易受到主观解读的影响。这些限制凸显了采用自动化、可扩展且透明的流程的必要性，以便以一致且可重复的方式将大型研究语料库映射到可持续发展目标类别。

1.1 采用基于查询和规则驱动的方法映射可持续发展目标

已采用的最广泛的可持续发展目标分类策略之一是基于规则或查询驱动的方法。这些方法基于预先设定的布尔查询关键词，依据官方可持续发展目标、可持续发展目标指标和政策文件制定 [3]。Aurora Universities Network 是最早开展此类机构倡议的组织之一，旨在制定结构化的可持续发展目标搜索查询，以便在书目数据库中识别与特定可持续发展目标相符的研究成果 [4]。

同样，爱思唯尔发布了多个版本的可持续发展目标映射查询集，这些查询集被广泛用于大规模文献计量分析和全球大学排名工作[5]。最新版本——爱思唯尔 2023年可持续发展目标映射数据集——可公开访问，提供经过整理的布尔查询定义，以及涵盖数十年学术成果的大规模带可持续发展目标标签的出版物元数据[9]。上述数据集可通过以下地址获取：

https://elsevier.digitalcommonsdata.com/datasets/y2zyy9vwzy/1 已被使用并成为复现可持续发展目标相关研究分析的重要参考资源。

在基于查询的方法中，主题关键词通常与符合可持续发展目标的行动导向术语结合使用。例如，可以借助贫困、不平等和社会保障等相关术语的组合，识别与可持续发展目标1（无贫困）相关的出版物 [6]。这些方法的主要优势在于透明度高，每项可持续发展目标分配都可以明确追溯到研究文本中的直接关键词。这种可解释性使其被应用于机构研究评估和国际基准模型中 [1]。

1.2 混合可持续发展目标分类与基于机器学习的方法

除基于规则的方法外，机器学习和自然语言处理（NLP）技术也被用于探索自动化可持续发展目标分类。基于标注数据集训练的监督学习模型能够捕捉超越精确关键词匹配的语义关系；当研究论文采用多样化或隐含的术语时，这有助于提高召回率 [7]。其中大多数模型都基于借助基于查询的系统创建的可持续发展目标标注语料库，例如爱思唯尔可持续发展目标映射数据集 [9], 这一最受欢迎的训练与评估数据集之一。

近期研究提出了集成方法和混合方法，将多个分类器的输出结合起来，或结合基于规则的方法与基于学习的方法 [8]。*Text2sdg*展示了将各种可持续发展目标标注系统（如爱思唯尔和Aurora映射）应用于并行研究研究方向的一致性 [7]。尽管这些方法前景可期，但可能需要大量计算资源和大规模参数优化。

基于机器学习的系统的一大缺点是缺乏可解释性。分类决策通常无法在单个文档层面清晰呈现，这限制了它们在审计敏感型应用中的适用性；在这类应用中，必须对可持续发展目标分配作出合理说明。

1.3 比较研究与揭示的研究空白

比较研究发现，根据所选方法的不同，可持续发展目标分类结果存在显著差异。对独立查询集的比较研究

和机器学习模型表明，替代方法可能会对同一研究语料库产生显著不同的可持续发展目标分布[2]。这些不一致性凸显了可持续发展目标映射缺乏统一规范，也强调了方法透明性的重要性。

多个来源强调，可解释性、可复现性和改进可行性是可持续发展目标相关研究评估工具极为重要的必要条件。尽管机器学习解决方案在某些情况下更具实用性，但对机构而言，基于查询的系统更易于使用和审计。尽管如此，目前有若干基于查询的系统以静态数据集或程序的形式发布，这限制了非技术背景的其他利益相关者对它们的应用。

1.4 研究动机与贡献

这些考量促使本研究采用清晰的、基于查询的方法来实现可持续发展目标的自动分类。所提出的框架采用公开可用且经专家策划的可持续发展目标查询集，例如爱思唯尔2023年可持续发展目标映射 [9] 并将其整合到一个模块化网络计算系统中。

本研究的主要贡献如下：

1. 构建一个可扩展流水线，通过易于使用且直观的界面采集和处理研究元数据。
2. 部署一个布尔分类引擎，用于对研究文本执行多个可持续发展目标相关子查询。
3. 提供交互式视觉分析，使用户能够探索可持续发展目标分布及相关的分类置信度分数。

2. 方法

2.1 总体系统架构

该架构由三个层级构成，即表示层、集成层和分类引擎。数据流从用户提供的研究元数据开始，最终生成排序后的可持续发展目标输出和可视化分析（见图 1）。

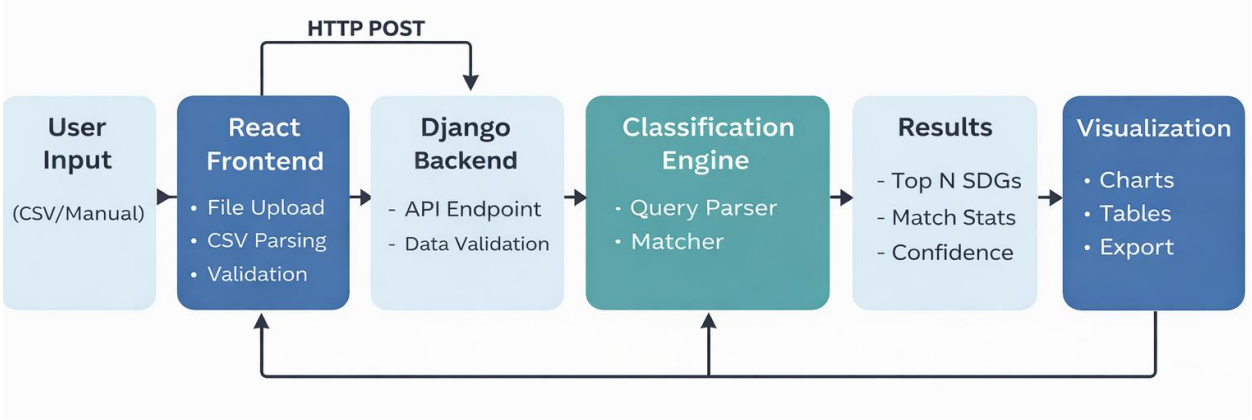


图 *1. O*　　可持续发展目标分类框架的整体系统架构和数据流

2.1.1 表示层（前端界面）

表示层是一个基于 Web 的前端，采用 React 和现代用户界面库。它为探索性和大规模的可持续发展目标分析提供了便捷的界面。前端由两个主要工作流程提供支持：

- 单篇论文分类，用户需要手动填写论文标题、摘要和关键词。
- 批量分类，用户提交包含多条记录的结构化书目文件（例如 CSV 或 TSV）。

用户可设置的参数包括要返回的可持续发展目标数量（Top-N 选择），并可即时查看分类结果。结果以可持续发展目标排序列表、置信度指标和交互式图表的形式呈现。图 2 和图 3 展示了界面布局。

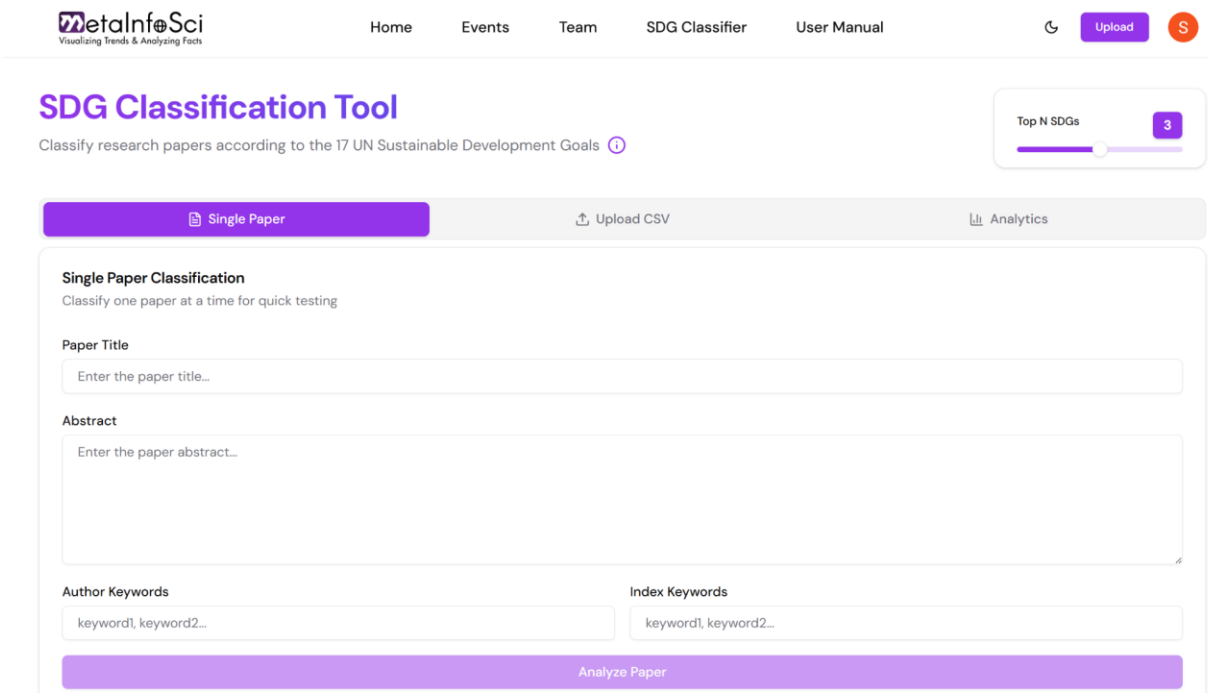


图 *2*。 单篇论文分类：可持续发展目标分类网络门户。

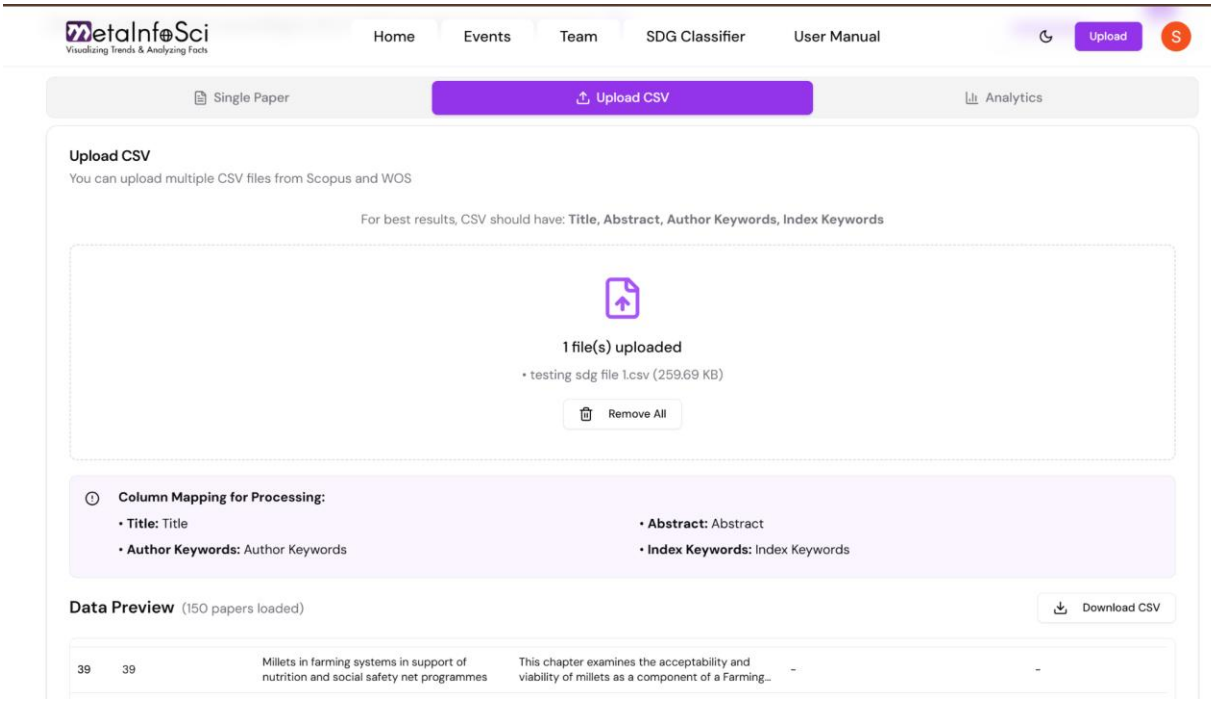


图 *3*。 批量分类（上传 *CSV*）： 可持续发展目标分类网络门户。

2.1.2 集成层（后端 API）

集成层是分类引擎与用户界面之间的接口。它通过基于 Python 和 Django 的 RESTful API 部署。后端 API 用于验证传入的请求、解析上传的书目文件、规范化元数据字段、调用分类引擎，并以 JSON 格式返回结构化的可持续发展目标结果。

(a) 数据摄取与预处理

系统支持 Scopus 和 Web of Science 等热门索引数据库的常见书目格式。这些格式为 CSV 和 TSV 文件，且必须包含以下字段：

- 论文标题
- 摘要
- 作者关键词
- 索引关键词

使用表头匹配启发式方法自动检测列。例如，任何包含“abstract”（不区分大小写）一词的列都会映射到摘要字段。如果自动检测失败，用户可以手动映射列。

导入后，每篇论文的记录都会通过串联所有可用的元数据字段，整合为单一文本表示。文本预处理刻意保持轻量：

- 转换为小写
- 移除标点符号（布尔运算符除外）
- 不进行全局停用词移除
- 不进行词干提取或词形还原

这种设计选择与布尔型可持续发展目标查询的结构一致，后者依赖明确的关键词模式和通配符运算符（例如 sustainab*），而非统计语言建模。

(b) 可持续发展目标查询库构建

该分类引擎基于权威且由专家精心编制的可持续发展目标查询集构建。本研究的主要数据来源是爱思唯尔可持续发展目标映射数据集（2023），所有人均可通过开放获取方式从以下来源获取：https://elsevier.digitalcommonsdata.com/datasets/y2zyy9vwzy/1。它是一个结构化数据集，采用

布尔查询的形式，与可持续发展目标及其指标相一致，并且目前已被机构研究评估系统广泛采用[9]。

一个可持续发展目标由若干布尔子查询表示，每个子查询对应该目标的一个主题方面。例如，可持续发展目标1（无贫困）包含涉及消除贫困、收入差距和社会保障的子查询。

每个子查询可能包含：

- 逻辑运算符（AND、OR）
- 通配符
- 地理限制（必要时）

2.1.3 分类引擎

(a) 布尔匹配逻辑

随后，分类引擎以循环方式运行，依次处理每个可持续发展目标，并将与之相关的所有子查询与整合后的论文文本进行匹配检查。当子查询的布尔条件得到满足时，即可视为匹配成功。

对于每个可持续发展目标，引擎都会计算：

- 匹配数：与给定可持续发展目标匹配的布尔子查询数量。
- 总数：为该可持续发展目标定义的布尔子查询总数。

(b) 可持续发展目标评分与排名

通过计算比例得出可持续发展目标相关性评分：

$$\text{评分} = \text{匹配项} / \text{总数}$$

每个可持续发展目标都根据此评分进行排名。系统默认显示每篇论文排名前三的可持续发展目标，但用户可以修改此设置。最重要的是，系统还会返回详细的匹配统计数据，用户可以查看哪些子查询生成了各项可持续发展目标的归类结果。

(c) 可视化分析

分析模块将分类结果转换为可视化摘要，如图4所示。它支持：

- 可持续发展目标频次条形图、饼图和圆环图。
- 分类论文汇总表。
- 可报告、可导出的结果文件。

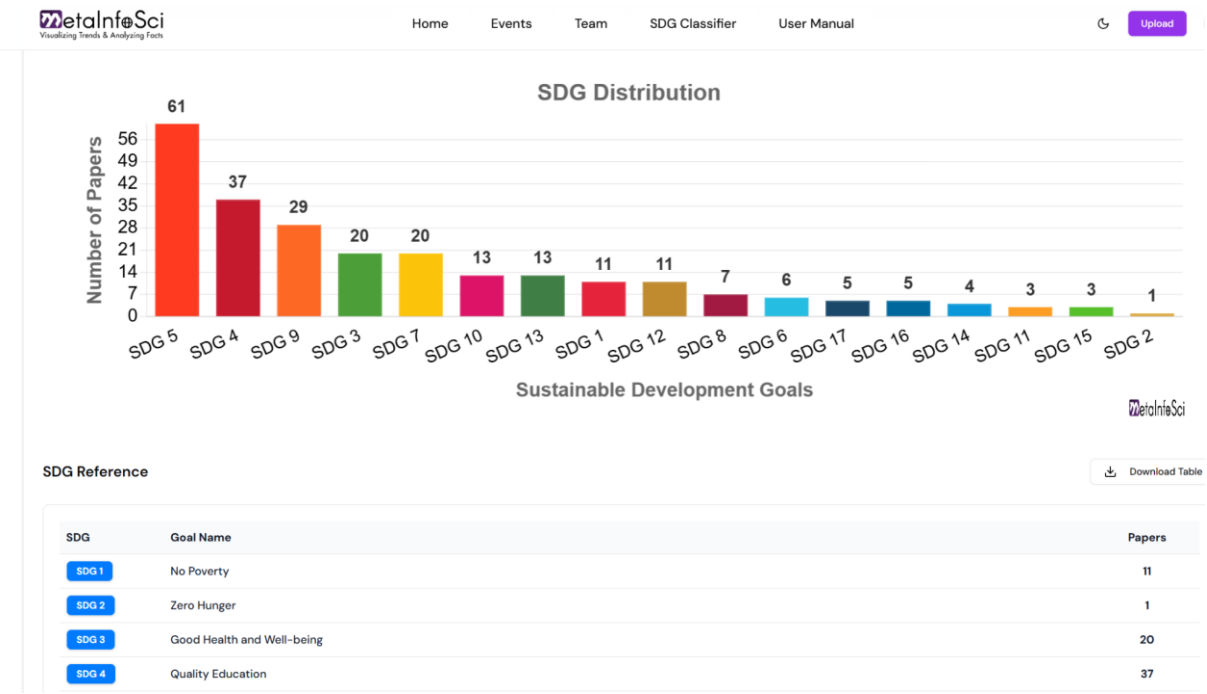


图*4.*分类论文按可持续发展目标划分的分布可视化，展示样本文献语料库中各可持续发展目标的出现频率。

2.2 系统概述与设计原则

拟议系统计划构建为一个高度可定制、基于查询的系统，能够根据联合国可持续发展目标自动对研究论文进行分类。以下四项原则构成了该方法的核心价值：

- 透明度：所有可持续发展目标分配均须落实为明确的关键词或布尔规则匹配。
- 可伸缩性：系统应既能进行单篇论文分析，也能进行大规模批量处理。
- 可复现性：在使用相同输入时，可持续发展目标输出应始终保持一致。
- 可扩展性：查询定义和分类逻辑必须能够在不修改底层代码的情况下进行维护。

为满足这些需求，该框架基于分层式 Web 架构，将用户交互、数据处理和分类逻辑分离。正是这种分离，使布尔分类引擎能够脱离用户界面运行，并可在不同的分析任务中重复使用。

2.3 实验设置与数据集构建

该研究没有在规模庞大且噪声较多的语料库上测试系统，而是构建了一个受控评估集，从而能够以非常准确且易于理解的方式评估分类性能。研究具体选取了三个可持续发展目标进行分析：

- 可持续发展目标1：无贫穷
- 可持续发展目标4：优质教育
- 可持续发展目标5：性别平等

针对每个可持续发展目标，分别在 Scopus 数据库中手动识别了 50 篇文章。由此得到一个包含 150 篇研究论文的评估语料库，其可持续发展目标相关性已知。所有论文均采用可持续发展目标分类系统进行分类，并设置了三种滑块设置：

- 滑块 = 1：返回 Top-1 可持续发展目标
- 滑块 = 2：返回排名前二的可持续发展目标
- 滑块 = 3：返回排名前三的可持续发展目标

这种受控设计能够有条不紊地分析增加返回的可持续发展目标数量对分类准确率及其覆盖率的影响（见图5）。

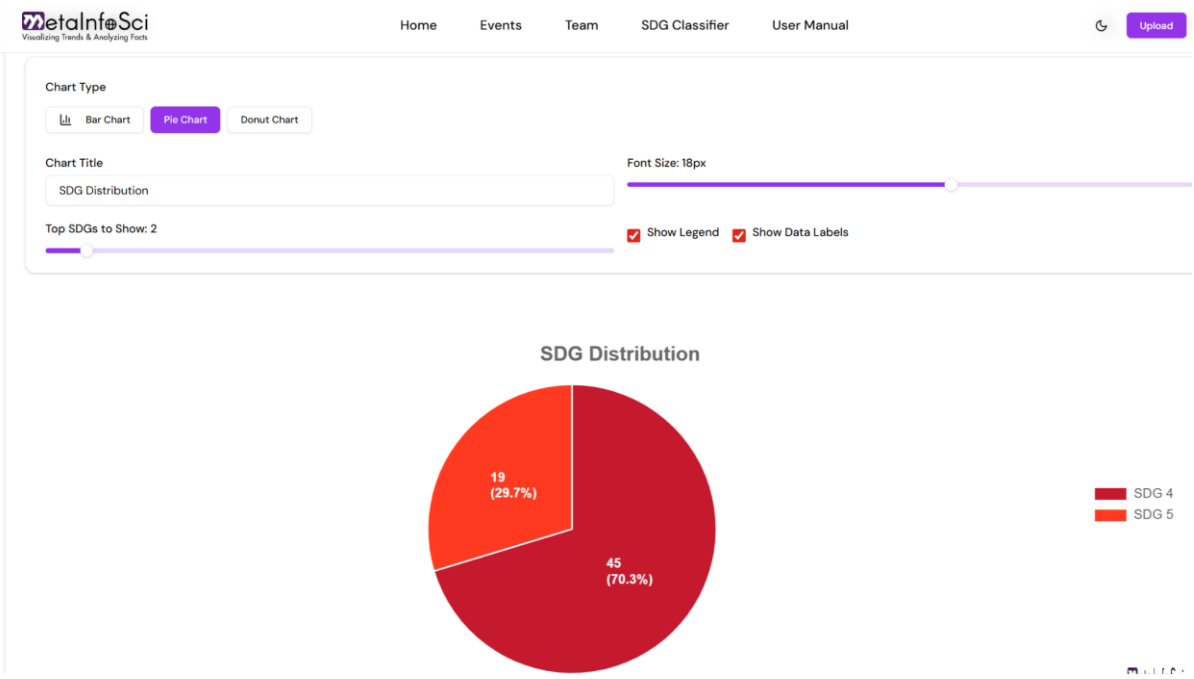


图5. 可持续发展目标 将滑块设为排名前二的可持续发展目标时，以饼图可视化可持续发展目标分布。

3. 结果与讨论

本部分对所提出的可持续发展目标分类框架在受控实验条件下进行实证分析。分析依据分类在不同滑块设置（Top-1、Top-2 和 Top-3 可持续发展目标输出）下的表现、与人工分配的可持续发展目标标签之间的定量一致性、定性错误倾向，以及研究结果的可解释性展开。

3.1 可持续发展目标优选

(a) Top-1 可持续发展目标分配

当滑块设置为仅返回排名最高的可持续发展目标时，该框架采取保守但准确的策略。在大多数情况下，返回的可持续发展目标与人工指定的真实标签直接对应。图6展示了 Top-1 设置下可持续发展目标的分布。分布高度集中于所评估的三个可持续发展目标（可持续发展目标1、可持续发展目标4和可持续发展目标5），这表明布尔评分系统能够准确识别给定论文中占主导地位的可持续性主题。这种形式侧重于准确性而非覆盖率，因此特别适用于需要使用一个主要可持续发展目标标签的场景，例如摘要仪表板、机构报告或高层次基准测试。

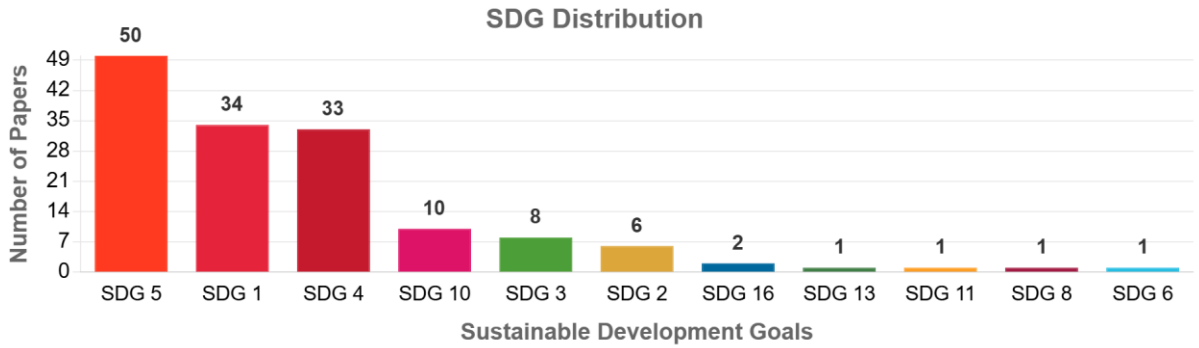


图*6.* 每篇论文的 *Top-1* 可持续发展目标。

(b) Top-2 可持续发展目标分配

滑块值越高，系统就能为每篇论文显示越多相关可持续发展目标。这种结构体现了可持续发展研究的多维性，尤其是对于横跨多个主题领域的论文。图7表明，尽管主要可持续发展目标（可持续发展目标1、可持续发展目标4、可持续发展目标5）仍占据主导地位，但次要可持续发展目标

开始变得更加常见。教育导向型论文（可持续发展目标4）在性别平等议题上往往与可持续发展目标5存在次要关联，而贫困导向研究（可持续发展目标1）则可能与教育或健康相关目标相联系。这种安排在召回率与可解释性之间取得了合理的平衡，因此是探索性分析和机构自我评估的有效分析工具。

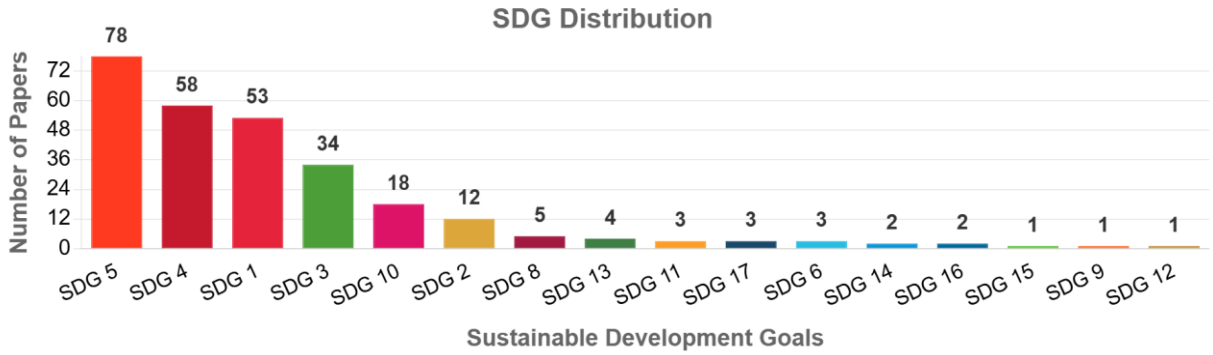


图7。 每篇论文排名前二的可持续发展目标。

(c) 排名前三的可持续发展目标分配

将滑块调至 3 时，系统会为一篇论文给出尽可能多的可持续发展目标，即 3 个，从而最全面地呈现其可持续发展目标相关性。如图8所示，这种设置还有一个额外优势：能够捕捉许多其他背景性和次要关联。虽然这些关联并非主导因素，但在可持续发展研究中仍具有重要意义。尽管目标之间的重叠会更加明显，该框架的归一化评分机制仍能让主要可持续发展目标与较弱或次要的关联清晰区分开来。该框架尤其适用于战略分析，因为深入了解次要和第三层级的可持续发展相互关联会带来益处。

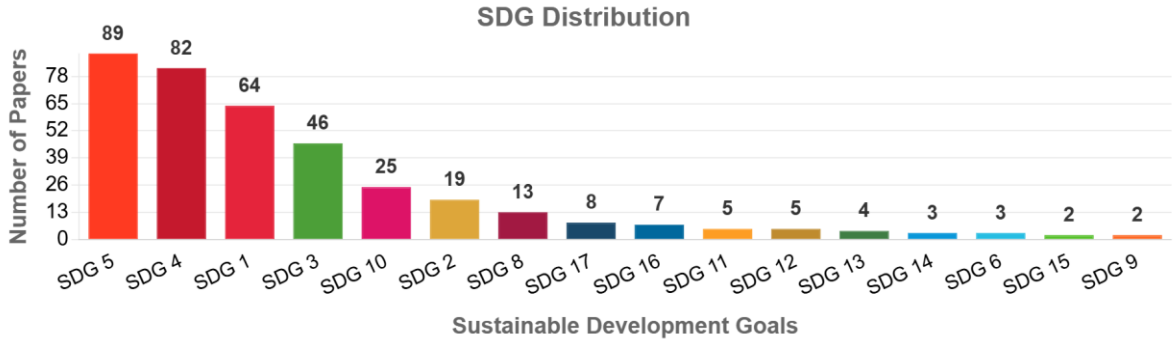


图8。 每篇论文排名前三的可持续发展目标。

3.2 定量评估指标

为定性评估分类性能，我们在不同的滑块设置下，评估了人工指定的可持续发展目标真实标注与系统输出之间的一致性。当真实标注中的可持续发展目标位于系统返回的前 N 个可持续发展目标中时，即认为该滑块等级下的预测正确。

表*1*。 可持续发展目标*1*、可持续发展目标*4*和可持续发展目标*5*的评估结果，包括每个滑块取值下的正确分类数量，以及未识别出相关可持续发展目标的论文数量。

SDG	总计论文	正确 @ Top-1	正确 @ Top-2	正确 @ Top-3	正确 @ 前四名	正确 @ 前五名	No 识别 n	准确率 @ Top-1 (%)	准确率 @ Top-2 (%)	准确率 @ 前五名 (%)
可持续发展目标1	50	32	46	50	50	50	0	64	92	100
可持续发展目标4	50	32	45	46	46	46	3	64	90	92
可持续发展目标5	50	39	49	50	50	50	0	78	98	100
总体而言	**150**	**103**	**140**	**146**	**146**	**146**	**3**	**68.67**	**93.33**	**97.33**

3.3 定性分析与错误模式

对误分类和边界案例的定性审查表明，主要差异更多是由可持续发展目标范围的重叠造成的，而非关键词的错误匹配。例如，一些以教育为重点的文章也关注社会包容或性别平等，因此会被次要地匹配到可持续发展目标5。同样，教育或健康主题也可能与贫困研究重叠。值得注意的是，系统会完整呈现此类案例，因为它会展示相同的布尔子查询和归一化分数，帮助用户理解为何系统会推荐某个可持续发展目标。

3.4 可解释性与实践意义

所提出框架的一项关键优势在于其方法清晰且一致。分类过程基于明确的布尔查询匹配，每项可持续发展目标分配均遵循预先制定并经专家审定的规则。尽管最终用户无法直接查看各个

匹配的子查询，但底层分类逻辑在反复执行时仍保持系统性、确定性和稳定性。

对于每篇研究论文，系统根据匹配的子查询数量及其归一化匹配分数对可持续发展目标进行排序。这种受规则约束的机制能够最大限度地减少差异，并支持在大型研究语料库中可靠地识别可持续发展目标。

4. 讨论与结论

本文提出了一个清晰且可推广的计算框架，用于依据联合国可持续发展目标对研究论文进行自动分类。受学术文献快速增长以及近年来可持续性相关科研评价兴起的启发，本文提出的系统旨在解决人工分类和不透明自动化解决方案的主要缺陷。

具体而言，该框架采用由专家编制的布尔查询集，即爱思唯尔可持续发展目标映射数据集（2023）[9], ，从而能够大规模进行可复现、可解释的可持续发展目标分配。系统架构由用户友好的 Web 界面、应用程序编程接口、功能强大的后端应用程序以及专用的布尔分类引擎组成；后端应用程序能够评估文档中的数百个可持续发展目标相关子查询。

总体而言，实验结果表明，这种基于查询的框架能够在精确率、覆盖率和透明度之间实现适当平衡。基于滑块的控制功能允许用户根据不同的分析需求定制系统输出，既可采用保守的单标签分配，也可开展覆盖更广的多可持续发展目标分析。

4.1 主要发现与观察

在所分析的数据集中观察到的可持续发展目标分布显示，研究成果明显集中于特定主题领域。与健康、基础设施、教育和气候措施相关的研究似乎最为常见，这意味着这些领域仍是学术研究人员关注的重点。

其次，输入元数据的完整性和质量对分类结果有着重大影响。研究发现，包含摘要和关键词的文章比仅包含标题的文章更容易被明确归类，且涵盖的可持续发展目标范围也最广。上述发现凸显了详尽的文本数据对于可靠的自动化可持续发展目标映射的重要性。

第三，定性核查表明，基于规则的分类结果在大多数情况下与专家判断相近。值得注意的是，该框架完全透明，因为它让用户清楚了解每项可持续发展目标分配所使用的具体查询组件。

与基于机器学习的可持续发展目标分类方法相比，所提出的基于查询的框架具有明显的实际优势。该系统不依赖不透明的模型行为，而是遵循由专家定义的布尔查询所支撑的确定性规则流程。这确保了分类结果稳定且一致；在机构研究评估、资助评估和政策导向分析中，清晰度与方法可靠性至关重要，因此这一点尤为宝贵。

4.2 局限性

尽管该框架具有若干优势，但仍存在一些局限性。目前，该系统仅适用于英文研究元数据。因此，尽管以可持续发展目标为高度导向，以其他语言发表的出版物可能无法得到恰当分类。这一限制源于底层布尔查询定义对语言的依赖，而这些定义目前以英文维护。

5. 参考文献

[1] Y. Kashnitsky、D. Meier、R. Mata 和 D. U. Wulff，《评估识别支持联合国可持续发展目标的研究的方法》，*Quantitative ScienceStudies*，第 5 卷，第 2 期，408–427 页，2024 年。[2] A. C. S. Armitage、B. Jayabalasingham、R. M. K. N. H. Gunasekara 和 E. Sugimoto，《绘制与可持续发展目标相关的学术出版物：独立的文献计量方法能得出相同结果吗？》，*Quantitative Science Studies*，第 1 卷，第 3 期，1092–1108 页，2020 年。[3] D. U. Wulff、D. Meier、R. Mata 和 Y. Kashnitsky，《运用新颖数据和集成模型改进可持续发展目标的自动标注》，*Sustainability Science*，第 19 卷，1773–1787 页，2024 年。

[4] Aurora Universities Network, 《将研究产出映射到可持续发展目标（SDGs）的搜索查询》，2020 年。[在线]。网址：https://aurora-network-global.github.io/sdg-queries/[5] B. Jayabalasingham、S. Ramanathan 和 E. Sugimoto, 《识别支持联合国可持续发展目标的研究》，*Mendeley Data*，版本 1，2019 年。doi:10.17632/87j5d7nh3b.1[6] 联合国, 《变革我们的世界：2030年可持续发展议程》，联合国大会，纽约，美国，2015 年。[7] D. Meier、R. Mata、D. U. Wulff 和 Y. Kashnitsky, 《text2sdg：用于通过文本监测可持续发展目标的 R程序包》，*The R Journal*，第 17 卷，第 50 页–70 页，2025 年。[8] D. Meier、D. U. Wulff、R. Mata 和 Y. Kashnitsky, 《比较和整合用于科研评价的可持续发展目标分类系统》，*Sustainability Science*，第 18 卷，第 1459 页–1474 页，2023 年。[9] 爱思唯尔 Digital Commons, 《可持续发展目标（SDG）映射数据集》，2023 年。doi:10.17632/y2zyy9vwzy.1[在线]。网址：https://elsevier.digitalcommonsdata.com/datasets/y2zyy9vwzy/1