arXiv:2601.01449v1 [cs.CL] 4 Jan 2026

基于 Open Legal Data 的德国法院判决分段与处理

Harshil DARJI a b, 1, Martin HECKELMANN a, Christina KRATSCH [a] 和 Gerard DE MELO [b]

[a]柏林应用技术大学，德国[b]哈索·普拉特纳研究所 / 波茨坦大学，德国

ORCiD 标识：Harshil Darji https://orcid.org/0000-0002-8055-1376, Martin Heckelmann https://orcid.org/0000-0003-1696-6849, Christina Kratsch https://orcid.org/0000-0003-0565-1112, Gerard de Melo https://orcid.org/0000-0002-2930-2059

摘要。结构化法律数据的可用性对于推动德国法律体系的自然语言处理（NLP）技术发展至关重要。Open Legal Data 是使用最广泛的数据集之一，提供了大规模的德国法院判决汇集。尽管这一原始数据集中的元数据结构一致，判决文本本身的格式却不统一，且往往缺少明确标记的章节。可靠地分离这些章节，不仅对修辞角色分类至关重要，也有助于检索和引文分析等下游任务。在这项工作中，我们基于官方 Open Legal Data 数据集，介绍了一个经过清理和分段的德国法院判决数据集，共**251,038** 份。我们系统地分离了德国法院判决中的三个重要部分，即裁判主文（判决主文）、事实部分（案件事实）和裁判理由（司法推理）；这些部分在原始数据集中通常呈现不一致。为确保提取过程的可靠性，我们采用 Cochran公式，以**95%** 的置信水平和**5%** 的误差范围，抽取了具有统计代表性的**384** 个案件随机样本，并逐一人工核验这三个部分是否均被正确识别。我们还将上诉权利告知（上诉须知）提取为单独字段，因为它属于程序性指示，并非判决本身的一部分。最终语料库以 JSONL 格式公开，便于获取，可用于进一步研究德国法律体系。

关键词。法律自然语言处理、德国法律体系、法院判决、Open Legal Data、数据集

1. 引言

法律数据量的不断增加以及德国法院判决数字化程度的日益提高，为推动法律领域的自然语言处理带来了新的机遇。诸如 *OpenLegal Data* [1] 和 *GerDaLIR* [2], 等大规模项目提供了跨越不同法院和年份的数千份法院判决的访问权限。然而，此类资

[1]通讯作者：Harshil DARJI，Harshil.Darji@HTW-Berlin.de。

源在法律自然语言处理中的可用性受到格式不一致、结构标记缺失以及HTML结构各异的影响 [3,4]。

德国法院判决遵循一致的法律写作格式，其中裁判主文 提供判决的执行部分，事实部分 概述事实和诉讼经过，而裁判理由阐述判决理由。对于法律案例检索 [5], 引文分析和修辞角色分类 [6], 等下游任务而言，可靠地区分这些部分是先决条件 [7]。缺乏清晰的分段会直接影响检索：理由与事实经过重叠可能导致错误匹配；也会影响引文分析，因为被引法条的重要性取决于它出现在裁判理由还是裁判主文中。在检索增强生成（RAG）系统中，感知章节的分块有助于提升可解释性，并防止模型混淆论证内容与裁判主文内容。

关于德国法院判决的既有研究也强调了结构化且机器可读的法院判决对于检索和分类的重要性 [8,9,10],，但现有数据集仍缺乏一致的章节边界。本文介绍一个数据集，其中包含从 Open Legal Data 官方数据转储中提取的 251,038 份德国法院判决。我们采用基于正则表达式的基于规则的提取方法，分离判决的三个部分，并识别法条引文和案例引文，以便在语料库中建立关联。为确保可靠性，我们使用 Cochran公式 [11] ，设定 95% 的置信水平和 5% 的误差幅度，随机抽取 384 个案件并进行人工核验。在该样本中，**97.40%** (±1.59%) 的案件处理正确，证实了提取流程的可靠性。该数据集以 JSONL 格式通过 Hugging Face数据集[2] 公开提供。

2. 相关工作

Ostendorff等人 [1] 推出的*Open Legal Data*平台创建了一个大型语料库，收录了可公开访问的德国法院判决，并包含法院标识符和日期等元数据。然而，尽管元数据结构一致，判决文本却以异构的 HTML 格式存储，缺乏标准化标记。此后，已有多个语料库基于这一来源构建。Glaser等人 [12] 创建了一个约含 100,000 份德国法院裁判的集合，用于评估摘要方法，并将文本切分为适合训练摘要模型的单元，但未强制遵循法律章节边界。Wrzalik等人 [2] 推出了 *GerDaLIR*，这是一个信息检索基准，依据判决之间的引文，将 123,000 个查询段落与 131,000 份文档关联起来。Darji等人 [3] 探讨了德国法院判决与法律条文之间的语义相似性，随后又发布了一个数据集，其中包含 1,944 条由人工标注、摘自德国法院判决的法律引文 [13]。尽管这些工作推动了德国法律领域的检索和相似性建模，但它们并未提供一个完全结构化的语料库，无法在所有判决中分别标出裁判主文、事实部分和裁判理由。

除数据集外，其他研究也指出，为法律文本添加清晰的结构有助于实现计算应用。Heckelmann [14] 探讨了如何将法律协议表示为可执行的智能合约，从而为其机器执行确立法律框架。尽管研究重点是如何将智能合约纳入

[2] harshildarji/openlegaldata。 https://huggingface.co/datasets/harshildarji/openlegaldata

现有的民法框架，动机却是相同的：结构是让法律数据能够用于下游任务的先决条件。

3. 数据集

我们根据官方 Open Legal Data 数据转储文件（截至 *2022-10-18*[3]）创建了数据集。虽然该原始数据集以 JSONL 格式提供，但判决文本嵌入在具有不同 HTML 结构的 `content` 属性中。我们的抽取流程专注于分离三个章节、规范化法院元数据，以及抽取法规引文和案例引文。

3.1. 抽取流程

我们首先使用 HTML 解析器解析 content 属性中的 HTML 字符串，并遍历限制为 p、h1–h4、td 和自定义 rd 标签的可见元素。随后通过合并空白字符对文本进行规范化，并跳过空行和重复行。

接下来，我们通过使用 Open Legal Data[4] 的公共 API（端点 /api/states/ 和 /api/cities/）解析城市和州标识符，来规范化法院元数据。缺失条目设为 未指定。

然后，我们通过识别特定的固定词汇标题来执行章节边界检测。每个章节的这些章节标题均通过两种精确的行级模式进行识别，并以不区分大小写的方式应用，同时要求整行匹配。模式如下：

- 紧凑形式： ^\s*<标记>\s*: *$
- 字母间隔形式： ^\s*<标 记>\s*: *$

章节词汇固定为裁判主文、事实部分、裁判理由和理由部分[5]。对于每一项，我们都会测试紧凑形式和字母间隔形式（例如，裁判主文与 t e n o r）。在遇到第一个标题之前，当前章节默认为裁判主文；非标题行则附加到当前章节中。这一假设符合德国法律文书起草惯例：所有法院判决通常都以裁判主文开头。

德国法院判决通常遵循两种起草模式。判决（经庭审作出的判决）会明确列出全部三个部分：裁判主文、事实部分和裁判理由。裁定 （未经庭审作出的裁定）虽然以裁判主文开头，但通常包含理由部分，并按罗马数字细分，其中理由一 对应事实部分 ，而理由二 对应裁判理由。如果没有细分，整个理由部分 均视为裁判理由。这一逻辑贯穿整个抽取流程。

此外，我们还识别并提取上诉权利告知，其中提供了有关可用上诉途径的程序性指示。尽管它属于已发布判决的一部分，

[3]Open Legal Data 转储文件。参见 https://static.openlegaldata.io/dumps/de/2022-10-18/ [4] Open Legal Data API。 https://de.openlegaldata.io/api/ [5]为进行边界检测而按要求纳入词汇表，但之后会划分为 事实部分 或 理由部分

但它不被视为司法推理的实质性部分；因此，我们将其单独存储在模式中。

完成章节分割后，所有收集的行都会使用法律参考文献提取工具进行处理[6]，，该工具能够识别法律引文并按类型（法律、案件）进行分类。随后，每份法院判决都会作为一个独立的 JSON对象进行存储，其中包含规范化的元数据、分章节文本和提取出的参考文献，从而无需额外预处理即可直接使用。网上提供了一个展示该结构的详细 JSON示例条目[7]。

3.2. 验证过程

对判决文本进行自动章节分割可能会引入细微错误，例如边界错置或章节之间部分重叠。为了评估我们的提取过程的可靠性，我们使用带有限总体校正的 Cochran公式估算了人工验证所需的样本量。该方法能够确保分类评估所需样本具有统计代表性，因为每份判决要么被正确分割，要么被错误分割。

无限总体的初始样本量如下：

$$n_0 = \frac{Z^2 \cdot p \cdot (1-p)}{e^2} = \frac{1.96^2 \cdot 0.5 \cdot 0.5}{0.05^2} \approx 384.16, \tag{1}$$

其中，Z 是所选置信水平对应的临界值（Z=1.96为 95%），p是正确提取的估计比例（设为 0.5，以使方差最大，并得出最保守的——即最大的——样本量估计值），e 是误差幅度（5% 时为 0.05）。对于 N=251,038 份判决应用有限总体校正后，得到

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \approx 383.58. \tag{2}$$

其中 N 在此处表示总体规模。实践中，我们采用保守取整后的样本量 $\mathbf{n = 384}$。

因此，我们均匀随机抽取了 384 个案件并进行人工审查。我们逐一检查每个抽样案件，确认三个部分（裁判主文、事实部分和裁判理由）是否识别正确；正确的定义严格限定为各部分之间不存在重叠。人工审查确认，97.40% 的案件分段正确，错误主要源于 HTML 中罕见的格式异常。根据这一结果，我们采用带有限总体校正的正态近似计算了 95% 置信区间；对于包含 251,038 份裁决的完整数据集，区间为（0.9581，0.9899）。这表明，正确分段案件的真实比例介于 95.8% 和 98.9% 之间，说明提取过程具有很高的可靠性。

[6] https://github.com/openlegaldata/legal-reference-extraction [7] 示例条目：https://huggingface.co/datasets/harshildarji/openlegaldata#example-entry

3.3. 章节覆盖率

表1 展示了数据集中三个部分的覆盖情况。裁判主文出现在 87.7% 的案件中，事实部分出现在 65.4% 的案件中，而裁判理由出现在 95.1% 的案件中。事实部分和裁判理由的覆盖率差异反映了不同类型裁决的起草惯例不同：判决通常会明确包含这两个部分；而在裁定中，如果没有划分为理由一和理由二，事实背景有时会并入理由部分。上诉权利告知提供有关可提起上诉的程序说明，共见于 8,335 份裁决（占语料库的 3.32%）。

表 1. 251,038 份裁决的章节覆盖率。

章节	数量
裁判主文	220,273 (87.7%)
事实部分	164,222（65.4%）
裁判理由	238,666（95.1%）

表2。 裁判文书的结构组成。

结构	数量
全部三个部分	144,383（57.5%）
仅裁判主文 + 裁判理由	63,720（25.4%）
仅裁判主文	11,388 (4.5%)

此外，还有**176** 份判决（占总数的**0.07%** ），其中所有部分均缺失。这些对应于原始内容字段为空的案件。表2展示了数据集中各部分的组合情况。大多数判决（57.5%）包含全部三个部分，而仅有25.4%的判决只包含裁判主文和裁判理由，另有少数判决（4.5%）仅包含裁判主文部分。

4. 结论与未来工作

本文介绍了一个包含251,038份德国法院判决的数据集，该数据集源自 Open Legal Data 数据集。我们通过将判决划分为裁判主文、事实部分和裁判理由，为数据集提供了一致的结构，从而解决了原始 HTML 格式不一致、标记不完整的问题。我们还使用 Cochran公式，通过具有统计代表性的随机样本评估了抽取流程。人工核验确认，约97%的案件实现了可靠的分离。该数据集采用 JSONL 格式，包含元数据和抽取出的参考文献，可直接用于进一步研究，无需额外预处理。

我们目前还在使用该数据集为德国法律文本构建检索增强生成系统。分段结构使我们的检索流程能够分别索引案件摘要、法条引用和裁判理由段落，并将其与用户查询进行对齐。未来的工作将重点通过加入排序和重排序策略来提升检索质量。我们还计划在各项子任务中评估检索增强生成系统，以进一步展示分段的价值。这些子任务包括法条检索，其中必须将裁判理由段落与所引用的法条进行对齐；推理覆盖度，其中需要区分事实背景与论证；以及可解释性，在这一任务中，用户能够从法院判决中仅查看最相关的部分。最后，随着更多法院判决的发布，我们将继续扩充该语料库，并进一步优化对法律文书撰写差异的处理。

参考文献

[1] Ostendorff M、Blume T、Ostendorff S。面向法律信息开放平台。载于：2020年ACM/IEEE数字图书馆联合会议论文集；2020年。第385-8页。[2] Wrzalik M、Krechel D。GerDaLIR：用于法律信息检索的德国数据集。载于：2021年自然法律语言处理研讨会论文集；2021年。第123-8页。[3] Darji H、Mitrovi´c J、Granitzer M。探索德国法律文本与所引用法律之间的语义相似性。载于：知识发现、知识工程与知识管理国际联合会议。Springer；2021年。第37-50页。[4] Darji H、Mitrovi´c J、Granitzer M。法律数据标注中的挑战与考量：全面综述。arXiv预印本arXiv:240717503。2024年。[5] Ma Y、Wu Y、Ai Q、Liu Y、Shao Y、Zhang M等。将结构信息融入法律案例检索。ACM信息系统学报。2023年；42(2)：1-28。[6] Santosh T、Isaia A、Hong S、Grabmair M。Hiculr：用于法律文档修辞角色标注的分层课程学习。arXiv预印本arXiv:240918647。2024年。[7] Bambroo P、Adhikary S、Bhattacharya P、Chakraborty A、Ghosh S、Ghosh K。MARRO：用于法律文档修辞角色标注的多头注意力。人工智能与法律。2025年：1-30。[8] Glaser I、Moser S、Matthes F。改进法律信息检索：德国法院裁判的元数据提取与分段。载于：KDIR；2021年。第282-91页。[9] Urchs S、Mitrovi´c J、Granitzer M。探索德国法律判决中德国法律写作风格各部分的分类。载于：2020年第十届高级计算机信息技术国际会议（ACIT）。IEEE；2020年。第451-4页。[10] Urchs S、Mitrovic J、Granitzer M。德国法律裁判语料库的设计与实现。载于：ICAART（第2卷）；2021年。第515-21页。[11] Cochran WG。抽样技术。John Wiley & Sons；1977年。[12] Glaser I、Moser S、Matthes F。德国法院裁判的摘要生成。载于：2021年自然法律语言处理研讨会论文集；2021年。第180-9页。[13] Darji H、Mitrovi´c J、Granitzer M。德国法律引用标注数据集。载于：第十九届人工智能与法律国际会议论文集；2023年。第392-6页。[14] Heckelmann M。智能合约。《新法学周刊》（NJW）。2018年：504-9页。