基于多粒度数值嵌入的法律刑期预测方法

李泽南[1] 邱锦宏[1]
LI Zenan QIU Jinhong

摘 要 现有模型在法律刑期预测任务中表现优异，但在处理涉案金额、犯罪次数等关键离散数值时，易出现数值语义鸿沟问题。传统基于 BERT 的纯文本处理方法难以捕捉法律条款中严格的阈值阶梯效应，导致金额敏感型案件的判决预测精度受限。为此，文章提出一种基于多粒度数值嵌入的法律判决预测网络。该模型构建可微数值表征学习模块，从全局量级（Magnitude）、局部法律分档（Legal Tier）、微观连续偏移（Continuous Offset）三个粒度对数值进行编码，并将多尺度数值特征与文本语义向量深度融合，使模型同时感知数值的绝对大小与相对法律定位。在 LAIC2021 数据集上的实验结果表明，该模型在刑期预测任务上达到 48.31% 的 Macro-F_1 值，较最优基线模型提升 1.02%。

关键词 自然语言处理；法律智能；数值嵌入；注意力机制

doi：10.3969/j.issn.1672-9528.2026.03.032

0 引言

随着人工智能技术在司法领域的深入应用，法律判决预测（legal judgment prediction, LJP）成为提升司法效率、促进裁判一致性的关键任务 [1-2]。其核心目标是通过解析案件事实文本，自动预测罪名、法律条款及刑期。其研究可追溯至 20 世纪中叶，早期学者尝试通过数学统计方法，如 Kort[3] 对美国最高法院“律师权”案件的预测；Nagel[4] 采用相关性分析的判决逻辑建模探索司法规律。但因法律文本的非结构化特性与复杂逻辑限制，成果多停留于理论阶段。直至 21 世纪以来，随着深度学习技术的突破与大规模法律数据集（如 CAIL2018[5]、ECHR）的构建，LJP 研究才进入快速发展期。

Luo 等 [6] 构建了基于双向门控循环单元（Bi-GRU）的联合建模框架，同步关联罪名预测与法律条文匹配任务。针对低频罪名识别难题；Hu 等 [7] 通过引入司法属性特征空

1. 江西理工大学软件工程学院 江西南昌 330013

案。显著提高了虚拟电厂的调度精度和运行效率。通过实际案例分析，验证了所提方法在降低运营成本和提高能源利用率的显著效果。

参考文献：

[1] 林顺富，郭宇霆，周波，等．计及电价不确定性的多虚拟电厂与共享储能协同优化调度方法 [J]. 智慧电力，2025, 53(7): 20-27.
[2] 易文飞，叶志刚．考虑储能和条件风险价值的虚拟电厂优化调度 [J]. 电力需求侧管理，2025,27(4):86-91.
[3] 高建强，蔡杜钟，刘春涛，等．基于多目标沙猫群算法的含碳捕集虚拟电厂优化调度 [J]. 动力工程学报，2025, 45(7): 1126-1133.
[4] 郑思纬，黄国平．碳交易机制下柔性负荷虚拟电厂日前优化调度研究 [J]. 徐州工程学院学报（自然科学版），2025, 40(2): 65-76.
[5] 丁君，秦浩庭，苏鹏，等．基于改进 Harris Hawk 优化算法的虚拟电厂优化调度研究 [J]. 可再生能源，2025,43(6):829-838.
[6] 邢耀敏，高春雨，廖丛林．考虑碳捕集及碳交易的虚拟电厂多目标优化调度 [J]. 分布式能源，2025,10(3):42-52.
[7] 李春锋．面向矿区电网新能源消纳的智能资源调度方法 [J]. 计算机仿真，2025,42(6):148-153.
[8] 刘超超，常玉娇．基于改进遗传算法的虚拟电厂分布式电源调度优化方法 [J]. 无线互联科技，2025,22(9):78-81.

【作者简介】

史渊源（1982—），男，宁夏固原人，本科，高级工程师，研究方向：大数据分析应用，email:yuany5819@163.com。

王亮（1981—），男，吉林松原人，本科，高级工程师，研究方向：大数据分析应用。

万鹏（1988—），男，宁夏石嘴山人，本科，高级工程师，研究方向：数据管理与运营。

（收稿日期：2025-08-14 修回日期：2026-02-27）

间，有效缓解了数据长尾分布的影响。值得一提的是，基于 Transformer 架构的预训练语言模型 LegalBERT[8] 在法律领域展现出独特优势，其通过大规模法律语料预训练，能够深度捕捉法律术语的语义特征，为复杂案情理解提供了新范式。

然而，法律文本中数值的强语义关联性以及刑期判决的阶梯式逻辑，使得传统 BERT 模型面临严峻挑战 。在司法实践中，数值往往不是独立的语义单元，而是通过特定的“阈值”触发不同的法律后果。例如，在《中华人民共和国刑法》关于盗窃罪的规定中，“3 万元”不仅是一个数字[8]，更是区分“数额较大”与“数额巨大”的刚性边界，直接决定了是判处“三年以下”还是“三年以上”有期徒刑 。现有的预训练模型通常采用子词（Subword）切分或简单的分箱策略处理数字 ，这种处理方式将连续的数值离散化为缺乏序数含义的符号，导致模型难以感知“29999 元”与“30000 元”在法律量刑上的质变差异[5, 9]。因此，如何有效地将离散的法律数值映射为具备法律阶梯属性的连续向量空间，成为当前法律判决预测研究中亟待解决的关键问题。

为此，本文提出一种基于多粒度数值嵌入的法律判决预测模型 ，其核心贡献如下：

提出多粒度数值表征学习机制（Multi-grained Numerical Embedding）： 针对法律数值的语义鸿沟问题，构建了一个可微的数值编码模块。该模块创新性地设计了量级嵌入（Magnitude）、法律分档嵌入（Legal Tier）与偏移量编码（Offset Encoding）的三层并行表征，将数值的全局尺度、离散法律属性以及区间内的连续相对位置融合为高维结构化特征。这不仅增强了模型对法律数值的语义感知能力，还有效解决了传统模型在临界数值预测上的不确定性，显著提升了刑期预测的准确性与逻辑一致性。

1 实验设计

1.1 数据集与预处理

本实验采用 LAIC2021 公开法律数据集。包含 98 962 份刑事裁判文书，涵盖事实描述、法院观点、指控罪名、法律条款及刑期等关键信息。为提升数据质量与模型训练效率，对原始数据进行如下预处理：

（1）文本过滤：剔除事实描述中有效单词数少于 10 的样本，保留语义完整的案情描述。

（2）标签筛选：仅保留出现频次大于 50 次的指控罪名和法律条款。

（3）刑期区间化：将刑期划分为非重叠区间，转化为分类任务以降低回归噪声。

（4）数据划分：按 8:1:1 比例随机切分数据集为训练集、验证集和测试集。

1.2 整体框架模型

本文提出基于多粒度数值嵌入的法律刑期预测方法，其模型架构图如图 1 所示。

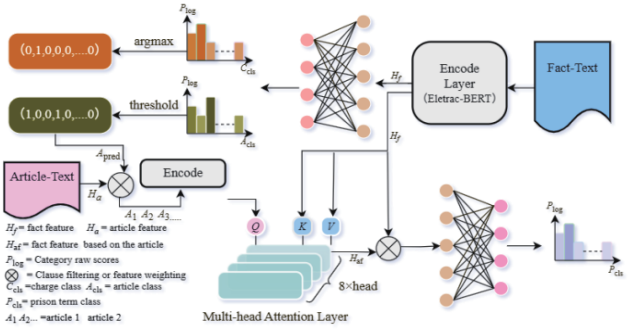

图 1 模型整体架构

1.3 文本语义编码层

在本模型的整体架构中（如图 1 所示），首先需要对案件事实描述文本（Fact Description）进行向量化表示。本文采用 Legal-BERT 作为基础编码器。给定输入案情文本序列 $X = \{x_1, x_2, \cdots, x_n\}$，其中 n 为序列长度通过 Legal-BERT 的多层 Transformer 结构，获取上下文感知的隐藏层表示：

$$\boldsymbol{H} = \text{LegalBERT}(X) \tag{1}$$

式中：$\boldsymbol{H} \in \mathbf{R}^{n \times d_h}$，$d_h$ 为隐藏维度，本文将 [CLS] 标记对应的 h_{cls} 作为整个案件的全局语义表征 H_f，该表征捕捉了案件的非数值语义信息，如作案手段、主观恶性。

1.4 多粒度数值嵌入

首先基于罪名构建分档规则库 D，后提取裁决文书中的金额 v，对数值 v 进行层次化嵌入。层次化嵌入分为三层，具体实现如下：

（1）量级嵌入（Magnitude Embedding）：提取数值 v 的量级表示，量级嵌入表示了数值的大小。

（2）法律分档（Legal Tier Embedding）：法律分档根据分档规则库 D 对数值进行分档，使模型能够抓取数值 v 在当前案件中对应的法律意义，用公式表示为：

$$\boldsymbol{E}_{\text{tier}}(v, c) = \text{Embed}^{c}_{\text{tier}}\left(\sum_{k=1}^{K} I\left(\geq t_k^c\right)\right) \tag{2}$$

式中：v 表示待编码的数值；c 表示对应的罪名；t_k^c表示罪名 c 对应的第 k 个法律分档阈值。。

计算步骤：根据罪名 c 的阈值序列$t_1^c t_2^c \cdots t_k^c$判断所属分档，例如：盗窃罪阈值 [3 万 , 10 万 , 30 万]，v = 48 万 →tier_{idx}=3，最后通过嵌入层映射为向量 $\boldsymbol{E}_{\text{tier}} \in \mathbf{R}^d$。

（3）偏移量编码（Offset Encoding）：偏移量用于捕捉同一法律分档区间内数值的具体相对位置。在法律判决预测中的作用类似于“微调器”，能够在保持法律分档框架的同时，为模型提供更细粒度的数值敏感性。设法律分档对应的区间为$[t_{k-1}^c, t_k^c)$，对于数值 $v \in [t_{k-1}^c, t_k^c)$偏移量用公式表示为：

$$\boldsymbol{E}_{\text{offset}}(v, c) = f_\theta\left(\frac{v - t_{k-1}^c}{t_k^c - t_{k-1}^c}\right) \tag{3}$$

式中：v 表示数值；c 表示罪名；f_θ 表示可学习的非线性编码

网络。

计算步骤：归一化偏移量（封闭区间）（例如盗窃罪 48 万，区间t_2^c=10 万至t_3^c=30 万，则偏移量$\text{offset}=\frac{48-10}{30\quad 10}=1.9$。最后经过神经网络映射为向量 $\boldsymbol{E}_{\text{offset}}\in\mathbf{R}^d$。

（4）特征拼接 $\boldsymbol{E}(v,c)\in\mathbf{R}^d$，用公式表示为：

$$\boldsymbol{E}(v,c)=\boldsymbol{E}_{\text{mag}}(v)\oplus\boldsymbol{E}_{\text{tier}}(v,c)\oplus\boldsymbol{E}_{\text{offset}}(v,c) \quad (4)$$

式中：⊕ 表示特征拼接操作。

（5）获得文本特征 $\boldsymbol{H}_f$ 与多粒度数值特征 $\boldsymbol{E}(v,c)$ 后，采用拼接（Concatenation）策略进行融合。为防止数值特征在维度上被文本特征淹没，在拼接前引入了特征加权机制。最终的融合特征 $\boldsymbol{H}_{\text{final}}$ 输入到全连接层进行刑期区间的分类预测：

$$P=\text{Softmax}(W_p\boldsymbol{H}_{\text{final}}+b_{\text{p}}) \quad (5)$$

模型训练采用交叉熵损失函数（cross-entropy loss），通过最小化预测概率分布与真实刑期标签之间的差异来优化所有参数。

2 实验结果与分析

2.1 实验配置

实验环境采用 Ubuntu 20.04.6 LTS 操作系统，硬件配置包括 Intel(R) Core(TM)i9-9820X CPU（主频 3.30 GHz）、NVIDIA Quadro RTX 5000 Max-Q（32 GB 显存）。软件环境基于 Python 3.8.8 与 PyTorch 1.12.1 框架。文本编码采用 Legal-BERT（bert-base-legal-cn）预训练模型，最大序列长度设置为 512，批量大小（Batch Size）为 32。数值嵌入中，量级与法律分档嵌入维度均为 64，偏移量编码通过两层全连接网络（输入维度 1→256→64，ReLU 激活）实现。学习率为 2e-5，权重衰减系数 0.01，训练过程采用早停策略 patience=3 与最大 epochs=200 限制。

2.2 评价指标

为全面评估模型在法律判决预测（LJP）任务中的性能，本实验针对法条推荐（Law Prediction）、罪名预测（Charge Prediction）和刑期预测（Term Prediction）三个子任务，采用了准确率（Accuracy）和宏平均 F_1 值（Macro-F_1）作为核心评价指标。

准确率（ACC）：衡量模型预测正确的样本占总样本的比例，反映模型的整体判决能力。

Macro-F_1：考虑到法律数据集中存在显著的长尾分布问题（即某些罪名或刑期区间样本极少），Macro-F_1 对每个类别单独计算 F_1 值后取算术平均，能够更公正地评估模型在稀疏类别上的表现。其计算公式为：

$$\text{Macro\_}F_1=\frac{1}{C}\sum_{i=1}^{C}\frac{2\times P_i\times R_i}{P_i+R_i} \quad (6)$$

式中：P_i 和 R_i 分别表示第 i 类别的精确率（Precision）和召回率（Recall）；C 表示类别总数。

2.3 结果与分析

2.3.1 消融实验

在上述基本环境配置下，为验证该模型核心组件的有效性，所有变体均使用相同的预训练模型（BERT-Legal），仅改变数值特征的处理方法，因此设计以下消融实验变体：Text-Only：移除层次化数值嵌入（量级、法律分档、偏移量编码），模型仅输入案情文本，不显示提取数值特征，验证“数值语义鸿沟的存在”；Text-Num：将提取的数值 v 进行简单归一化后拼接在文本特征后验证数值重要性；w/o Offset：保留量级嵌入和法律分档嵌入，但移除偏移量编码（Offset）。相当于把数值看作纯粹的离散类别；完整模型，包含量级 + 法律分档 + 偏移量，结果如表 1 所示。

表 1 不同数值特征处理模块的消融实验结果

Methods	Law-Article		Charge		Term	
	ACC	Macro-F_1	ACC	Macro-F_1	ACC	Macro-F_1
Text-Only	74.16	71.11	95.11	93.95	40.14	41.21
Text-Num	75.15	74.88	96.55	94.25	41.52	41.77
w/o Offset	**77.25**	75.26	**97.88**	94.41	46.11	45.29
完整模型	77.24	**75.31**	97.84	**94.44**	**49.39**	**48.31**

消融实验结果清晰地揭示了多粒度数值嵌入机制在跨越“数值语义鸿沟”中的递进式增益效应。实验数据显示，单纯引入原始数值（Text-Num）相较于纯文本基线（Text-Only）仅带来微弱的性能提升（Macro-F_1 增幅不足 0.6%），实证了深度神经网络难以仅凭连续数值自动习得复杂的非线性量刑规则。然而，随着法律分档嵌入（Legal Tier）的引入，模型性能实现了质的飞跃（Macro-F_1 激增至 45.29%），表明将外部专家知识注入模型，把连续数值离散化为具备明确法律属性的“档位向量”，是捕捉“数额巨大”等刚性阈值及其阶梯式量刑逻辑的关键。在此基础上，完整模型进一步叠加偏移量编码（Offset Encoding），将 Macro-F_1 提升至最优的 48.31%；偏移量编码有效充当了“语义微调器”，通过精确刻画数值在分档区间内的相对位置，填补了离散化带来的信息损失。综上所述，这种“粗粒度分档定性 + 细粒度偏移定量”的组合策略，完美契合了司法判决中“定性分析与定量计算相结合”的双重逻辑，显著增强了模型对临界数值及区间内细微变化的敏感度。

2.3.2 多模型对比

在深度学习领域，法律判决预测任务需兼顾文本语义理解与法律逻辑建模。当前主流的几个模型中：CNN 通过卷积层提取文本局部特征，但难以捕捉长距离法律条文关联；BERT 基于 Transformer 编码器建模全局上下文；Electra 通过替换词检测预训练优化词向量质量；R-former 创新性融合图结构与自注意力机制，挖掘罪名 - 条款 - 刑期间拓扑关系，

ML-LJP 通过建模法律条款间关联关系并融合对比学习，实现罪名与法律条款的联合预测。

为验证本文模型的有效性，本文在相同实验环境下与多类主流架构进行对比，其结果如表 2 所示。

表 2 多模型对比结果表

Methods	Law-article		Charge		Term	
	ACC	Macro-F_1	ACC	Macro-F_1	ACC	Macro-F_1
CNN	71.47	60.41	96.41	91.03	34.14	32.94
BERT	71.91	66.49	96.72	90.98	41.83	41.65
Electra	71.97	67.09	96.29	92.91	42.56	41.84
R-former	—	—	97.49	94.04	43.62	43.17
ML-LJP	75.71	73.21	97.54	94.44	47.63	47.52
本文模型	77.24	**75.31**	97.84	**94.64**	**49.39**	**48.31**

为验证本文在复杂法律语境下的判决预测能力，在相同实验设置下与 CNN、BERT、Electra、R-former 及 ML-LJP[10] 主流基线模型进行了对比。实验结果显示，本文模型在所有核心指标上均取得最优性能。特别是在对数值敏感度要求极高的刑期预测（Term）任务中，BERT 和 Electra 等通用预训练模型受限于其对离散数值的表征缺陷，Macro-F_1 值分别停留在 41.65% 和 41.84%。相比之下，本模型的刑期预测 Macro-F_1 提升至 48.31%，超越 BERT 模型约 6.66%。这一压倒性优势有力地证明了，单纯依赖文本语义无法精准捕捉量刑的阶梯式逻辑，而本文提出的多粒度数值嵌入模块（量级 - 分档 - 偏移）成功地将离散数值转化为具备法律语义的稠密表征，显著增强了模型在金额敏感型案件中的判决可靠性。

3 结论

本文针对现有预训练模型在法律数值理解上的语义鸿沟，提出了基于多粒度数值嵌入的判决预测模型。该模型创新性地构建了包含“全局量级 - 法律分档 - 区间偏移”的三层表征学习机制，将离散数值转化为兼具法律逻辑与连续语义的稠密向量，有效弥补了传统方法在处理“阶梯式”量刑逻辑时的缺陷。实验表明，该模型在刑期预测任务中 Macro-F_1 达 48.31%，显著优于 BERT、Electra 等基线模型，验证了将领域先验知识融入深度表征的有效性。未来工作将进一步探索复杂案情下的多模态数值推理机制（如累计金额计算），以及模型在不同司法体系与罪名中的泛化适应能力，以推动司法智能化的精细化发展。

未来工作将集中在以下两个方面：一是探索复杂案情下的多模态数值推理机制。目前的模型主要处理单一的关键金额，但在实际案件中，往往涉及多次犯罪金额的累加、不同种类财物的折算等复杂情况，未来可以考虑引入图神经网络（GNN）来处理数值间的运算关系。二是提升模型的泛化适应能力。目前的实验基于 LAIC2021 数据集，未来将验证该模型在民事赔偿金额预测、行政处罚金额预测等其他司法领域的有效性，以推动法律智能技术的全面落地。

参考文献：

[1] 黄欢欢 . 基于深度学习的法律判决预测研究 [D]. 成都 : 四川大学 ,2022.

[2] 张晗 , 郑伟昊 , 窦志成 , 等 . 融合法律文本结构信息的刑事案件判决预测 [J]. 计算机工程与应用 ,2023,59(3):253-263.

[3]KORT F. Predicting supreme court decisions mathematically: a quantitative analysis of the "right to counsel" cases [J]. American political science review, 1957, 51(1): 1 - 12.

[4]NAGEL S S. Applying correlation analysis to case prediction [J].Texas law review, 1963, 42: 1006.

[5]XIAO C J, ZHONG H X, GUO Z P, et al. CAIL2018: a large-scale legal dataset for judgment prediction [EB/OL]. (2018-07-04)[2026-01-22].https://doi.org/10.48550/arXiv.1807.02478.

[6]LUO B F, FENG Y S, XU J B, et al. Learning to predict charges for criminal cases with legal basis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2017: 2727-2736.

[7]HU Z K, LI X, TU C C, et al. Few-shot charge prediction with discriminative legal attributes[C]//Proceedings of the 27th International Conference on Computational Linguistics. Brussels: ACL, 2018:487–498.

[8]CHALKIDIS I, FERGADIOTIS M, MALAKASIOTIS P, et al. LEGAL-BERT: the muppets straight out of law school[C]// Findings of the Association for Computational Linguistics: EMNLP 2020.Brussels: ACL, 2020: 2898-2904.

[9] KATZ D M, BOMMARITO II M J, BLACKMAN J. A general approach for predicting the behavior of the Supreme court of the United States [J]. Plos one, 2017, 12(4): e0174698.

[10] LIU Y F, WU Y Q, ZHANG Y T, et al. ML-LJP: multi-law aware legal judgment prediction[C]//SIGIR'23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. NewYork: ACM, 2023: 1023-1034.

【作者简介】

李泽南（2000—），男，河南焦作人，硕士研究生，研究方向：机器学习与自然语言处理。

（收稿日期：2025-12-21 修回日期：2026-03-11）