人工智能辅助法官决策研究
——基于量刑偏差识别视角

周 静 杨玲燕 刘 喆 王 芳*

摘 要：为实现“让人民群众在每一个司法案件中都感受到公平正义”，国家持续推进量刑规范化改革，规范刑罚裁量权，促进量刑公正。如何把握刑罚裁量权的合理范围，努力实现同案同判，是量刑规范化改革的核心。本文以发现量刑畸轻畸重等量刑偏差判决为研究目标，提出一种量刑偏差的识别方法：基于中国裁判文书网460 486条数据，采用长短期记忆（Long Short-Term Memory，简称LSTM）模型对案件的量刑进行预测，并提出了异质性系数度量，用于量刑偏差案件的识别。研究发现，制作、复制、出版、贩卖、传播淫秽物品牟利罪，窝藏、包庇罪，以及挪用资金罪是最容易产生量刑偏差的三种罪。本文以挪用资金罪为例，采用多元线性回归模型分析了量刑偏差的原因，分析发现坦白对挪用资金罪量刑有减轻作用，挪用金额大小与量刑长短成正比，然而，“挪用资金用于营业活动”这一变量系数存在异常，可能会导致量刑偏差的出现。最后，本文选取了两个有代表性的量刑偏差案件，通过法律专家的案例分析来判定偏差案件识别是否准确，并提供法律依据。该方法以既有判决的量刑共识为基础，以期为国家量刑规范化改革提供辅助参考。

关键词：司法人工智能；刑期预测；量刑偏差识别；异质性系数；LSTM模型
中图分类号：C81；C39
JEL分类号：K14；C45；B23

一、引 言

量刑，是刑罚公正的终极体现。为实现“让人民群众在每一个司法案件中都感受到公平正义”，国家持续推进量刑规范化改革。为此，最高人民法院自2008年起颁布《人民法院量刑指导意见（试行）》（以下简称《指导意见》），随后进行了6次修订。《指导意见》对量刑的基本方法和步骤、常见量刑情节的适用范围、

* 周静，中国人民大学统计学院，E-mail: zhoujing_89@126.com；杨玲燕，山东大学数据科学学院，E-mail: yanglingyan@mail.sdu.edu.cn；刘喆，中国人民大学统计学院，E-mail: lzhe0029@126.com；王芳（通信作者），山东大学数据科学学院，E-mail: wangfang226@sdu.edu.cn。作者感谢国家自然科学基金项目（T2293773）、中国人民大学科学研究基金面上项目（21XNA027）对本文研究的支持。作者感谢匿名审稿人和编辑部的宝贵意见，当然文责自负。

常见犯罪的量刑提供了全面的适用指南。但现实世界千差万别，《指导意见》不可能穷尽所有情况，加之区域经济社会发展水平差异、法官个体差异、被告人所具有的个性化特征等各种原因，同案不同判的情况仍有发生。这可能导致较低的服判息诉率，以 2020 年为例，中国最高人民法院共审理了 112 万件一审案件，其中分别有约 11% 和 2% 的案件经历了二审和发回重审。这表明，仍有大量诉讼并未终止（无上诉或反上诉），其中许多案件可能都是争议案件，不同法官对诸如刑期、罚金等具体判罚可能持有不同的意见（孙海波，2017）。同时，也可能在一定程度上影响司法公正，不利于维护法律的权威和公信力。有学者指出，应该将法官个体的刑罚裁量与法官量刑集体经验进行对比，对靠近集体经验量刑的法官的自由裁量权采取肯定和尊重的态度，而对量刑显著偏离集体经验的法官的量刑决策进行识别并纠正其偏差（吴雨豪，2021）。2016 年起，国家大力推进智慧法院建设，希望通过大数据、人工智能技术的运用发现审判共识，进而提高案件受理及审判的准确度和公平性，为推动司法公平正义贡献力量（白建军，2017；左卫民，2021）。

本文站在审判监督的角度，提出了一种能够自动发现量刑偏差的技术方法。自 2021 年起，中国裁判文书网[1] 已向公众开放了超过 1 亿份法律判决文书，为基于司法判决的分析研究提供了海量的数据基础，也为开发先进的机器学习算法来实现量刑偏差案件的自动识别奠定了数据基础。本文以 2018 年全年刑事裁判文书数据为样本，基于 62 种罪名、共 460 486 条法律文书数据进行分析，提出了一种能够准确发现司法审判中量刑畸轻畸重的异常情况的方法。具体包括以下三个方面：首先，以刑期为因变量，以法律文书中“经审理查明”和“法院认为”提取的文本作为案件事实描述，构建长短期记忆模型用于刑期的预测；其次，基于模型的预测结果，计算预测刑期与真实刑期的差值，并构建异质性系数用于识别量刑偏差的罪名，计算发现，制作、复制、出版、贩卖、传播淫秽物品牟利罪，窝藏、包庇罪，以及挪用资金罪是异质性系数得分最高的三种罪，说明这三种罪最有可能产生量刑偏差案件；最后，为了探究影响量刑的具体因素，本文以挪用资金罪为例，构建了影响刑期长短的回归模型，分析了挪用资金罪量刑偏差的具体原因。

二、文献综述

随着大数据和人工智能的不断发展，国内外学术界都在积极推动相关技术与司法实践的融合，以促进司法智能化的发展（Aletras et al.，2016；Zhong et al.，2014；

1 资料来源：https://wenshu.court.gov.cn/。

张玉洁，2021）。例如，以深度学习为代表的自然语言处理（Natural Language Processing，简称 NLP）技术在司法领域就得到了极大的发展。研究人员设计了以循环神经网络（Recurrent Neural Network，简称 RNN）为基础的不同算法框架以适应不同的学习任务，不仅可以根据给定的案情事实描述对判决结果进行预测（Luo et al.，2017；Hu et al.，2018），还可以专注于预测刑期等程度更为精细的判决预测任务（舒洪水，2020；李大鹏等，2022）。在 RNN 的诸多变体中，LSTM 是使用最为广泛的模型结构之一，它可以解决很多 NLP 任务，例如，Johnson and Zhang（2003）探索了使用 LSTM 模型的文本区域嵌入方法。Zhou et al.（2016）集成了双向 LSTM（Bidirectional-LSTM，简称 BiLSTM）模型和二维最大池化来提取文本特征，Wan et al.（2016）使用双向 LSTM 模型捕获每个句子的上下文信息并进行表示，以探索语义匹配关系。在司法领域，Chen et al.（2019）指出，被告人在司法实践中可能被同时指控多项罪名，因此可以使用深度门控网络通过提取事实描述和特定罪名之间的复杂关系，建立基于罪名的刑期预测模型，该方法能有效地提高模型的预测精度。马建刚和马应龙通过构建图长短期记忆（Graph Long Short-Term Memory，简称 Graph LSTM）模型，实现了语义驱动下的司法文书分类。而 Li et al.（2019b）另辟蹊径，将有期徒刑的样本数据按照刑期长度划分为五类，以犯罪事实信息为自变量，使用 FastText 和 TextCNN 建立文本分类模型。王治政等（2021）则认为相较于深度学习模型，司法知识图谱可以展现案件核心要素的联结情况，更有利于对量刑预测结果进行解释。还有学者建议通过一个完整的框架执行所有司法预测任务，例如，Li et al.（2019a）提出了一种多通道注意力机制的神经网络，按照人类的思维逻辑同时实现对罪名、法条和刑期的预测，提高了司法裁决预测任务的可信度和解释性。

围绕刑罚与量刑影响因素的因果关系研究，孙道萃（2020）提出应当建立精准的人工智能辅助预测量刑系统，通过挖掘数据中的量刑规律，从而进一步提高传统量刑实践的公平性和正义性。白建军（2016）选取了 147 229 个交通肇事罪的案件，以刑期长度为因变量，以手动提取的法定量刑情节为自变量，建立多元回归模型，结果显示：通过限缩量刑情节的裁量幅度，可以将此类案件的量刑确定性由原来的 30.5% 提高到 51.1%；在此基础上控制样本的离散程度，可以将量刑确定性由 51.1% 进一步提高到 73.4%。还有部分学者则以危险驾驶罪为研究对象，从不同角度构建了量刑模型并对量刑特征进行了研究（章桦和李晓霞，2014；文姬，2016；樊祜玺和万力，2019；江溯，2021；文姬和黄雪，2020；白建军，2020；Lee，2006）。除研究如何建立模型来预测刑期长度以外，高通等（2020）更关注某一种特定的酌定情节如何影响量刑结果，研究结果表明，赔偿这一酌定情节对故意伤害罪的量刑结果有显著影响，且随着案件严重程度的加深，影响呈下降趋势。同年，章桦（2020）也对贪污罪的数额与情节做出了实证研究，指出明确数额与严重情节

之间、严重情节与从宽情节在定罪量刑中的影响程度，应是未来理论研究、立法修正和司法解释的着力方向。除此之外，还有学者以某种罪名为例研究了更具体的量刑差异影响因素。例如，胡昌明（2018）以盗窃罪为例探究了被告人身份差异对量刑的影响；王剑波（2018）研究了行政级别、身份性质与受贿罪的量刑差异。

量刑合理与否关系到司法是否公正运行，从现有文献看，利用算法进行量刑纠偏的文献甚少，但已经有学者开始关注引起量刑偏差的因素并予以控制。如吴雨豪（2021）基于北京地区五类案件近 5 万份刑事判决书，对比个案的刑罚裁量与全样本量刑集体经验，识别出了量刑显著偏离集体经验的判决。赵学军（2019）通过对 4 354份裁判文书进行统计分析发现，不同地域、不同时期和不同个案间的量刑偏差现象依然存在。谭红叶等（2020）提出了偏差区间划分方法，保证刑期区间划分的准确性，避免因刑期区间划分错误带来的预测偏差。

从以上对现有文献的回顾不难看出，目前已有诸多关于人工智能量刑的研究，但大多集中在预测及量刑影响因素探究的层面，鲜有研究量刑偏差案件的识别，已有研究主要从理论角度论证疑难案件的定义、成因、重要性、处理方式等内容，却未提出识别量刑偏差案件的具体量化方法。为填补该理论空白，本文拟通过前沿的深度学习模型建立刑期预测模型，并开发出一个量化指标，用于量刑偏差案件的识别，为司法量刑实践提供参考，使其具有一定的实用价值。

三、数据介绍与描述性分析

本文选取刑事判决文书作为研究对象。其优势在于，刑事案件审判过程的规范性和严谨性相对较高，犯罪构成要件与量刑情节的内在逻辑与司法人工智能模型的决策机制更加符合。

（一）数据介绍

本文选取了 2018 年 1 月 1 日至 2018 年 12 月 31 日的刑事裁判文书，数据总量为 1 361 354 份，包含 268 种不同的罪名。根据研究目标，本文仅选择刑期类型为有期徒刑的判决作为样本。由于不同罪名包含的样本数量存在较大差异，部分罪名仅包含几个样本，为方便后续建模，本文进一步选取样本量大于 500 的罪名进行研究。处理后的样本最终涉及 62 种罪名，共 460 486 条数据。

表 1 展示了样本量最多和最少的十种罪名，从中可以看到，盗窃罪、故意伤害罪等常见犯罪对应的样本量较大。这里排除了刑期类型不是有期徒刑的罪名，例如大部分的危险驾驶罪、代替考试罪、环境监管失职罪和逃避商检罪四种罪名。

表 1　样本量最多和最少的十种罪名

罪名	样本量	罪名	样本量
盗窃罪	107 211	组织、利用会道门、邪教组织、利用迷信破坏法律实施罪	626
故意伤害罪	60 387	强制猥亵、侮辱妇女罪	623
交通肇事罪	53 064	集资诈骗罪	607
走私、贩卖、运输、制造毒品罪	38 728	猥亵儿童罪	607
诈骗罪	29 072	伪造公司、企业、事业单位、人民团体印章罪	589
寻衅滋事罪	23 757	窝藏、包庇罪	540
容留他人吸毒罪	13 671	妨害信用卡管理罪	530
开设赌场罪	11 765	非法采伐、毁坏国家重点保护植物罪	523
抢劫罪	7 660	危险驾驶罪	518
妨害公务罪	6 916	非法行医罪	503

注：本表展示了样本区间为 2018 年 1 月 1 日至 2018 年 12 月 31 日的刑事判决文书中，样本量最多的十种罪名与样本量最少的十种罪名以及对应的样本数。

（二）描述性分析

根据研究目的，本文选取刑期长度作为因变量（单位：月），其分布见图 1，可以看到原始刑期（左图）呈右偏分布，大多数犯罪的刑期集中在 30 个月以内。此外，图 2 还展示了案件数最多的五种罪名（盗窃罪，故意伤害罪，交通肇事罪，走私、贩卖、运输、制造毒品罪，诈骗罪）的刑期分布箱线图，从中可以看到，不同的罪，其刑期分布差异较大，其中走私、贩卖、运输、制造毒品罪和诈骗罪的刑期方差较大，说明这两种罪个案刑罚差异较大。

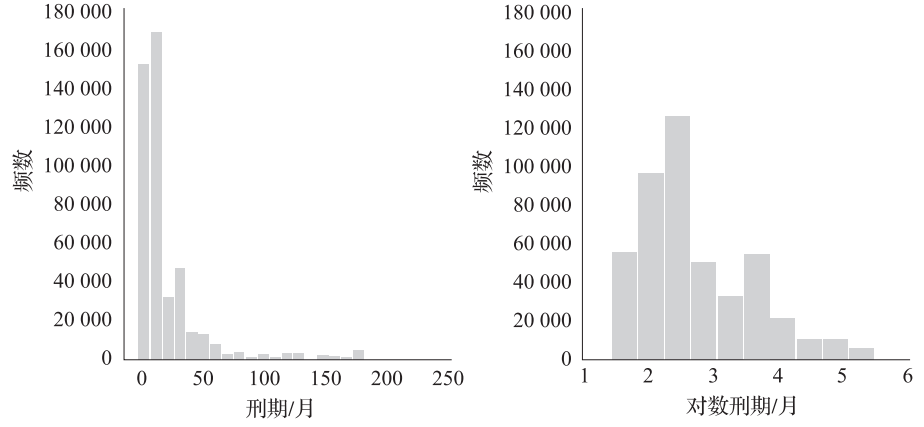

图 1　刑期分布直方图

注：本图展示了样本数据的刑期分布状况。其中，左图为原始刑期分布直方图，右图为经过对数变换后的刑期分布直方图。

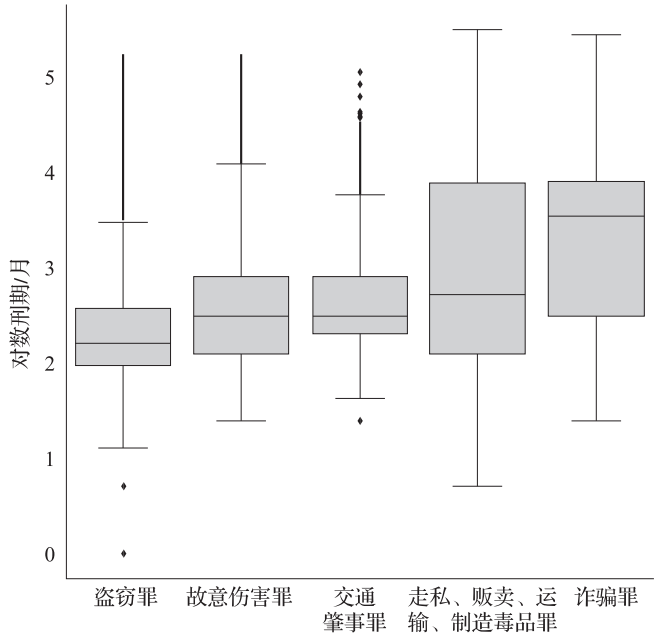

图 2 案件数最多的五种罪名的分组箱线图

注：本图展示了频数最多的五种罪名的刑期分布状况。其中，横坐标为罪名，纵坐标为刑期的对数值。

由于案件的事实信息是研究定罪量刑的基础，因此，只有保证案件内容真实且客观，才能确定被告人的犯罪性质并据此裁决刑期的长度。由于裁判文书的“案件内容”通常篇幅较长，且包含对案件详细情节和判决结果等各种信息的描述，因此，本文仅选取对刑期影响较大的客观事实部分进行研究。具体而言，刑事判决文书是半结构化文书，完整的文书由案号、被告人基本信息、检察院指控、经审理查明、法院认为和法院判决等几部分组成。其中，被告人基本信息、检察院指控部分仅代表各方的观点，未经法庭质证，尚未被认定为实际影响判决的法律事实。因此，本文选择案件内容中“经审理查明”与“本院认为”两部分由法院认定的客观事实信息，用于后续对刑期的建模预测。

为了更直观地理解该数据集，表 2 给出了样本量最多的十种罪名的基本描述统计（犯罪人年龄中位数、男性犯罪人占比、高中学历及以下占比、刑期中位数、罚款金额中位数和民事赔偿金额中位数）。从表 2 可以总结出以下一些结论：首先，这十种罪名的犯罪人年龄的中位数都不超过 40 岁；其次，尽管这十种罪名的法定刑罚各不相同，但大约有 80% 的案件判处的刑期在 30 个月以下，大约 90% 的案件判处的刑期在 44 个月以下；最后，研究发现，民事赔偿金额的中位数大于罚款金额的中位数。

表 2　样本量最多的前十种罪名的描述统计结果

罪名	年龄	男性（%）	高中学历及以下（%）	刑期（月）	罚款（元）	民事赔偿（元）
盗窃罪	33	91.57	81.09	9	3 000	2 721
故意伤害罪	36	91.01	77.83	12	5 000	24 403
交通肇事罪	37	90.63	75.04	12	4 000	121 116
走私、贩卖、运输、制造毒品罪	36	83.05	79.94	18	5 000	14 337
诈骗罪	33	80.85	66.21	36	10 000	19 387
寻衅滋事罪	30	91.81	75.96	12	5 500	14 240
容留他人吸毒罪	33	85.41	78.25	8	3 000	40 000
开设赌场罪	37	82.26	78.97	12	10 000	20 000
抢劫罪	28	94.87	78.93	42	4 000	10 000
妨害公务罪	38	80.77	75.14	8	5 000	6 322

注：本表展示了样本量最多的十种罪名的犯罪人年龄中位数、男性犯罪人占比、高中学历及以下占比、刑期中位数、罚款金额中位数和民事赔偿金额中位数。

四、量刑偏差案件发现

实现量刑偏差案件自动识别的基础是能够对案件的刑期进行准确的预测，而案件的事实信息是刑期裁决的前提和基础。对于事实情节较为复杂的案件，法官更容易受到认知能力和工作经验等主观因素的影响，从而可能会做出量刑不一致的裁决，导致量刑偏差案件的产生。基于此背景，这一部分使用简单易训练的长短期记忆模型（LSTM）建立刑期长度与犯罪事实的关系，并以此对刑期进行预测。基于模型的预测结果，提出异质性系数用于识别量刑偏差案件，为法官的司法裁定提供辅助。同时，为了对比 LSTM 和其他深度学习模型结论的一致性，参考 Kim（2014）提出的 TextCNN 模型、Graves and Schmidhuber（2005）提出的双向 LSTM（BiLSTM）模型及 Vaswani et al.（2017）提出的 Transformer 模型，本文还训练了 TextCNN、BiLSTM 和 Transformer 三种模型用于结果对比。

（一）LSTM 模型

长短期记忆模型，也叫 LSTM 模型，由 Hochreiter and Schmidhuber（1997）提出，属于循环神经网络（RNN）的一个变种。RNN 模型主要用于处理文本序列，可被视为状态空间模型在文本序列数据上的一种具体实现方法，其核心思想是通过状态变量不断保留、传递历史信息。LSTM 模型则是对 RNN 模型的拓展，其核心是要

同时兼顾长期记忆性和短期记忆性。LSTM 模型设置了隐含层来更新，其中包含三种门控单元：输入门、遗忘门和输出门。输入门决定长期状态变量对当前信息的更新，遗忘门决定对历史状态信息的继承，输出门决定从长期历史状态到当前短期状态的输出。下面具体介绍 LSTM 模型关于刑期预测的模型搭建。

首先，本文将刑期预测问题表述如下。设数据集中共有 N 个案件，经预处理后[2]，案件 $i(1\leqslant i\leqslant N)$ 的事实描述表示为词语序列 $X_i=\{X_{it}:1\leqslant t\leqslant T_i\}$，其中，$T_i$ 为 X_i 的长度，X_{it} 为来自一个给定语料库 W 的词根。由于本文提取了“经审理查明”和“本院认为”两部分文本作为每个案件的事实描述，因此语料库 W 由该事实描述文本中不重复的词根构成，词根则通过中文分词技术获得。例如本文选取的是 Jieba 分词技术，图 3 展示了语料库 W 中词频数最高的前 100 个词根的词云图。令 Y_i 为第 i 个案件的刑期，则 LSTM 模型的目的就是根据文本序列 X_i 对刑期 Y_i 进行预测。值得注意的是，每个案件的事实描述的文本序列 X_i 的长度会不一样，这不符合后续 LSTM 模型对输入的要求（该模型要求输入序列的长度一样），因此，本文计算了每个案件的文本序列长度 T_i 的值及相应的分位数值。结果发现，如果将文本序列的最大长度设置为 3 000 个词根，那么原始文本序列中有超过 99% 的文本信息会被保留下来。因此，我们将 T_i 统一设置为 3 000，对于不足 3 000 的文本序列，本文采取用空格补全的办法使其长度达到 3 000。

图 3　语料库 W 中词频数最高的前 100 个词根的词云图

注：本图展示了样本数据中事实描述部分（即语料库 W）词频最高的 100 个词根。

其次，考虑如何基于 X_i 构建本文的 LSTM 模型以用于刑期预测。参考 Mikolov et al.（2013）提出的词向量嵌入的方法，将语料库 W 里的每一个词根通过词向量技术映射为一个维度为 d 的词向量，本文中令 $d=256$。因此，可以基于某个映射函数 f 得到向量序列 $Z=f(X_{it})$。接下来，考虑如何基于 t 时刻的文本信息和两种历史状态（长期状态和短期状态）预测 $X_{i(t+1)}$，这就产生了如图 4 所示的 LSTM 模型结构示意图。该模型结构可以保证法律文书中有关案件的事实情况在第 t 时刻的文本描述可以

2　即选取“经审理查明”和“本院认为”两部分文本内容作为事实描述。

全部被考虑进模型并用于预测$X_{i(t+1)}$，在模型的最后一层，我们构建了X_{it}和Y_i之间的函数关系。至此，LSTM 模型构建完毕。模型共包含 7 层，除“经审理查明”与“法院认为”、文本序列外，还包括一个维度为 3 000 的输入层，一个维度为 256 的嵌入层，一个维度为 128 的 LSTM 层，一个维度为 50 的全连接层，最后一个是输出层（因为预测的是刑期，因此维度为 1）。为防止模型的过拟合，模型中参考并使用了 Krizhevsky et al.（2012）提出的 dropout 技术，并设置随机失活概率为 0.01，最终，该模型一共需要消耗 25 863 021 个参数。

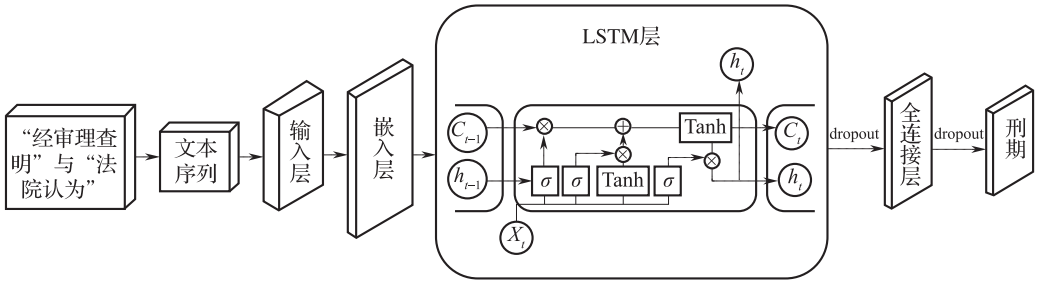

图 4　LSTM 模型结构示意图

注：作者基于研究结果自行整理。

（二）其他深度学习模型

除了 LSTM 模型，本文还考虑了其他三种经常用于 NLP 任务的深度学习模型，分别是 TextCNN 模型、BiLSTM 模型和 Transformer 模型。其中，TextCNN 的结构见图 5，文本序列首先通过输入层和嵌入层，然后并行连接三个不同大小的卷积层和池化层，通过拼接和拉直后，送入全连接层，得到刑期输出。BiLSTM 的模型结构类似于前文的 LSTM，区别在于将 LSTM 层替换为双向 LSTM 层，具体模型示意图见图 6。本文构建的类 Transformer 模型结构见图 7，其主体部分是 Transformer 的一个编码器块（Transformer 块），可将输入序列转化为含有全局注意力信息的向量表示，最后同样将向量表示通过全连接层得到刑期输出。以上三种模型结构，采取与 LSTM 模型一致的预处理方式，分别构建刑期预测模型。

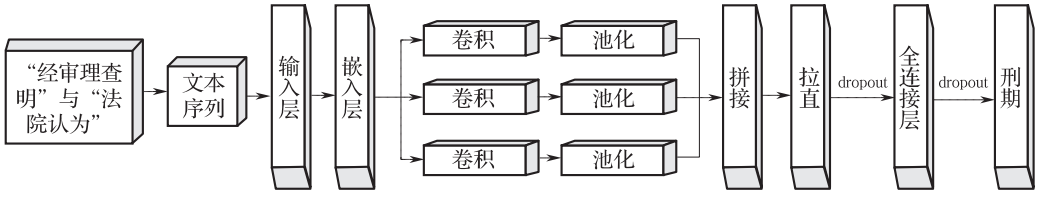

图 5　TextCNN 模型结构示意图

注：作者基于研究结果自行整理。

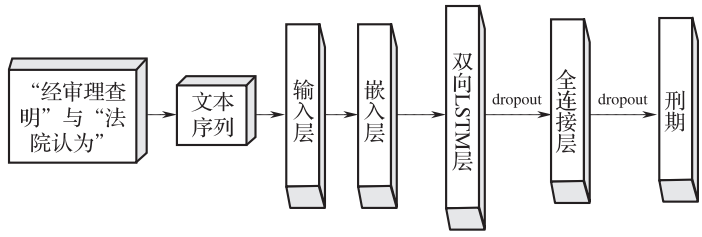

图 6　BiLSTM 模型结构示意图

注：作者基于研究结果自行整理。

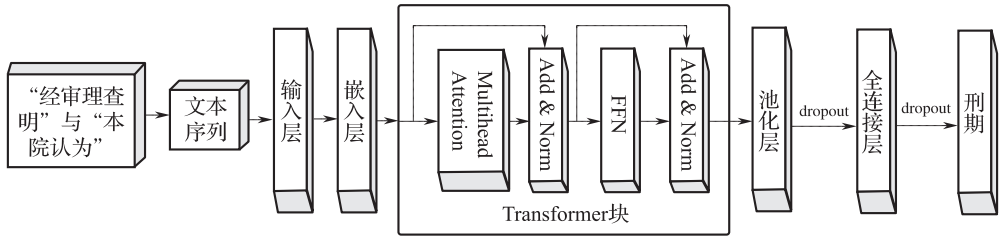

图 7　Transformer 模型结构示意图

注：作者基于研究结果自行整理。

（三）模型训练结果

本文将全样本数据集随机切分为 80% 的样本作为训练集，用于模型训练；20% 的样本作为测试集，用于验证模型效果。由于本文的研究对象刑期长度是连续性变量，因此选择均方误差（MSE）作为损失函数，并将 MSE 作为模型精度的验证指标。在实际训练中，本文对因变量刑期进行了标准化处理，使得方差变为 1，根据回归模型的R^2计算公式（$R^2 = 1 - RSS/TSS$，其中 RSS 为回归模型不能解释的方差，即残差平方和，TSS 为因变量的总平方和）可知，如果将 RSS 和 TSS 都分别除以样本量，那么R^2的计算公式就变为$R^2 = 1 - MSE/Var$（Y），其中 Var（Y）就是因变量的方差 1，因此R^2可以计算为 1 - MSE，用以评估模型的拟合优度。接下来，为了极小化损失函数的值，采取了标准的小批量梯度下降算法，设定批量数 batch size 为 200，使用 Adam 算法进行优化，并且将学习率设置为 0.01。所有实验均在 NVIDIA P100 的 GPU（内存为 16GB）上进行训练。其中，LSTM、TextCNN 和 BiLSTM 模型训练了 50 个批次；由于 Transformer 模型训练一个批次大概需要 50 分钟，因此为了结果对比，只训练了 20 个批次，即便是 20 个批次，从训练曲线上看，模型也基本达到了收敛，见图 8d。图 8 展示了四个模型分别在训练集和测试集上的损失曲线，可以看到，四个模型在训练集中损失函数的值最后在 0.18 上下徘徊，在测试集中损失函数的值最后在 0.22 上下徘徊，模型基本收敛。根据前文的分析，可以计算出 LSTM 模型的R^2为 78% 左右，说明本文构建的 LSTM 模型能够解释刑期变化的 78% 左右，模型具有一定的预测能力。其他三个模型的结论也基本一致。

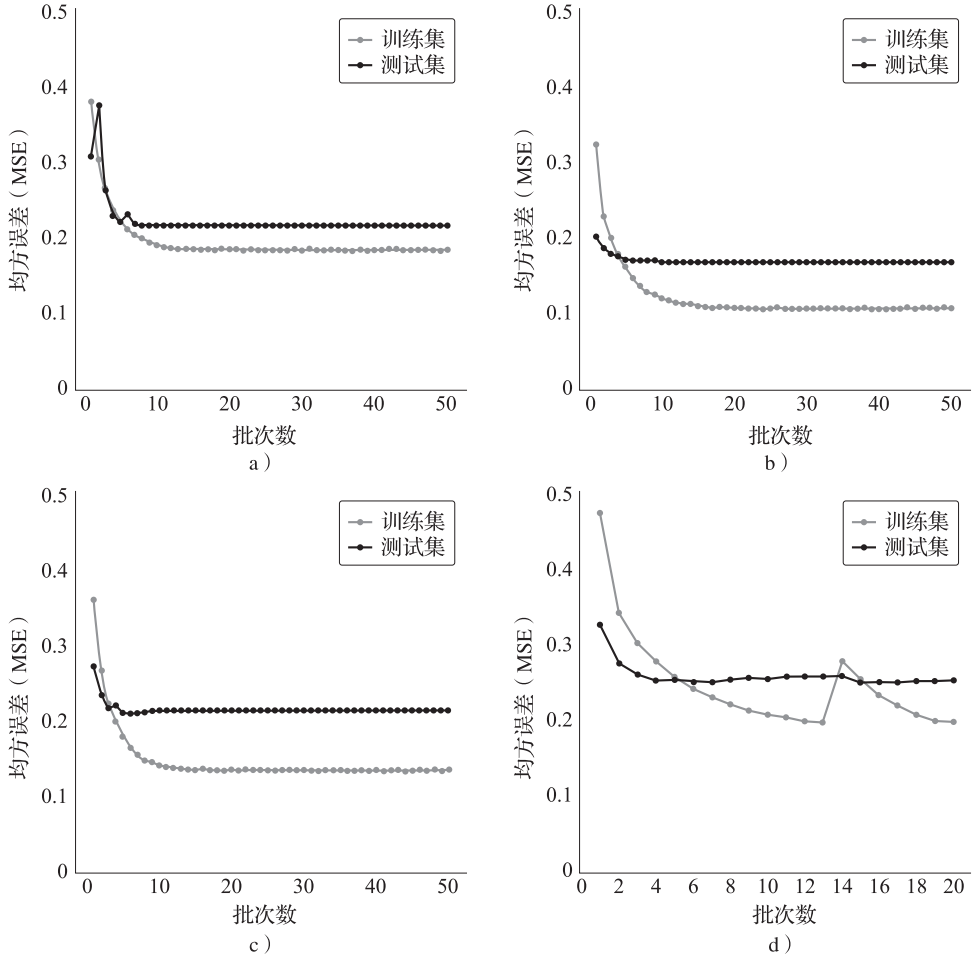

图 8　四个模型在训练集和测试集上的损失曲线

注：本图展示了四种深度学习模型的训练效果。其中，a 为 LSTM 模型在训练集和测试集上的损失曲线，b 为 TextCNN 模型在训练集和测试集上的损失曲线，c 为 BiLSTM 模型在训练集和测试集上的损失曲线，d 为 Transformer 模型在训练集和测试集上的损失曲线。

（四）异质性系数的提出

以 LSTM 模型为例，根据模型的训练结果，可以计算出第 i 个案件的预测值 $\hat{Y}_i$，它表示根据 LSTM 模型给出的预测刑期，该预测值与真实值之间的差异记为残差 $e_i = Y_i - \hat{Y}_i$。残差越大，说明法官和模型对该案件的刑期认识越不一致，这种不一致很可能是由于案件本身比较复杂、存在量刑困难造成的。但是，一个案件的残差值存在着随机性，且无法判断其相对大小，因此很难对案件的量刑偏差程度给出评判，因此，本文首先考虑从犯罪类型上评判哪些罪名更有可能产生量刑偏差案件，然后针对每一种罪名，根据其残差绝对值的大小再进一步判断量刑偏差程度。

在样本数据中，不同罪名的刑期分布有很大的差异。例如，容留他人吸毒罪的刑期分布在 0 个月 ~174 个月之间，均值为 8.7 个月，标准差为 4.7 个月，分布较为集中；而走私、贩卖、运输、制造毒品罪的刑期分布在 0 个月 ~240 个月之间，均值为 48.2 个月，标准差为 56.4 个月，分布较为分散。这说明不同的罪名，量刑范围存在较大的差异。这种差异既可能来自案件本身由于严重程度的不同而造成量刑差异过大，也有可能是由于法官的个人主观判断而产生的量刑偏差。本文参考统计学中离散系数的构造思想，提出了异质性系数，其计算公式为：

$$\text{第 } R \text{ 种罪名的异质性系数} = \frac{\text{第 } R \text{ 种罪名的残差绝对值的中位数}}{\text{第 } R \text{ 种罪名的刑期中位数}}$$

离散系数是用观测资料的标准差除以均值，来度量观察值离散程度的统计量，而本文试图通过异质性系数来度量模型判断和法官判断的不一致程度。具体地，对于第 R 种罪名的异质性系数，其分母“刑期中位数”表示的是过往司法实践中该罪名的一个平均刑期程度，可以近似认为是法官判断的一个平均量刑。分子度量的则是具体到该罪的每个案件模型判断与法官判断的平均差异，近似于离散系数构造中的标准差。这样分子除以分母构造出来的指标（即异质性系数）可以在一定程度上表示模型判断和法官判断的不一致程度。在此背景下，如果模型给出的预测值和法官判决基本一致的话，那么可以说明人和机器对刑期的判断基本不存在争议；如果两者相差较大，则说明人和机器的判断存在不一致，那么这时就要警惕有可能存在量刑偏差案件。

由于深度学习模型的输入变量是案件的客观事实信息，因此模型的预测结果与实际刑期长度的偏差基本反映了客观事实信息解释刑期长度变化的水平。对于每个具体的案件而言，若模型预测结果与实际刑期长度非常接近，说明案件的客观事实可以充分解释刑期长度的变化，此案件应为常规案件；若预测结果与实际刑期长度偏差较大，即客观事实无法充分解释刑期长度变化，说明该案件事实情节复杂，法官在认定事实和适用法律时可能掺杂了较多的主观因素，极有可能是量刑疑难案件。本文在具体计算异质性系数时，删除了缓刑与数罪并罚情况较为集中的罪名，如失火罪、非法持有毒品罪等。某种罪名的异质性系数取值越大，那么该罪名对应案件的量刑预测结果与实际量刑结果偏离较大，说明人与模型存在不一致的看法，极有可能产生量刑偏差案件；某种罪名的异质性系数取值越小，那么该罪名对应案件的量刑预测结果与实际量刑结果倾向于一致，说明人与模型的判断比较统一，那么此类案件比较容易判决，出现量刑偏差案件的可能性较低。表 3 展示了分别根据四种深度学习模型计算出来的异质性系数最高的 5 种罪名，其中有 3 种罪名都被四个模型判断为异质性较高，分别为：制作、复制、出版、贩卖、传播淫秽物品牟利罪（0.592），窝藏、包庇罪（0.550）和挪用资金罪（0.515），其中括号里的数字为根据四种模型计算出来的平均异质性系数得分。

表 3　四种深度学习模型计算出的异质性系数最高的五种罪

排名	LSTM	TextCNN	BiLSTM	Transformer
1	制作、复制、出版、贩卖、传播淫秽物品牟利罪	窝藏、包庇罪	制作、复制、出版、贩卖、传播淫秽物品牟利罪	制作、复制、出版、贩卖、传播淫秽物品牟利罪
2	危险驾驶罪	制作、复制、出版、贩卖、传播淫秽物品牟利罪	挪用资金罪	窝藏、包庇罪
3	窝藏、包庇罪	挪用资金罪	危险驾驶罪	非法行医罪
4	挪用资金罪	挪用公款罪	组织、利用会道门、邪教组织、利用迷信破坏法律实施罪	挪用资金罪
5	组织、利用会道门、邪教组织、利用迷信破坏法律实施罪	职务侵占罪	窝藏、包庇罪	组织、利用会道门、邪教组织、利用迷信破坏法律实施罪

注：本表展示了四种深度学习模型计算的异质性系数得分最高的五种罪名及排序。

五、挪用资金罪量刑偏差的实证分析

通过上述异质性系数的分析，本文首先对最有可能出现量刑偏差的罪名进行了识别，综合四种深度学习模型的预测结果，制作、复制、出版、贩卖、传播淫秽物品牟利罪，窝藏、包庇罪和挪用资金罪是同时被四个模型预测为异质性系数较高的三种罪名。由于制作、复制、出版、贩卖、传播淫秽物品牟利罪和窝藏、包庇罪的样本量较小（前者为 672，后者为 540），为保证实证结果的可靠性，进一步量化影响量刑偏差的因素，识别量刑偏差案件的司法特征，这部分以挪用资金罪为例，进行实证分析。

（一）数据与变量

1. 样本

在 2018 年全部样本中，存在文本内容缺失、数据重复以及文本不符合研究需求的问题，在剔除瑕疵数据后，得到 913 份挪用资金罪的样本数据，空间跨度为 31 个省级行政区[3]。

3　中国共有 34 个省级行政区，样本覆盖不包括台湾省、香港特别行政区、澳门特别行政区。

2. 变量说明

这里的被解释变量为“挪用资金罪的有期徒刑刑期”，解释变量是影响挪用资金罪定罪、量刑的影响因素。根据《刑法》《关于常见犯罪的量刑指导意见（试行）》[4]《关于办理贪污贿赂刑事案件适用法律若干问题的解释》[5]，参考判决文书和相关实证研究文献，删除分布过于稀疏的变量（频次少于 20），共确定了 20 个自变量。其中，犯罪构成情节 2 个，影响量刑的法定量刑情节 4 个、酌定量刑情节 6 个，以及 8 个非法定因素（即本文的控制变量）。具体如下：

犯罪构成情节：挪用资金用途（具体分为用于非法活动、用于经营活动、用于个人且超过三个月未归还）、挪用金额。

法定量刑情节：自首、坦白、立功、累犯。

酌定量刑情节：认罪认罚、退还挪用资金、前科、是否取得被害人谅解、当庭认罪、初犯。

非法定因素：GDP、地域（省份）、是否有辩护人、被告人是否有工作、被告是否为党员、教育年限、性别、年龄。

（二）模型与回归结果

考虑到选取的自变量较多，为了避免自变量之间存在多重共线性，同时也为了提高模型精度，本文采用基于似然比检验（LR）的向后逐步回归方法进行估计。此外，由于挪用资金罪采用的是“数额 + 情节”的量刑方式[6]，考虑到不同地区经济水平差异，模型同时关注挪用资金罪量刑是否具有地域性。基于上述设想，本文分别建立了四个回归模型：模型一（M1）将因变量刑期（有期徒刑，以月为单位）标准化，未控制省份效应；模型二（M2）将因变量刑期（有期徒刑，以月为单位）标准化，并控制省份效应；模型三（M3）的因变量为原始刑期（有期徒刑，以月为单位），未控制省份效应；模型四（M4）的因变量为原始刑期（有期徒刑，以月为单位），并控制省份效应。

通过逐步回归进行变量选择，模型最终保留 2 个犯罪构成情节、2 个法定量刑情节、3 个酌定量刑情节和 6 个非法定因素。针对可能存在的多重共线性问题，本文采用相关系数矩阵和方差膨胀因子（VIF）两种方法来检验多重共线性程度。参照 Lee（2006）的判断方法，变量两两之间的相关系数均小于 0. 85，因此认为不存在严重的多重共线性。各变量的 VIF 值均小于 3，表明不存在明显的多重共线性。四个模型的回归结果和R^2值见表 4。其中，R^2为模型拟合度，R^2越高表示模型

4 资料来源：法发〔2021〕21 号。

5 资料来源：法释〔2016〕9 号。

6 《关于办理贪污贿赂刑事案件适用法律若干问题的解释》规定挪用资金金额、情节均为法定刑升格条件。

表 4　模型回归结果

变量	M1	M2	M3	M4
	标准化刑期	标准化刑期	原始刑期	原始刑期
挪用金额（对数）	0.025 6***	0.026 7***	0.022 2***	0.023 1***
	(0.005 0)	(0.005 2)	(0.004 9)	(0.005 0)
挪用资金用途 - 用于营利活动	-0.124***	-0.134***	-0.137***	-0.150***
	(0.042 8)	(0.044 7)	(0.044 7)	(0.046 1)
挪用资金用途 - 用于非法活动	0.331***	0.326***	0.283***	0.290***
	(0.063 5)	(0.064 5)	(0.061 4)	(0.063 0)
自首	-0.058 0	-0.049 2	-0.056 1	-0.059 4
	(0.061 1)	(0.064 1)	(0.059 1)	(0.062 7)
坦白	-0.125**	-0.113*	-0.126**	-0.128**
	(0.057 9)	(0.059 8)	(0.056 0)	(0.058 5)
前科	0.065 7	0.045 6	0.026 5	0.008 9
	(0.090 2)	(0.091 3)	(0.087 2)	(0.089 2)
退还挪用资金	-0.292***	-0.256**	-0.162	-0.122
	(0.107)	(0.112)	(0.104)	(0.109)
被害人谅解	-0.247***	-0.281***	-0.238***	-0.257***
	(0.059 7)	(0.060 7)	(0.057 7)	(0.059 3)
GDP（对数）	-0.026 9*	0.041 4	-0.017 8	0.028 2
	(0.016 2)	(0.068 1)	(0.015 6)	(0.066 6)
是否有辩护人	0.311***	0.311***	0.271***	0.271***
	(0.046 3)	(0.048 2)	(0.044 7)	(0.047 1)
性别	0.014 9	0.014 7	-0.037 0	-0.041 5
	(0.058 6)	(0.059 5)	(0.056 6)	(0.058 1)
被告学历	0.026 9***	0.025 7***	0.026 3***	0.025 1***
	(0.007 9)	(0.008 1)	(0.007 6)	(0.007 9)
被告是否为党员	-0.125	-0.172*	-0.095 3	-0.109
	(0.092 0)	(0.093 8)	(0.088 9)	(0.091 7)
省份效应	否	是	否	是
样本量	913	913	913	913
R^2	0.16	0.21	0.14	0.17

注：本表展示了四个回归模型的实证结果，其中 *、**、*** 分别表示在 10% 、5% 和 1% 的水平上显著，系数下方括号内的数值为标准误差。

拟合度越好。在本文模型中，当实际刑期与预测刑期之间的残差越大时，量刑结果的确定性越小，模型拟合度越低。挪用资金罪各自变量回归系数见表 4，其中括号中的数字为标准误差。

对比四个模型的回归结果，解释变量在各模型中的影响方向、显著性、系数大小大体相同，这说明模型具有一定的稳健性。

（三）量刑偏差原因分析

从回归结果来看，大部分具有显著性的回归结果符合一般认知，如坦白对量刑有减轻作用、挪用金额大小与量刑长短成正比等，但“挪用资金用途 – 用于营利活动”这一变量系数存在异常。在挪用资金罪中，根据资金使用去向的不同，挪用资金可以分为三种：挪用资金用于个人且超过三个月未归还、挪用资金用于营利活动、挪用资金用于非法活动。从法律规定上看，三者刑法评价的严厉程度逐步增加，刑罚逐步严厉。但根据回归结果，“挪用资金用途 – 用于营利活动”的回归系数为负数，刑罚相较于“挪用资金用途 – 用于个人且超过三个月未归还”却更为轻缓。

1. 异常系数原因分析

模型二（M2）结果表明，控制其他变量不变，与挪用资金用于个人且超过三个月未归还相比，挪用资金用于营利活动量刑刑期减少。挪用资金用于营利活动是以获得金钱或物质回报为目的的行为，天然具有较大的风险性，犯罪嫌疑人挪用资金进行营利活动的风险明显高于归个人使用的挪用行为。挪用资金用于营利活动行为的法律评价比挪用资金归个人使用要更为严厉，因此挪用资金用于营利活动的系数应当为正数，但回归结果却显示为负。

造成回归系数异常的原因可以从客观因素和法官心理因素两方面来分析。从客观因素看，对挪用资金进行营利活动所获取的利息、收益进行追缴，强制剥夺被告人对吸收资金的占有和使用，有助于打击和预防挪用资金的犯罪行为，对量刑起到了缓和作用。从法官心理因素看，被告人积极退还挪用资金并与被害人达成赔偿协议等悔罪表现，使得法官认为其主观恶性较小，从而有从宽处罚倾向。

2. 案例佐证

案例一：（2018）琼 9023 刑初 86 号

基本案情：2018 年 2 月 9 日澄迈县人民检察院提起公诉，指控被告人朱绵武犯挪用资金罪。被告人朱绵武私自将集体资金 22 万元挪用给他人用于营利活动，数额较大，追诉前已退还。海南省澄迈县人民法院判决被告人朱绵武犯挪用资金

罪，判处有期徒刑九个月，缓刑一年。

案例二：（2017）云 0502 刑初 388 号

基本案情：2017 年 7 月 26 日重庆市开州区人民检察院提起公诉，指控被告人张志良犯挪用资金罪。被告人张志良挪用村民建房相关费用人民币 265 500 元归个人使用，数额较大，未退还。云南省保山市隆阳区人民法院判决被告人张志良犯挪用资金罪，判处有期徒刑三年。

案例一、二中两个被告分别为挪用资金用于个人且超过三个月未归还、挪用资金用于营利活动已退还两种法律情形，两个被告人挪用金额大致相当。以上案例能够明显看出，在挪用金额及其他量刑情节基本相同的情况下，两案量刑差异较大。

六、结论与研究局限

本文以量刑为研究对象，选取 2018 年 1 月 1 日至 2018 年 12 月 31 日全年的刑事裁判文书数据，对量刑影响因素进行相关研究，并提出了异质性系数指标，用于辅助识别量刑偏差案件。具体地，本文以刑期作为因变量，构建了四种深度学习模型（LSTM、TextCNN、BiLSTM、Transformer）用于刑期的研究，基于模型结果可以为每个罪名计算异质性系数得分，该指标得分越高说明该种罪名越有可能产生量刑偏差案件。研究发现，制作、复制、出版、贩卖、传播淫秽物品牟利罪，窝藏、包庇罪和挪用资金罪是异质性系数得分较高的三种罪名。为进一步量化影响量刑偏差的因素，识别量刑偏差案件的司法特征，本文以挪用资金罪为例，进行了相关实证分析。多元线性回归分析结果表明，坦白等法定或酌定量刑情节对量刑有减轻作用，挪用资金金额大小与量刑长短成正比。然而，也存在与司法解释相悖的一些现象，例如挪用资金用途不同对量刑影响不同，这有可能是挪用资金罪异质性突出的原因。本文的意义并非通过建立模型精确预测刑期长度，从而取代法官的裁决结果，而是在确定量刑机制的基础上，对量刑偏差案件进行识别，为量刑司法实践提供参考和支撑，提升量刑体系的规范性。

尽管本文的研究结论对识别量刑偏差案件具有一定的辅助作用，但仍然存在一些不足，值得后续持续研究。首先，本文在样本的选取上并未包含案件数较少的案件（例如，本文选取的都是案件数在 500 以上的罪名），对于案件数较少的罪名的量刑偏差识别应作为一个单独的研究在未来进行探索。其次，本文只考虑了刑事犯罪的量刑问题，并未考虑其他法律范畴（例如民事）的量刑问题，本文的方法论是否仍然适用于民事或其他法律范畴也值得进一步研究。再

次，本文在实证分析阶段仅使用了法定量刑情节、酌定量刑情节等关键词，这些关键词的选取极度依赖研究人员的专业知识，难免会造成提取不全面的问题，因此如何基于大量的司法文书数据，采取前沿的机器学习算法自动提取对量刑有重要影响的关键因素，也是一个非常重要且值得未来单独进行研究的一个问题，该问题的研究或许能对实证研究中的 R^2 提升具有重要帮助。最后，本文利用了前沿的人工智能方法对刑期进行了预测，并发展了异质性系数用于量刑偏差案件的识别，但目前还无法对量刑偏差的具体原因做自动化的识别。本文在量刑偏差原因的识别上采取的是一种事后分析的方法，即通过回归模型去探索影响偏差产生的因素。通过人工智能的方法同时分析关于刑期预测、异质性系数以及量刑偏差的具体原因，在未来仍然是一个值得研究的问题。

综上，尽管本文存在一定的局限，所提模型虽并不能给出一个绝对正确的量刑建议，但是可以客观地描述司法实践中最重要的共识，以及共识一致的程度。本文在此基础上，尝试为量刑偏差问题提供一些参考建议，为刑罚理论的完善提供支撑。

参考文献

[1] 白建军，2016. 基于法官集体经验的量刑预测研究[J]. 法学研究，38(6)：140－154.

[2] 白建军，2017. 法律大数据时代裁判预测的可能与限度[J]. 探索与争鸣，(10)：95－100.

[3] 白建军，2020. 法秩序代偿现象及其治理——从妨害公务罪切入[J]. 中外法学，32(2)：418－443.

[4] 樊祐玺，万力，2019. 危险驾驶罪基本特征及量刑影响因素实证研究——基于 700 份醉酒型危险驾驶案件一审裁判文书的分析[J]. 医学与法学，11(2)：86－90.

[5] 高通，2020. 故意伤害案件中赔偿影响量刑的机制[J]. 法学研究，42(1)：154－170.

[6] 胡昌明，2018. 被告人身份差异对量刑的影响：基于 1060 份刑事判决的实证分析[J]. 清华法学，12(4)：91－110.

[7] 江溯，2021. 以危险方法危害公共安全罪认定规则研究[J]. 中国法学，(4)：221－246.

[8] 李大鹏，赵琪珲，邢铁军，赵大哲，2022. 基于分层注意力循环神经网络的司法案件刑期预测[J]. 东北大学学报(自然科学版)，43(3)：344－349.

[9] 马建刚，马应龙，2019. 语义驱动的司法文档学习分类方法[J]. 计算机应用，39(6)：1696－1700.

[10] 舒洪水，2020. 司法大数据文本挖掘与量刑预测模型的研究[J]. 法学，(7)：113－129.

[11] 孙道萃，2020. 人工智能辅助精准预测量刑的中国境遇——以认罪认罚案件为适用场域[J]. 暨南学报(哲学社会科学版)，42(12)：64－78.

[12] 孙海波，2017. 疑难案件裁判的中国特点：经验与实证[J]. 东方法学，(4)：52－63.

[13] 谭红叶，张博文，张虎，李茹，2020. 面向法律文书的量刑预测方法研究[J]. 中文信息学报，34(3)：107－114.

[14] 王剑波，2018. 行政级别、身份性质与我国受贿罪的量刑差异[J]. 政法论坛，36(1)：93－107.

[15] 王治政，王雷，李帅驰，孙媛媛，陈彦光，许策，王刚，林鸿飞，2021. 基于多视角知识图谱嵌入的量刑预测[J]. 模式识别与人工智能，34(7)：655－665.

[16] 文姬，2016. 醉酒型危险驾驶罪量刑影响因素实证研究[J]. 法学研究，38(1)：165－186.

[17] 文姬，黄雪，2020. 行贿罪量刑与省域经济水平关系实证研究[J]. 辽宁师范大学学报(社会科学版)，43(2)：49－53.

[18] 吴雨豪，2021. 量刑自由裁量权的边界：集体经验、个体决策与偏差识别[J]. 法学研究，43(6)：109－129.

[19] 章桦，2020. 贪污罪"数额与情节"关系实证研究——基于全国 18392 例量刑裁判[J]. 法学，(6)：175－192.

[20] 章桦，李晓霞，2014. 醉酒型危险驾驶罪量刑特征及量刑模型构建实证研究——基于全国 4782 份随机抽样判决书[J]. 中国刑事法杂志，(5)：99－108.

[21] 张玉洁，2021. 智能量刑算法的司法适用：逻辑、难题与程序法回应[J]. 东方法学，(3)：187－200.

[22] 赵学军，2019. 量刑偏差的司法表现与量刑规范的实现路径——基于抢劫罪刑事判决书的实证考察[J]. 天津法学，35(3)：57－63.

[23] 左卫民，2021. AI 法官的时代会到来吗——基于中外司法人工智能的对比与展望[J]. 政法论坛，39(5)：3－13.

[24] ALETRAS N, TSARAPATSANIS D, PREOTIUC-PIETRO D, LAMPOS V, 2016. Predicting judicial decisions of the European court of human rights: a natural language processing perspective [J]. Peerj Computer Science, 2(10). DOI: 10. 7717/cs. 93.

[25] CHEN H J, CAI D, DAI W, DAI Z H, DING Y D, 2019. Charge-based prison term prediction with deep gating network [J]. ArXiv Preprint. DOI: 10. 48550/arXiv. 1908. 11521.

[26] GRAVES A, SCHMIDHUBER J, 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 18(5－6): 602－610.

[27] HOCHREITER S, SCHMIDHUBER J, 1997. Long short-term memory [J]. Neural Computation, 9(8): 1735－1780.

[28] HU Z K, LI X, TU C C, LIU Z Y, SUN M S, 2018. Few-shot charge prediction with discriminative legal attributes [C] //Proceedings of the 27th International Conference on Computational Linguistics, 487－498.

[29] JOHNSON R, ZHANG T, 2003. Supervised and semi-supervised text categorization using LSTM for region embeddings [C] // Proceedings of the 33rd International Conference on Machine Learning, 526－534.

[30] KIM Y, 2014. Convolutional neural networks for sentence classification[C] //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746－1751.

[31] KRIZHEVSKY A, SUTSKEVER I, HINTON G, 2012. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 25(2): 84－89.

[32] LEE G, 2006. The effectiveness of international knowledge spillover channels [J]. European Economic Review, 50(8): 2075－2088.

[33] LI S, ZHANG H, YE L, GUO X, FANG B, 2019a. MANN: a multichannel attentive neural network for legal judgment prediction [J]. IEEE Access, 7(1): 151144－151155.

[34] LI Y, HE Y K, YAN G, ZHANG S, WANG H, 2019b. Using case facts to predict penalty with deep learning [C]// IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C).

[35] LUO B F, FENG Y S, XU J B, ZHANG X, ZHAO D Y, 2017. Learning to predict charges for criminal cases with legal basis [J]. ArXiv Preprint. DOI: 10. 18653/v1/D17－1289.

[36] MIKOLOV T, CHEN K, CORRADO G, DEAN J, 2013. Efficient estimation of word representations in vector space [J]. Procedia Computer Science, 80: 2205－2210.

[37] VASWANI A, SHAZEER N M, PARMAR N, USZKOREIT J, JONES L, GOMEZ A N, KAISER L, POLOSUKHIN I, 2017. Attention is all you need [J]. ArXiv Preprint. DOI: 10. 48550/arXiv. 1706. 03762.

[38] WAN S X, LAN Y Y, GUO J F, XU J, PANG L, CHENG X Q, 2016. A deep architecture for semantic matching with multiple-positional sentence representations[C]. ArXiv Preprint. DOI: 10. 48550/arXiv. 1511. 08277.

[39] ZHONG H X, XIAO C J, TU C C, ZHANG T Y, LIU Z Y, SUN M S, 2014. How does NLP benefit legal system: a summary of legal artificial intelligence [J]. ArXiv Preprint. DOI: 10. 48550/arXiv. 2004. 12158.

[40] ZHOU P, QI Z, ZHENG S, XU J, BO X, 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling[C] // The 26th International Conference on Computational Linguistics, 3485 – 3495.

Research on Artificial Intelligence Assisted Judge Decision-Making —From the Perspective of Sentencing Deviation Identification

Jing Zhou
(Center for Applied Statistics and School of Statistics, Renmin University of China)

Lingyan Yang
(Data Science Institute, Shandong University)

Zhe Liu
(Center for Applied Statistics and School of Statistics, Renmin University of China)

Fang Wang*
(Data Science Institute, Shandong University)

Summary: Sentencing is the ultimate embodiment of penal justice. To achieve the goal of "making people feel fairness and justice in every judicial decision," the Chinese Supreme People's Court continues to reform the standardization of sentencing. For this purpose, the Supreme People's Court issued the "Sentencing Guidance for the People's Court" (referred to as the Guidance) in 2008, which has since been revised six times. The Guidance provides comprehensive guidelines for the basic methods and steps of sentencing, the scope of common sentencing circumstances, and the sentencing of common crimes. However, real-world scenarios are complex, and the Guidance cannot cover all situations. Furthermore, differences in regional economic and social development levels, divergent rulings among individual judges, and the personal characteristics of defendants may lead to inconsistent sentences for similar cases. This may lead to a low rate of settlement, as seen in 2020 when the Supreme People's Court heard a total of 1.12 million first-instance cases, of which about 11% and 2% went through a second trial and remand for retrial, respectively. This indicates that numerous cases have not yet been settled (without appeal or reverse appeal), and many of these may be controversial cases where judges may hold divergent opinions on specific sentencing such as imprisonment or fines. This low rate of settlement may also influence the judicial process and is detrimental to safeguarding the authority and credibility of the law. Some scholars have suggested that the individual judge's discretion in sentencing should be compared with the collective experience of judges in sentencing. Judges who make rulings that reflect the collective experience should be supported and respected for their discretion, while judges who deviate significantly from the collective experience should have their decisions identified and corrected. Since 2016, the Chinese Supreme Court has vigorously promoted the construction of smart courts, hoping to use big data and artificial intelligence technology to discover judicial consensus. This would improve the accuracy and fairness of case acceptance and trials, and enable the judicial system to make fair and consistent rulings.

* Corresponding Author: Fang Wang, Data Science Institute, Shandong University, E-mail: wangfang226@sdu.edu.cn.

From the perspective of trial supervision, this article proposes a technical method for automatically detecting sentencing deviations. Since 2021, China Judgments Online (https://wenshu.court.gov.cn/) has released more than 100 million legal judgment documents to the public. This undertaking has provided a massive data foundation not only for research related to judicial judgments, but also for developing advanced machine learning algorithms that can automatically detect deviations in sentencing. In this article, we take the criminal judgment documents from 2018 as the sample and analyze a total of 460 - 486 legal documents based on 62 charges. We then propose a method that can accurately detect abnormal situations of sentencing deviation in judicial trials. Specifically, the model includes the following three aspects. First, using the sentence as the dependent variable and the text extracted from the "as determined through trial" and "considered by the court" in the legal documents as the description of case facts, four deep learning models (LSTM, TextCNN, BiLSTM, Transformer) are constructed for sentence prediction. Second, based on the predicted results of the model, the difference between the predicted sentence and the actual sentence is calculated. Finally, we propose a heterogeneity index that is used to identify the charges that have deviations in sentencing. Inspired by the coefficient of variation in statistics, the heterogeneity index is constructed as

$$\text{Heterogeneity index of the R}^{\text{th}}\text{ crime} = \frac{\text{Median of the absolute residual value of the R}^{\text{th}}\text{ crime}}{\text{Median of the sentence of the R}^{\text{th}}\text{ crime}}$$

In this article, the heterogeneity index is used to measure the degree of inconsistency between the model judgment and the judge's judgment. Specifically, for the heterogeneity index of the R^{th} crime, the denominator "median sentence" represents an average sentence level for the crime in past judicial practices and can be roughly considered the average sentencing by judges. The numerator measures the average difference between the model judgment and the judge's judgment for each case of the crime, which is very similar to the standard deviation in the construction of the coefficient of variation. Thus, the heterogeneity index can to some extent represent the degree of inconsistency between the model judgment and the judge's judgment. In this context, if the prediction values given by the model and the judge's ruling are basically consistent, it can be concluded that no dispute between human and machine judgments exists for that particular sentence. However, if the difference between the two is large, then inconsistency exists between human and machine judgments, and sentencing bias may be present. The calculation shows that the crimes of producing, copying, publishing, selling, spreading obscene materials for profit, crimes of harboring and sheltering, and crimes of misappropriation of funds are the top three charges with the highest heterogeneity index, which indicates that these three charges are most likely to result in cases of sentencing deviation.

To further quantify the factors that influence sentencing deviation and identify the judicial characteristics of such cases, this article takes the crime of misappropriation of funds as an example and conducts related empirical analysis. The results of regression analysis show that mitigating circumstances, such as confession, and statutory or discretionary sentencing factors have a mitigating effect on sentencing. The size of the misappropriated funds is directly proportional to the length of the sentence. However, there are also some phenomena that contradict judicial interpretations. For example, the defendant's purpose for misappropriating funds has an impact on sentencing, which may be due to the prominent heterogeneity of this specific crime. The goal of this article is not to replace the judge's ruling by accurately predicting the length of the sentence through establishing a model. Rather, it is to identify cases of sentencing deviation based on the established sentencing mechanism, provide reference and support for sentencing judicial practice, and enhance the normativity of the judicial system.

Keywords: Judicial Artificial Intelligence; Sentencing Prediction; Sentencing Disparity; Heterogeneity Measurement; LSTM

JEL Classification: K14; C45; B23