



开放科学
(资源服务)
标识码
(OSID)

刑事案件上诉原因文本挖掘 ——基于二审文书数据

温书豪¹ 黄学军² 周良³ 赵英³

1. 中国共产党钦州市委员会政策研究室 钦州 535000;
2. 四川大学计算机学院 成都 610065;
3. 四川大学公共管理学院 成都 610065

摘要: [目的/意义] 从刑事二审案件裁判文书中挖掘上诉理由和相关影响因素, 给法院和智慧量刑系统提供相关数据。[方法/过程] 以北大法宝网近一年的刑事二审案件裁判文书作为基础数据, 用信息抽取、word2vec 训练词向量和聚类等文本挖掘方法对文本内容进行挖掘。[结果/结论] 在传统的上诉理由之外, 发现了基于上诉人自身态度的上诉理由。信息抽取、word2vec 训练词向量和聚类等文本挖掘方法可用于裁判文书相关内容挖掘。

关键词: 刑事二审案件; 上诉原因; 裁判文书; 文本挖掘

中图分类号: G202, G35

The Influencing Factors Analysis of Second Criminal Instance Based on Text Mining

WEN Shuhao¹ HUANG Xuejun² ZHOU Liang³ ZHAO Ying³

1. Policy Research Office, CPC Qinzhou Committee, Qinzhou 535000, China;
2. College of Computer Science, Sichuan University, Chengdu 610065, China;
3. School of Public Administration, Sichuan University, Chengdu 610065, China

Abstract: [Purpose / significance] This study aims to explore the reasons for appeal and related influencing factors from the judgment documents of criminal cases of second instance, so as to provide relevant data for the court and intelligent sentencing

基金项目 国家重点研发计划“高质高效的审判支撑关键技术及装备研究”(2018YFC0830300)。

作者简介 温书豪(1995-), 硕士, 研究方向为信息技术应用; 黄学军(1971-), 硕士, 副教授, 硕士生导师, 研究方向为语义网, 知识图谱等; 周良(1984-), 博士, 助理研究员, 研究方向为在线个性化定制、用户信息行为分析, E-mail: zhouliang_bnu@163.com; 赵英(1968-), 博士, 教授, 博士生导师, 研究方向为信息管理、信息无障碍、信息系统等。

system. [Methods/Process] This study takes the judicial documents of criminal cases in the second instance of Beijing University's magic treasure website in recent one year as the basic data, and the text content was mined by using text mining methods such as information extraction, word2vec training word vector and clustering. The results show that in addition to the traditional reasons for appeal, the appellant's own attitude was found. [Results /Conclusions] The results also indicated that the text mining methods such as information extraction, word2vec training word vector and clustering can be used to mine relevant content of judicial documents.

Keywords: Criminal cases of second instance; reasons for appeal; judicial documents; text mining

引言

2017年3月12日,最高人民法院院长周强在全国人大会议上提出“加快建设智慧法院,努力提供更多优质司法服务”的目标。法院系统的司法裁判体系越来越重视对互联网、大数据和人工智能等技术的运用,并借助这些技术对现有的司法裁判体系进行改革。

但目前人工智能在司法领域中的应用存在缺陷。一方面,法律数据匮乏、低质、结构化不足,算法存在缺陷,目前法律人工智能在实践中的效果不佳^[1]。另一方面,我国目前申请二审的案件多,在一审案件基数大、缺乏智慧量刑系统辅助的情况下,现有的司法资源与要求的司法资源相比严重不足,导致法官办案压力大^[2],可能会在一定程度上影响着审判结果;另外,众所周知,一些二审上诉的案件最后还是维持了原判,而二审程序却耗费了大量的人力物力和时间,大大增加了法院的工作强度。

裁判文书是对中国司法实践全面貌、原始面貌最为具体和客观的反映^[3],目前用作人工智能开发的法律数据基本都是裁判文书。2010年、2013年和2016年,最高人民法院前后三次发布了《关于人民法院在互联网公布裁判文书的规定》,裁判文书上网公开进程大大加快,使得裁判文书的获取更加便利。

因此,随着裁判文书获取的便利性提高,为了在司法实践中减少刑事二审上诉的发生,使司法资源的配置使用更加规范合理,实现二审程序的公平公正,提升二审的效率,也为了支持智慧量刑系统的研发和修改等。本文探索了基于二审裁判文书的刑事案件上诉原因的挖掘方法,通过获取互联网上的刑事二审裁判文书并提取相关内容,利用文本挖掘的方法分析上诉理由集中在哪些方面,有什么特征,从而可将得到的上诉理由相关信息提供给法官、智慧量刑系统工作者等,为法官审判和智慧量刑系统的开发提供信息服务和辅助参考。

1 文献综述

1.1 裁判文书研究现状

马超等^[4]对中国裁判文书网公布的所有文书的全面性、及时性和已经公开内容的妥当性进行了全面考察。郭叶等^[5]以最高人民法院发布的指导性案例为研究对象,北大法宝网司法案例库的裁判文书作为数据样本,从不同角度分析了指导性案例司法应用的特点和规律。张忠民^[6]利用五千多份环境裁判文书作为样本,对生态破坏案件的时间、地点、数量、审理程序、案由、举证责任分配、因果关系等方面进行了

全面的分析,旨在完善生态破坏的司法救济。黄启辉^[7]通过对40家法院2767份裁判文书的统计分析,对行政诉讼一审的审判状况进行了研究。朱春华^[8]基于8家法院的3980份裁判文书的统计分析,对行政诉讼二审审判状况进行了研究。Callander等^[9]通过对无过错赔偿涉及旅游事故判决案件的分析,探讨了冒险旅游经营者在旅游事故中可能遭受的后果。Blynch等^[10]利用1984—1994年间的500多个案例分析了医疗专家对潜在医疗法律索赔的意见。Chandler等^[11]通过对加拿大2008年至2012年间与刑事罪犯的责任和累犯风险有关的神经科学证据的裁判文书,讨论了加拿大刑事司法系统目前如何使用神经科学证据的问题。

综上,在以裁判文书为对象的研究中,目前国内学者主要从两个方面展开:一是基于数据本身,主要是对裁判文书公开制度、问题和数据分布情况等问题的研究;二是基于数据内容的分析,主要是对裁判文书具体内容的分析。

1.2 刑事案件上诉理由研究现状

程荣斌、邓云等^[12]认为我国两审终审制存在弊端,应区分不同案件,分别实行两审终审和三审终审的二元化审级制度,并科学地解决上诉审的审查范围、方式和上诉理由限制等问题。孙远^[13]认为在立法没有对上诉制度作出根本性变革之前,应以上诉理由为标准,对二审案件进行分流,从而尽可能实现刑事上诉审构造的合理化。李昌林、王丹^[14]认为要保障上诉人的权利,应对上诉理由对二审案件进行分流,因此要将上诉理由应规范化、明确化和法定化。杨杰辉^[15]认为上诉理由具有防止滥上诉、提高

诉讼效率以及明确审判对象、限制审判范围等功能,因此应该规定上诉理由,防止滥上诉,保障上诉案件能够公正处理。

综上,对刑事案件上诉理由的研究可以分为两个方面:一是将刑事案件的上诉理由作为研究的一部分,分析刑事案件的上诉制度和程序等;二是对刑事案件中上诉理由功能的分析。

1.3 判决书文本挖掘研究现状

文本挖掘是通过计算机、机器学习、自然语言学以及统计学知识和技术,快速地从海量的文本信息中抽取出事先未知的、可理解的、隐藏的有价值的知识和过程^[16]。其能批量处理文本,大大提高数据分析的效率。文本挖掘流程一般包括数据预处理、知识发现、模式评价和结果解释与呈现四个部分^[17],分析可以从文本结构分析、文本摘要、文本聚类、情感分析和趋势预测等方面展开。

王景峰^[18]利用数据挖掘方法从裁判文书中提取案件事实及其适用法律条文信息,构建数据集进行学习得到了案件适用法律的自动识别模型。王业沛^[19]从裁判文书中抽取关键特征,利用深度学习技术对判决结果的倾向性进行了分析,发现基于深度学习的模型准确率普遍优于基于词典规则的模型。王海亮^[20]通过对大量裁判文书的挖掘构建了法律咨询系统,向用户推荐律师。梁楠^[21]利用文本挖掘技术从裁判文书中获取律师的表现信息,对律师进行排名从而向用户推荐律师。高雅婷^[22]基于部分裁判文书和相关数据利用文本挖掘技术设计和开发了刑事案件文本挖掘系统。舒洪水^[23]以毒品判决书为文本挖掘对象,利用TF-IDF、N-Gram等

技术方法对量刑进行了预测,构建了量刑预测模型。

综上,对裁判文书进行文本挖掘的研究主要是利用相关方法和手段对文书相关内容进行提取,构建模型,从而预测判决结果、量刑和律师推荐等。

从以上三个方面可知,目前缺乏利用相应文本挖掘技术对上诉理由的专门分析。随着裁判文书获取更加便捷,本文选择从互联网相关裁判文书网上获取刑事案件二审裁判文书,利用相关工具和方法,对刑事案件上诉理由及相关内容进行挖掘和分析。

2 研究思路和框架

为了从裁判文书网站获取刑事案件二审裁判文书并利用文书内容提取上诉理由及相关影响因素,本文探索了基于文本挖掘的刑事案件二审上诉原因的方法。该方法主要包括三个方面:数据收集阶段、数据预处理阶段和上诉理

由提取阶段。具体如下:

(1) 数据收集阶段

在数据收集阶段,确定裁判文书获取的网站,解析网页的结构,通过相关爬虫协议获取裁判文书全文和相关数据字段。

(2) 数据预处理阶段

在数据预处理阶段,首先对数据中存在字段空缺、数据错位、条目重复、内容重复等问题进行数据清洗;其次使用信息抽取的相关方法提取裁判文书中的案由、上诉人身份、上诉理由文本、审判结果;最后对上诉理由文本进行分词。

(3) 上诉理由提取

在上诉理由提取阶段,首先使用相关工具对分词好的上诉理由文本训练词向量模型并保存;然后将上诉理由文本中不同词性的词分类,找出最能代表上诉理由词性的词文本;最后通过聚类算法基于保存的词向量模型对词性文本中的词进行聚类。

具体流程如图1所示。

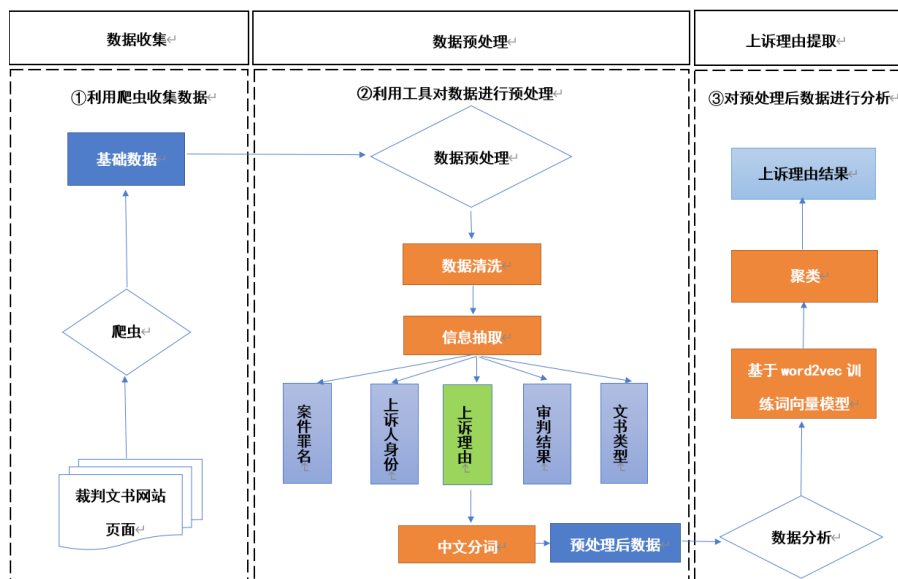


图1 基于文本挖掘的裁判文书分析流程图

总体而言,本文的方法是从两个方面进行分析:一是利用信息抽取的方法抽取出刑事案件二审裁判文书文本中的关键信息,包括案件包含的案由、上诉人身份和审判结果等,并分析这些相关结果中哪些因素会更多的出现在刑事案件二审上诉中,即说明这些因素出现时刑事案件二审上诉的可能性更大。二是利用文本挖掘的方法对上诉理由进行提取和分类,进而分析刑事案件二审的具体上诉原因主要出现在哪些方面,它们又具有什么样的特征。

3 刑事案件二审裁判文书数据收集与预处理

3.1 数据收集

“北大法宝”是由北京大学法制信息中心

与北大英华科技有限公司联合推出的智能型法律信息一站式检索平台。系统现存司法案例已近千万篇,是国内现存裁判文书数据量较多、覆盖面较广的数据库之一。因此在裁判文书网站的选取上,本文选择了“北大法宝”网。

裁判文书网上公开的范围和数量是随着时间变化而逐渐增加的,也就是说在相关裁判文书网站,年份距现在越近,裁判文书的数量和类型会更全面,每种类型的文书数量也越多,即越近年份的数据越具有代表性。为了较全面地分析刑事案件二审的上诉原因,在北大法宝网上检索了2019年1月1日—2019年8月22日(论文数据收集截止时)之间的所有刑事案件二审裁判文书一共48346篇。最后由于北大法宝网站的Ip浏览数量限制等原因,共爬取到2019年3月18日—2019年8月22日的刑事二审裁判文书28411篇(图2)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	文书名称	案由分类	案件字号	文书类型	时间	全文										
2	刘德华非法经营罪	(2019)湘02判决书	2019.08.22	刘德华非法经营二审刑事判决书	湖南省株洲市中级人民法院刑事判决书(2019)湘02刑终186号	原公诉机关湖南省株洲市										
3	刘实强、文抢劫罪	(2019)湘13裁定书	2019.08.22	刘实强、刘实强抢劫、盗窃二审刑事裁定书	湖南省娄底市中级人民法院刑事附带民事裁定书(2019)湘13刑终513号	原公诉机关										
4	刘大安信用卡诈骗罪	(2019)吉02裁定书	2019.08.22	刘大安信用卡诈骗罪二审刑事裁定书	吉林省吉林市中级人民法院刑事裁定书(2019)吉02刑终195号	原公诉机关吉林省吉										
5	蔡某1交通肇事罪	(2019)吉02裁定书	2019.08.22	蔡某1交通肇事罪二审刑事裁定书	吉林省吉林市中级人民法院刑事附带民事裁定书(2019)吉02刑终200号	原公诉机关吉林										
6	党振华、诈骗罪	(2019)湘02裁定书	2019.08.21	党振华、党某某诈骗、掩饰、隐瞒犯罪所得、犯罪所得收益二审刑事裁定书	湖南省株洲市中级人民法院刑事裁定书(2019)湘											
7	成子文、包某聚众斗殴罪	(2019)湘02裁定书	2019.08.21	成子文、包某聚众斗殴二审刑事裁定书	湖南省株洲市中级人民法院刑事裁定书(2019)湘02刑终256号	原公诉机关湖南省株										
8	郑雪飞与唐中波组织、强迫、引诱、容留、介绍卖淫罪	(2019)青01裁定书	2019.08.21	郑雪飞与唐中波组织、强迫、引诱、容留、介绍卖淫罪二审刑事裁定书	青海省西宁市中级人民法院刑事裁定书(2019)青01刑											
9	被告人交某贪污罪	(2019)青01裁定书	2019.08.21	被告人交某贪污罪二审刑事裁定书	青海省西宁市中级人民法院刑事裁定书(2019)青01刑终161号	原公诉机关湟源县人民										
10	牛浩然危险驾驶罪	(2019)内01裁定书	2019.08.21	牛浩然危险驾驶罪二审刑事裁定书	呼和浩特市中级人民法院刑事裁定书(2019)内01刑终178号	原公诉机关呼和浩特市玉										
11	吕沁波与张东方生产、销售假药罪	(2019)晋05裁定书	2019.08.21	吕沁波与张东方生产、销售假药罪二审刑事裁定书	山西省晋城市中级人民法院刑事裁定书(2019)晋05刑终224号	原公诉机关										
12	王某某贪污罪	(2019)晋08裁定书	2019.08.21	王某某贪污罪、职务侵占罪郑某某贪污罪二审刑事裁定书	山西省运城市中级人民法院刑事裁定书(2019)晋08刑终367号	原公诉机关										
13	于会来寻衅滋事罪	(2019)吉01判决书	2019.08.20	于会来寻衅滋事罪二审刑事判决书	吉林省长春市中级人民法院刑事判决书(2019)吉01刑终200号	原公诉机关吉林省德惠										
14	王某1、付某1交通肇事罪	(2019)京11裁定书	2019.08.20	王某1、付某1交通肇事罪二审刑事裁定书	北京市高级人民法院刑事附带民事裁定书(2019)京11刑终96号	上诉人暨被										

图2 数据存储表格(部分)

3.2 数据预处理

(1) 信息抽取

在裁判文书全文中,大多数文书都具有相似的结构。正则表达式是一种工具,由特定意义的字符组成,表示某种匹配的规则^[24]。其能按照设定的表达式从文本数据或网络页面信息中提取部分特定信息,因此本文利用正则表达

式,基于爬取到的刑事二审裁判文书内容结构,编写代码匹配相应关键词,提取刑事案件二审裁判文书全文中的上诉人身份、上诉理由全文和二审审判结果。

(2) 分词

为了理想的分词结果,本文首先利用清华大学开放中文词库中的法律词表对上诉理由文

本进行了分词,该词表来主流网站的社会标签、搜索热词、输入法词库等,包含词频统计信息DF值(Document Frequency),经过多轮人工筛选,且一直不断更新。然后依据上面的分词结果,利用jieba分词自定义词典的方式构造了本文的自定义词典,对一些上诉理由的复合词进行添加,然后再对所有的裁判文书中的上诉理由进行了分词。

(3) 上诉理由的词性选择

本文首先利用word2vec模型对上诉理由的分词结果进行了词向量模型的聚类,然后利用jieba分词将上诉理由文本中不同词性的词进行分类并做了词频统计。为了找到能真正代表上诉理由的词性,本文对不同词性词频前10的词做了统计(表1),发现上诉理由可以很清晰的被未知词(x)代表,而其他词性的词并不能显示出上诉理由。因此,本文基于统计结果从未知词提取了上诉理由。

表1 上诉理由词性分布图(部分)

上诉理由名词(n)	上诉理由形容词(a)	上诉理由动词(v)	上诉理由未知词(x)
上诉人	准确	提出	量刑过重
原审	合同	上诉	上诉理由
被告人	明显	认定	从轻处罚
原判	严重	请求	申请撤回上诉
辩护人	伤害	没有	不构成
法院	具体	不服	事实不清
本院	重大	行为	适用缓刑
意见	清楚	辩护	减轻处罚
被害人	虚假	判决	依法改判
证据	重新	认为	自首情节

(4) 基于word2vec训练词向量

传统的聚类方法通常的文本表示方法是空间向量模型,文本中的词组成特征向量,特征向量构成空间向量模型,通过计算空间向量模型中向量之间的距离来评估文本或词之间的相似性,距离越近,文本/词之间越相似。但若两个文本或词字面表达不一样,但意义相似,即使它们的特征词不相同,它们也是相似的。传统的方法并不能解决一词多义、一义多词等现象。

裁判文书文本中由于不同法院、不同撰稿人有不同的表述用语习惯,特别是在上诉理由中,涉及量刑等的词存在着大量意思相近而表述不同的词,如“量刑过重”和“量刑畸重”等,为了更好的对上诉理由进行提取与分类,解决传统聚类方法中一义多词不能识别的问题,更好地对上诉理由进行分类,本文使用了一种基于word2vec^[25]的文本/词聚类方法,word2vec可以基于上下文训练文本的词向量,从而识别相似词,给予它们相近的词向量,然后再利用传统的聚类算法计算向量空间上的相似度就可以将词/词组分类。

本文首先利用word2vec将上诉理由文本中的词表示成词向量的形式,训练词向量模型,word2vec中的CBOW模型能够基于上下文给出词的词向量,这利于表述不一样但意思相近的词聚类。因此,为了进一步分析刑事案件二审上诉原因构成,本文使用word2vec中的CBOW模型对刑事二审裁判文书中上诉理由分词后的数据训练词向量,保存词向量模型,以便于后面的利用。

(5) 基于词向量模型进行聚类

本文选取了使用最广泛的 K-means 算法来对上诉理由进行聚类。在使用聚类方法时,选择合适的聚类数目往往是关键的一步,这会直接影响聚类的效果,对应到 k-means 算法中,就是 k 值的选取,本文使用了目前常见的手肘法。手肘法可以通过对聚类文本误差平方和 (SSE) 的计算,来确定 k-means 聚类的 k 值^[26, 27],其实现简单,处理大量数据的效率高。

本文选择了最能代表上诉理由的词性,确定该词性的词的合集作为聚类的文本,然后基于训练好的词向量模型用 k-means 算法对该文本进行聚类。并利用轮廓系数法^[25]对结果进行了检验。

4 刑事案件二审裁判文书文本挖掘结果

4.1 相关因素分析

(1) 文书类型

从文书类型(表2)可以看出,刑事案件二审裁判文书基本由裁定书和判决书两种类型的文书构成。其中,裁定书占的比例最大,达到了文书总量的 90.5% 以上,判决书占文书总量的比例也接近 9.5%。

表 2 刑事二审裁判文书类型

文书类型	数量	占比
裁定书	25639	90.50%
判决书	2682	9.47%
调解书	5	0.02%
其他文书	3	0.01%
总数	28329	100%

(2) 二审审判结果

为了更好的突出二审的审判情况,本文将“维持原判”、“撤回上诉”、“撤回抗诉”和“双方调解”这些原审审判结果未发生改变的称为“保持原判”。而“撤销”、“判处”和“重新审判”这些原审判决发生变化的称为“重新审判+改判”。从结果(表3)中可以看出,刑事案件二审保持原判的数量占了大多数,为 84.53%,而改判的只有 15.47%。

表 3 审判结果总体分布

审判结果	数量	占比
保持原判	23324	84.53%
重新审判+改判	4269	15.47%

(3) 案由

从词频统计的结果(图3)所示,刑事二审涉及的案由最多的为“盗窃”,有 5409 件,约占裁判文书总数的 19.09%;其次是“故意伤害”,有 3387 件,约占裁判文书总数的 11.20%;此外,“诈骗”、“寻衅滋事”、“走私贩卖运输制造毒品”等都占有一定的比率,说明这些案由刑事一审后上诉的可能性大。后面的案由数量有所差别但差别不大,并且案件数量不算少,说明上诉涉及的案由多种多样。

(4) 上诉人身份

从结果(图4)可以看出,上诉人身份的大部分集中于“原审被告”、“原审附带民事诉讼原告人”、“原审附带民事诉讼被告人”等身份。其中,“原审被告”的占比达到了 92%。

盗窃 5409
故意伤害 3387
诈骗 2860
寻衅滋事 2527
走私贩卖运输制造毒品 2057
危险驾驶 1560
交通肇事 1417
抢劫 1192
开设赌场 1051
容留他人吸毒 879
敲诈勒索 765
非法吸收公众存款 692
非法拘禁 652
聚众斗殴 632
合同诈骗 574
妨害公务 491
非法持有毒品 454
强奸 444
非法经营 367
贪污 356
故意杀人 356
受贿 341
故意毁坏财物 328
掩饰隐瞒犯罪所得收益 324
赌博 322
职务侵占 298
拒不执行判决裁定 255
组织领导传销活动 251
组织卖淫 202
信用卡诈骗 198

图3 案由词频统计结果前 30

原审被告人 44719
原审附带民事诉讼原告人 2248
原审附带民事诉讼被告人 672
原审自诉人 496
一审被告人 350
一审自诉人 178
附带民事诉讼原告人 124
原审附带民事诉讼被告 77
原审被告单位 67
原审自诉人暨附带民事诉讼原告人 62
附带民事诉讼被告人 52
原审刑事附带民事诉讼原告人 46
被告人 34
原审附带民事诉讼被告单位 25
原审附带民事诉讼原告 18
附带民事诉讼被告 18
原审被告 12
原审附带民事诉讼原告人 11
被告 9
原审刑事附带民事诉讼被告人 8

图4 上诉身份词频统计前 20

4.2 上诉理由分析

从未知词词频统计结果（图5）所示，“量刑过重”一共出现了13819次，出现次数比后面的都多出许多，在前30的词频中，还出现了“量刑过重”、“量刑畸重”、“量刑偏重”等近义词，可见“量刑过重”是最主要的上诉原因。此外，“从轻处罚”、“不构成”、“事实不清”、

“适用缓刑”等词出现频率也较高（这里需要特殊说明的是，“不构成”在上诉理由文本中一般后接罪名，表明上诉人不构成**罪，因为涉及罪名过多，本文为了统计的方便，在自定义词典中只添加了“不构成”这个词），说明上诉的理由构成要素众多，存在于方方面面。但显然，词频统计虽然反映了上诉理由的分布情况，但分布的情况太散，不利于规律的总结，从而也得出普适性的结论。因此，为了更好的反映上诉理由的分布，本文对上诉理由的未知词基于训练好的词向量模型进行了聚类，以期发现它们之间存在的关联。

量刑过重,13819
从轻处罚,6845
申请撤回上诉,6571
不构成,3958
事实不清,3591
适用缓刑,3187
减轻处罚,3035
依法改判,2626
自首情节,2444
量刑适当,1980
认罪悔罪,1332
量刑重,1261
认罪态度好,1238
适用法律错误,1231
审判程序合法,986
认定事实错误,882
请求改判,851
积极赔偿,832
量刑畸重,724
不构成犯罪,717
上诉意见,695
量刑偏重,676
适用法律正确,665
定罪准确,569
程序违法,550
改判缓刑,537
立功表现,505
从轻判处,476
判处缓刑,467
量刑畸轻,459

图5 上诉理由未知词词频统计前 30

在利用训练好的词向量模型对上诉理由未知词文本进行聚类的时候，为了更好的聚类效果，本文先用手肘法基于词向量模型计算了对

未知词文本聚类的k值(图6),拐点在k=2的位置。因此,将上诉理由未知词聚成了两类。

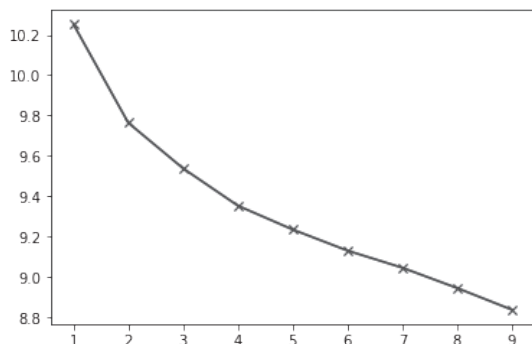


图6 K-means 手肘法拐点图

['量刑过重', '从轻处罚', '申请撤回上诉', '不构成', '事实不清', '减轻处罚', '认罪悔罪', '认罪态度好', '适用法律错误', '审判程序合法', '认定事实错误', '请求改判', '不构成犯罪', '上诉意见', '量刑偏重', '适用法律正确', '定罪准确', '程序违法', '改判缓刑', '立功表现', '判处缓刑', '量刑畸轻', '缴纳罚金', '犯罪未遂', '定性错误', '不是主犯', '代理意见', '事实错误', '适用法律不当', '事实认定错误', '认定事实有误', '情节轻微', '量刑过轻', '服从原审判决', '认罪悔罪态度好', '提交新证据', '程序严重违法', '量刑错误', '法定代理人', '程序错误', '无法法律依据', '赔偿数额过高', '适用非监禁刑', '适用不当', '依据错误', '使用缓刑']
['适用缓刑', '依法改判', '自首情节', '量刑适当', '量刑重', '积极赔偿', '量刑畸重', '从轻判处', '次要作用', '认罪态度较好', '犯罪情节轻微', '犯罪情节较轻', '支持抗辩', '积极配合', '危害性小', '愿意赔偿', '免于刑事处罚', '服从原判', '情节较轻', '委托辩护人', '事实有误', '定罪错误', '态度良好', '较轻的刑罚', '危害小', '危害较小', '刑罚过重', '不应构成犯罪', '事实不当', '定罪不准确', '行为错误', '认罪悔罪']

图7 上诉理由聚类结果

传统的刑事二审案件上诉人的上诉理由一般包括四个方面^[14]:

- (1) 认定事实错误;
- (2) 适用法律错误
- (3) 量刑过重
- (4) 违反诉讼程序

从聚类的结果中可以看出,类别一中的大部分词都与量刑、事实、适用法律、诉讼程序等有关,如“量刑过重”、“事实不清”、“适用法律错误”和“程序违法”等,这些词与传统的二审案件上诉理由基本是一致的,表明本文所用的提取上诉理由的方法是可行的。而类别二中的还存在一些词与传统二审案件的上诉理由不一样,如“自首情节”、“积极赔偿”、“次要作用”、“危害小”等,这些词虽然有别于传统的上诉理由,但经过查证法律界相关

如上诉理由的聚类结果(图7)所示,上诉理由可以分为两类:第一个类大多数是上诉人对法院的审判结果的不满,如“量刑过重”、“适用法律错误”、“事实不清”、“程序错误”等;第二个类则是上诉人自身情况的陈述,如“自首情节”、“危害小”、“积极赔偿”、“态度良好”等。虽然总体都是对审判结果的不满意,但一方面是从法院审判结果的角度出发,另一方面是从陈述自身态度的角度出发。

专业文章,犯罪后的态度如“自首”、“悔罪”、“积极赔偿”等都会对量刑产生一定的影响,在案件审判时当事人也会以此为理由为自己减轻刑期^[29],可以说这些词也能代表当事人的一部分上诉理由,目前,相关制度如认罪认罚从宽制度等的研究也有较高的热度^[30],但这些词在传统的上诉理由中是并没有具体说明的,而本文所使用的办法将这部分上诉理由挖掘了出来,表明了本文方法不但能够挖掘出传统的上诉理由,还能挖掘出其他有别于传统上诉理由的而且同样重要的上诉理由,这进一步证明了本文方法的可行性。

然后,本文使用轮廓系数法对聚类效果进行了评估,其综合考虑了类中聚集程度和类间离散程度,取值 $S(i)$ 介于 $[-1,1]$ 之间,值越靠近1,表明聚类效果越好。所有数据的 $S(i)$ 均值被称

为聚类结果的轮廓系数。具体公式如下:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

如表4所示,基于word2vec的k-means聚类方法比传统的k-means聚类方法结果要好。

表4 聚类效果 $S(i)$ 值对比

方法	$S(i)$ 值
基于word2vec的k-means聚类方法	0.16446176
传统的k-means聚类方法	0.04718446606648576

本文能获取新发现的基于上诉人态度的这部分上诉理由,是由于上诉理由包含一系列复合词。为了获取这些复合词,本文采取了自定义词典的方式,在自定义词典的过程中,为了更好更准确地将这些词提取出来,研读了部分上诉理由的文本,将上诉理由文本中上诉人陈

述的主要信息中筛选了重复出现次数多、有代表性的复合词添加到自定义词典进行分词,从而得到了这部分新发现的上诉理由。从词频统计的结果也可以看出,上诉人自身情况的陈述出现的次数较多,是上诉理由的重要组成部分。

4.3 重新审判和改判案件上诉理由分析

本文基于同样方法对重新审判和改判案件的上诉理由进行了分析。表5是全部文书与重新审判和改判文书相关信息统计对比,如表5所示,在案由、上诉人身份和上诉理由统计中,全部二审案件与重新审判和改判案件的上诉理由前10名有大多数是相同的,且上诉理由的词与上文传统的上诉理由和文献中涉及的词是高度重合的,这进一步表明了本文思路和方法的可行性。

表5 全部文书与重新审判和改判文书相关信息对比

案由统计(前10)		上诉人身份统计(前10)		上诉理由统计(前10)	
全部二审案件	重新审判和改判案件	全部二审案件	重新审判和改判案件	全部二审案件	重新审判和改判案件
盗窃	诈骗	原审被告入	原审被告入	量刑过重	量刑过重
故意伤害	故意伤害	原审附带民事诉讼原告人	原审附带民事诉讼原告人	从轻处罚	从轻处罚
诈骗	交通肇事	原审附带民事诉讼被告人	原审附带民事诉讼被告人	申请撤回上诉	不构成犯罪
寻衅滋事	盗窃	原审自诉人	原审被告单位	不构成犯罪	事实不清
走私贩卖运输制造毒品	寻衅滋事	原审被告单位	原审自诉人	事实不清	自首情节
危险驾驶	走私贩卖运输制造毒品	原审自诉人暨附带民事诉讼原告人	原审刑事附带民事诉讼原告人	适用缓刑	减轻处罚
交通肇事	危险驾驶	原审刑事附带民事诉讼原告人	原审附带民事诉讼被告单位	减轻处罚	依法改判
抢劫	非法吸收公众存款	原审附带民事诉讼被告单位	原审附带民事诉讼被告单位	依法改判	适用法律错误
开设赌场	合同诈骗	原审附带民事诉讼原告人	原审自诉人暨附带民事诉讼原告人	自首情节	量刑适当
盗窃	诈骗	原审附带民事诉讼被告人	原审刑事附带民事诉讼被告人	量刑适当	认罪悔罪

5 讨论与总结

本文从互联网相关裁判文书网站上获取刑事案件二审裁判文书全文和相关信息,并基于裁判文书的全文对刑事案件二审的上诉人身份、上诉理由和审判结果进行了抽取,对文书类型、二审审判结果、上诉人身份和案由等相关信息进行了分析,然后对上诉理由进行了提取和分类,并进一步将重新审判和改判案件的上诉理由和全部裁判文书的上诉理由进行了对比分析。

结果表明,利用词性提取出上诉理由关键信息是可行的,利用 word2vec 的 k-means 聚类方法对上诉理由进行聚类时效果好于传统方法。同时,实验证明,刑事二审的案由中,“盗窃”、“故意伤害”、“诈骗”、“寻衅滋事”和“走私贩卖运输制造毒品”等案由出现的次数最多。刑事案件二审的上诉人身份中,“原审被告”、“原审附带民事诉讼原告人”、“原审附带民事诉讼被告人”等上诉人身份出现的次数最多。刑事案件二审的上诉理由中,“量刑过重”、“从轻处罚”、“不构成犯罪”、“事实不清”和“适用缓刑”等理由出现的次数最多,并且上诉理由主要从上诉人对于法院判决不满和上诉人自身情况和态度两个方面来陈述。重新审判和改判的案件占全部案件的比率较低,并且重新审判和改判案件与全部二审案件的上诉理由大多数相同。结果也证明了不仅可以从文本挖掘的角度来分析裁判文书,也可以利用该方法从不同方面对刑事案件二审的上诉原因进行挖掘。文本挖掘的结果也可以为一审审判、智慧量刑等提供更多的决策参考,为法务工作者提供更好的信息服务。

由于不同地区、不同法院的文书格式并不固定,内容表述不详尽,用语上存在一定的差异,并且还存在错别字等情况,导致文本提取相关信息时存在一定阻碍和困难,由此,分析结果可能会产生一定误差。希望未来在裁判文书获取权限更开放,数据更规整的情况下做出更好、更深入的研究。

参考文献

- [1] 左卫民. 关于法律人工智能在中国运用前景的若干思考[J]. 清华法学, 2018, 12, 66(2):109-125.
- [2] 李婷. 司法改革背景下刑事二审庭审实质化问题思考——以刑事二审审判权的运行实践为出发点[J]. 法律适用, 2016(7):101-106.
- [3] 左卫民, 王婵媛. 基于裁判文书网的大数据法律研究: 反思与前瞻[J]. 华东政法大学学报, 2020, 23(2):64-76.
- [4] 马超, 于晓虹, 何海波. 大数据分析: 中国司法裁判文书上网公开报告[J]. 中国法律评论, 2016(4): 195-246.
- [5] 郭叶, 孙妹. 指导性案例应用大数据分析——最高人民法院指导性案例司法应用年度报告(2016)[J]. 中国应用法学, 2017(4):40-62.
- [6] 张忠民. 生态破坏的司法救济——基于 5792 份环境裁判文书样本的分析[J]. 法学, 2016(10):111-124.
- [7] 黄启辉. 行政诉讼一审审判状况研究——基于对 40 家法院 2767 份裁判文书的统计分析[J]. 清华法学, 2013, 7(4):73-85.
- [8] 朱春华. 行政诉讼二审审判状况研究——基于对 8 家法院 3980 份裁判文书的统计分析[J]. 清华法学, 2013(4):87-103.
- [9] Callander M, Page S J. Managing risk in adventure tourism operations in New Zealand: A review of the legal case history and potential for litigation[J]. Tourism Management, 2003, 24(1):13-23.
- [10] Christopher B - Lynch, Coker A, Dua J A. A clinical analysis of 500 medico - legal claims evaluating

the causes and assessing the potential benefit of alternative dispute resolution[J]. BJOG: An International Journal of Obstetrics & Gynaecology, 1996, 103(12):1236-1242.

- [11] Chandler J A. The use of neuroscientific evidence in Canadian criminal proceedings[J]. Journal of Law & the Biosciences, 2015, 2(3):550-579.
- [12] 程荣斌, 邓云. 审级制度研究 [J]. 湖南省政法管理干部学院学报, 2001, 17(5):4-12.
- [13] 孙远. 论刑事上诉审构造 [J]. 法学家, 2012(4):128-145.
- [14] 李昌林, 王丹. 刑事上诉理由研究 [J]. 中国刑事法杂志, 2011(7):71-77.
- [15] 杨杰辉. 刑事上诉理由研究 [J]. 法治研究, 2017(6):130-137.
- [16] Feldman R, Dagan I. Knowledge Discovery in Textual Databases (KDT)[C]. KDD. 1995, 95: 112-117.
- [17] 鹿小明. 文本挖掘及其在信息检索中的应用 [J]. 情报资料工作, 2004(6):26-28.
- [18] 汤铭. 基于领域知识库的校园智能问答系统关键技术研究 [D]. 南京: 东南大学, 2018.
- [19] 王景峰. 基于层次多标签分类的适用法律识别 [D]. 南京: 南京大学, 2016.
- [20] 王海亮. 基于文本挖掘的法律咨询系统研究与实现 [D]. 北京: 北京化工大学, 2017.
- [21] 梁楠. 基于文本挖掘的律师推荐方法研究与应用 [D]. 成都: 电子科技大学, 2015.
- [22] 高雅婷. 基于刑事案件要素的文本挖掘系统研究与实现 [D]. 西安: 长安大学, 2019.
- [23] 舒洪水. 司法大数据文本挖掘与量刑预测模型的研究 [J]. 法学, 2020(7):113-129.
- [24] 沙金. 精通正则表达式 [M]. 北京: 人民邮电出版社, 2008.
- [25] Tomas Mikolov. Word2vec project [EB/OL]. [2014-09-18]. <https://code.google.com/p/word2vec/>.
- [26] 成卫青, 卢艳红. 一种基于最大最小距离和 SSE 的自适应聚类算法 [J]. 南京邮电大学学报 (自然科学版), 2015, 35 (2): 102-107.
- [27] 王建新, 王柏人, 曲鸣, 等. 基于改进欧式距离的硬件木马检测 [J]. 计算机工程, 2017, 43(6):92-96.
- [28] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics, 1987(20):53-65.
- [29] 张明楷. 论犯罪后的态度对量刑的影响 [J]. 法学杂志, 2015(2): 1-10.
- [30] 熊秋红. 认罪认罚从宽的理论审视与制度完善 [J]. 法学, 2016(10): 97-110.