南京大学
NANJING UNIVERSITY

研 究 生 毕 业 论 文

（申请硕士专业学位）

论 文 题 目	基于领域特征的文本数据扩增技术
作 者 姓 名	李卓阳
专业学位类别（领域）	工程硕士（软件工程领域）
研 究 方 向	软件工程
指 导 教 师	刘嘉 副教授

2021 年 5 月 20 日

学　　号：MF1932107

论文答辩日期：2021 年 5 月 20 日

指 导 教 师：  　　（签字）

Text Data Augmentation Technique Based on Field Features

by

Zhuoyang li

Supervised by

Associate Professor **Jia Liu**

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
Master of Engineering
in
Software Engineering

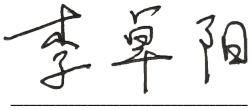
Software Institute
Nanjing University

May 20, 2021

学 位 论 文 原 创 性 声 明

任何收存和保管本论文的单位和个人，未经作者本人授权，不得将本论文转借他人并复印、抄录、拍照或以任何方式传播，否则，引起有碍作者著作权益的问题，将可能承担法律责任。

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不句含其他个人或集体已经发表或撰写的作品成果。本文所引用的重要文献，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 

日期： 2021 年 5 月 20 日

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于领域特征的文本数据扩增技术

工程硕士（软件工程领域）专业 2019 级硕士生姓名：李卓阳

指导教师（姓名、职称）：刘嘉 副教授

摘　要

随着深度神经网络技术的发展，基于领域文本数据集训练得到的深度神经网络模型逐渐应用到社会各个领域，用来解决各个领域中的实际问题。深度学习模型的构建需要大规模、高质量领域文本数据作为训练集。在实践中，领域文本获取成本高等原因会造成缺乏训练数据、样本分布不均衡等问题，并会导致深度学习模型的泛化能力较差。数据扩增是一种可以提高训练集大小的技术。

目前，常用文本数据扩增技术处理文本数据时易影响体现文本领域特征的词语及语义结构信息，导致扩增后文本质量差，对于模型泛化能力的提高作用有限。鉴于此，本文以司法领域数据集为例，设计并实现了基于领域特征的文本数据扩增技术，包括对领域文本数据集的预处理步骤和四种特征扩增方法。

数据集预处理是为后续基于领域特征的文本数据扩增提供支撑。基于TF-IDF 权重的特征裁剪扩增方法是以文本分词在数据集中的 TF-IDF 值为依据，结合依存句法分析技术进行剪枝操作；基于主题模型的特征融合扩增方法是使用主题模型技术聚类数据集中相似文本，将待扩增文本与相似目标文本进行内容交换；基于依存句法的特征变换扩增方法是使用依存句法分析技术解构文本，将句法树中依存关系相同的树枝进行交换；基于词频词性的特征替换方法是基于领域数据集分析构建高频词表和词向量模型，将文本中符合高频词和相关词性的词语使用词向量模型推荐领域近义词进行替换。

本文通过设计对比实验，在司法数据集上构建高质量文本分类模型，将特征扩增文本和 EDA 扩增文本作为测试集，实验表明特征扩增文本在保持类别标签方面表现较好，有效地保持了文本的领域特征。其次，在司法和媒体领域原始数据训练集中加入使用特征扩增方法和 EDA 方法扩增的数据，相比于原始数据训练的 CNN 和 RNN 模型，加入扩增数据后的模型准确率提升。总体而

言，加入特征扩增文本的模型比加入 EDA 扩增文本的模型在测试集上的准确率提高幅度更大。实验表明，基于领域特征的文本数据扩增技术具有一定的实用性和有效性。

关键词： 文本数据扩增，自然语言处理，TF-IDF 算法，主题模型，领域特征

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Text Data Augmentation Technique Based on Field Features

SPECIALIZATION: Software Engineering
POSTGRADUATE: Zhuoyang li
MENTOR: Associate Professor **Jia Liu**

Abstract

With the development of deep neural network technology, deep neural network models trained based on field text data sets are gradually applied to various fields of society to solve practical problems in various fields. The construction of deep learning models requires large-scale, high-quality field text data as a training set. In practice, reasons such as the high cost of acquiring domain text will cause problems such as lack of training data, unbalanced sample distribution, and will lead to poor generalization capabilities of deep learning models. Data augmentation is a technique that can increase the size of the training set.

At present, the commonly used text data augmentation technology is easy to affect the words and semantic structure information that reflect the features of the text field when processing text data, resulting in poor text quality after augmentation, and its effect on improving the generalization ability of the model is limited. Given this, this paper takes the judicial field data set as an example, designs and implements a text data augmentation technology based on field features, including preprocessing steps for the field text data set and four feature augmentation methods.

Data set preprocessing is to provide support for the subsequent augmentation of text data based on field features. The feature pruning and augmentation method based on TF-IDF weight is based on the TF-IDF value of the text segmentation in the data set, combined with dependency syntax analysis technology for pruning operation; the feature fusion augmentation method based on the topic model is to use topic model technology Clustering similar texts in the data set, exchange the content of the text to be augmented with similar target texts; the feature transformation augmentation method

based on dependency syntax uses dependency syntax analysis technology to deconstruct the text, and exchange branches with the same dependency relationship in the syntax tree; The feature replacement method based on word frequency and part-of-speech is to construct a high-frequency vocabulary and word vector model based on field data set analysis and replace words that meet high-frequency words and related parts of speech in the text using the word vector model to recommend field synonyms.

In this thesis, a comparative experiment is designed to build a high-quality text classification model on the judicial data set. The feature augmented text and EDA augmented text are used as the test set. The experiment shows that the feature augmented text performs well in maintaining the category label, and effectively maintains The field features of the text. Secondly, adding data augmented using feature augmentation methods and EDA methods to the original data training set of the judicial and media fields, compared with the CNN and RNN models trained on the original data, the accuracy of the model after adding the augmented data is improved. In general, the model with feature augmented text has a greater improvement in accuracy on the test set than the model with EDA augmented text. Experiments show that the text data augmentation technology based on field features has certain practicability and effectiveness.

keywords: Text data augmentation, Natural language processing, TF-IDF algorithm, Topic model, Field features

目　　录

目　　录 ····· **v**

图目录 ····· **vii**

表目录 ····· **ix**

第一章　绪论 ····· **1**

1.1　选题背景与意义 ····· 1

1.2　国内外研究现状及分析 ····· 2

1.2.1　文本数据扩增研究现状 ····· 2

1.2.2　文本领域特征提取研究现状 ····· 5

1.3　本文主要研究工作 ····· 6

1.4　本文的组织结构 ····· 7

第二章　技术综述 ····· **9**

2.1　自然语言处理技术 ····· 9

2.1.1　中文文本分词 ····· 9

2.1.2　词性标注 ····· 10

2.1.3　依存句法分析 ····· 11

2.1.4　词向量 ····· 13

2.2　LDA 主题模型 ····· 14

2.3　TF-IDF 算法 ····· 16

2.4　本章小结 ····· 17

第三章　基于领域特征的扩增方法 ····· **19**

3.1　文本预处理 ····· 19

3.2　基于 TF-IDF 权重的特征裁剪 ····· 22

3.3　基于主题模型的特征融合 ····· 28

3.4　基于依存句法的特征变换 ····· 40

3.5　基于词频词性的特征替换 ····· 44

3.6　本章小结 ····· 48

第四章 基于领域特征的扩增数据生成及实验评估 ····· 51
4.1 扩增数据标签评估 ····· 51
4.2 扩增实验数据准备 ····· 55
4.3 训练数据实验评估 ····· 57
4.4 本章小结 ····· 65
第五章 总结与展望 ····· 67
5.1 总结 ····· 67
5.2 进一步工作 ····· 68
参考文献 ····· 69
致 谢 ····· 75
《学位论文出版授权书》 ····· 77

图目录

2-1 依存句法树示例图 ····· 13
2-2 CBOW 和 Skip-gram 模型结构图 ····· 14
2-3 LDA 模型图 ····· 15

3-1 司法裁判文书数据集所涉罪名类型分布图 ····· 20
3-2 文本预处理分词操作核心代码 ····· 20
3-3 司法裁判文书数据集高频词语词云图 ····· 21
3-4 基于 TF-IDF 权重的特征裁剪扩增方法流程图 ····· 22
3-5 TF-IDF 模型构建核心代码 ····· 23
3-6 特征裁剪扩增方法待扩增示例文本依存句法树 ····· 24
3-7 待扩增文本分词词语 TF-IDF 赋值计算核心代码 ····· 25
3-8 生成待裁剪树枝列表核心代码 ····· 26
3-9 特征裁剪扩增方法待扩增示例文本扩增结果依存句法树 ····· 28
3-10 基于主题模型的特征融合扩增方法流程图 ····· 29
3-11 司法数据集主题模型在不同 K 值下的模型困惑度曲线图 ····· 30
3-12 LDA 主题模型构建核心代码 ····· 31
3-13 主题模型的关键词统计图 ····· 31
3-14 基于 MDS 的主题距离分布图 ····· 32
3-15 主题模型 6 号主题的关键词分布图 ····· 33
3-16 待扩增文本的主题模型预测结果条形图 ····· 34
3-17 基于主题模型的相似文本筛选核心代码 ····· 36
3-18 依存句法树树枝结点层次遍历核心代码 ····· 38
3-19 特征融合扩增方法待扩增示例文本扩增结果依存句法树 ····· 39
3-20 基于依存句法的特征变换扩增方法流程图 ····· 40
3-21 基于依存句法的特征变换待扩增示例文本依存句法树 ····· 41
3-22 依存句法树树枝筛选核心代码 ····· 41
3-23 依存句法树包含关系树枝合并核心代码 ····· 42

3-24 获取包含不同结点树枝配对元组核心代码 ····· 43
3-25 特征变换扩增方法待扩增示例文本扩增结果依存句法树 ····· 43
3-26 基于词频词性的特征替换扩增方法流程图 ····· 45
3-27 司法裁判文书数据集词频统计图 ····· 45
3-28 词向量训练核心代码 ····· 46
3-29 基于词频词性的特征替换方法待扩增示例文本依存句法树 ····· 47

4-1 模型在测试集 testFeature 和测试集 testEDA 上的精确率 ····· 53
4-2 分类模型在测试集 testFeature 和测试集 testEDA 上的召回率 ····· 53
4-3 模型在测试集 testFeature 和测试集 testEDA 上的 F1 值 ····· 54
4-4 基于领域特征的文本数据扩增方法一般步骤 ····· 55
4-5 媒体领域数据集词频统计图 ····· 56
4-6 不同 K 值下的主题模型困惑度曲线图 ····· 56
4-7 司法领域特征扩增与 EDA 扩增数据的 CNN 模型准确率结果图 ···· 61
4-8 司法领域特征扩增与 EDA 扩增数据的 RNN 模型准确率结果图 ···· 62
4-9 媒体领域特征扩增与 EDA 扩增数据的 CNN 模型准确率结果图 ···· 63
4-10 媒体领域特征扩增与 EDA 扩增数据的 RNN 模型准确率结果图 ···· 63

表目录

1-1 EDA 方法 …… 3

2-1 部分词性标注编码表 …… 11
2-2 依存句法分析部分标签说明表 …… 11

3-1 文本预处理的标准结构化文本格式表 …… 19
3-2 基于 TF-IDF 权重的特征裁剪扩增方法参数表 …… 27
3-3 特征裁剪扩增方法待扩增示例文本扩增结果对照表 …… 27
3-4 基于主题模型的相似文本推荐表 …… 35
3-5 基于主题模型的特征融合扩增方法参数表 …… 38
3-6 特征融合扩增方法待扩增示例文本扩增结果对照表 …… 39
3-7 基于依存句法的特征变换扩增方法参数表 …… 42
3-8 特征变换扩增方法待扩增示例文本扩增结果对照表 …… 44
3-9 特征替换扩增方法词性选择表 …… 47
3-10 基于词频词性的特征替换扩增方法参数表 …… 48
3-11 特征替换扩增方法待扩增示例文本扩增结果对照表 …… 48
3-12 基于领域特征扩增方法介绍简表 …… 49

4-1 司法领域分类模型评估指标结果表 …… 52
4-2 司法领域数据集基于领域特征扩增结果表 …… 57
4-3 司法领域数据集 EDA 扩增结果表 …… 58
4-4 媒体领域数据集基于领域特征扩增结果表 …… 59
4-5 媒体领域数据集 EDA 扩增结果表 …… 59
4-6 模型训练集数据表 …… 60
4-7 领域原始基础数据集模型预测准确率结果表 …… 64

第一章　绪论

1.1　选题背景与意义

随着人工智能技术的快速发展，深度学习作为新兴技术为解决现实领域中的问题提供了全新的方法，并取得了显著成果。深度学习技术的实现是建立在大规模高质量数据训练得到模型的基础上的，模型性能的评估也依靠广泛而具有代表性的大量数据。数据质量的高低和规模决定了 AI 系统是否能够更好地满足人们的现实需求。

一般来讲，监督学习作为一种模型训练方式其目的很明确，标签的存在可以衡量最终训练效果，监督学习模型相对半监督或无监督学习模型可以对模型进行有效评估。但是，监督模型需要大量的标注数据，在构建深度学习模型的过程中，由于现实领域中的客观条件和人们的主观差异，在数据的选择方面普遍存在数据质量参差不齐、数据获得成本高、数据可用量较少等问题。在典型的分类任务中，这些问题的存在会导致深度神经网络模型出现过拟合现象，严重影响模型的分类效果。然而，很多领域获取大量高质量数据的成本较高，如司法领域的司法裁判文书，媒体领域新闻文本等。如何利用有限的标注数据生成更多的数据来减少神经网络模型中的过拟合现像，进而训练出拥有更强泛化能力的模型成为一个亟需解决的现实问题。数据扩增技术提供了一套解决该问题的方案。

数据扩展是通过人工扩展数据集，从有限的数据生成更多的等价数据的技术。该方法能够提高训练数据集的规模和质量，是解决训练数据不足的有效方法 [1]。目前在各种深度学习领域得到了广泛的应用。在计算视觉领域中，图像扩增相对比较容易，有一系列简单有效的方法可供选择，如几何变换、颜色变换、旋转裁剪等，这些方法已经被证明有很好的效果，被封装成机器学习库来方便进行计算视觉领域中的图像数据扩增 [2]。然而，自然语言是离散的抽象符号，任何微小的变化都可能会导致语义的偏差，因此，图像领域中的数据扩增技术不能简单地迁移到自然语言处理领域。同时，在实践中开发基于文本的深度学习技术时，往往会遇到文本数据量不足、不同类别的文本数据量差别较

大、数据标注成本较高等问题。因此，研究文本数据扩增技术对解决基于文本的深度学习问题具有重要意义。

当前，国内外在文本数据扩增领域提出了多种扩增方法，如回译 [3]、简单数据扩增技术（EDA）[4]、基于 GAN 网络的扩增 [5]、情景增强 [6] 和无监督数据扩增 [7] 等，这些广泛应用的方法在降低数据获取成本，抑制过拟合，提高模型泛化能力发挥了重要作用。然而，这些方法大都是对文本进行单句字符级别的处理，本质上对文本字词的删除、替换和位置交换。在进行文本分类的任务中，这些对文本字符级别的处理方法易影响体现文本领域特征的词语以及体现领域特征的语义结构信息，导致扩增后的文本不能很好地体现其所在领域特征，扩增文本质量较低，对文本分类模型性能的提升效果不明显。因此，基于领域特征的文本数据扩增技术对于扩增后文本体现领域特征、获得高质量扩增文本并提高文本分类效果具有重要意义。

综上所述，基于文本的深度学习算法对高质量数据有较大需求。本文将吸取传统文本数据扩增技术的优点，基于已有的有限数据集，使用 TF-IDF 算法、词向量、依存句法分析、主题模型等技术充分挖掘数据集的领域特征，提出一种简便的基于领域特征的扩增技术，提高深度学习文本分类模型的性能。该技术为深度学习模型开发者提供更广泛的高质量数据集，辅助模型开发者提高文本分类模型性能，具有较强的实用意义。

1.2 国内外研究现状及分析

基于领域特征的文本数据扩增技术主要研究内容有两个部分，分别为文本数据扩增技术研究和文本领域特征提取研究。为了充分了解目前这两个内容的国内外研究现状，为技术的研究提供理论和实践支撑，分别调研了目前对于文本数据扩增的研究进展和领域特征抽取的相关研究进展

1.2.1 文本数据扩增研究现状

深度神经网络模型的训练需要大量数据，训练数据越多，泛化能力越强，数据量少会导致模型过于拟合有限的训练集，难以对测试集进行泛化。因此，模型会存在过拟合的问题。虽然可以通过 Early Stoping 早停减少迭代次数 [8]、

在权重范数上加入正则化项降低模型复杂度 [9]、Dropout 随机失活 [10] 等方法在模型训练过程中减轻模型的过拟合问题，但是使用数据扩增技术得到大量的高质量训练数据才是最根本、最有效的解决过拟合问题的方法。

表 1-1: EDA 方法

扩增方法	含义	例子	缺点
同义词替换（SR）	忽略停用词，在句子中随机抽取 n 个词，在同义词词典中随机抽取同义词替换	我每天早上都会去公园运动 ——> 我每天早上都会去公园锻炼	同义词拥有相似的词向量，扩增前后的句子对于模型训练来说几乎一样，实际上没有对数据集有效扩增
随机插入（RI）	将句子中某一非停用词的同义词随机插入到句子的某一位置，该操作做 n 次	我每天早上都会去公园运动 ——> 我每天早上都会去清晨公园运动	扩增后的数据不能保持原有的语义结构和语义顺序，没有包含太多有价值的信息
随机交换（RS）	随机交换句子中某两个词语的位置	我每天早上都会去公园运动 ——> 公园每天早上都会去我运动	没有对句子中的语素进行改变，对模型泛化能力提升效果有限
随机删除（RD）	句子中的每个词都有概率 p 的可能性被删除	我每天早上都会去公园运动 ——> 我每天都会去公园运动	可能会删除对分类影响较大的特征词，甚至改变标签的正确性

现有的数据扩增技术主要有两条技术路线，回译和加噪。

回译是将原有有限数据集通过机器翻译软件翻译为其他语言，再翻译回原语言的方法。由于不同语言的逻辑顺序不同，回译方法可以增加文本数据的多样性，甚至可以改变语法结构，保留基本的语义信息，从而得到与原数据差别较大、质量比较高的新数据，这是一种比较理想的扩增方法。但是回译方法依赖机器翻译的质量。J Ma 等 [11] 使用回译方法扩增中文文本数据集在减少样本分布的不均衡性，提高分类性能上取得了很好的效果。

回译扩增示例：

[原句] 我希望每个人都能度过愉快的一天。

[扩增后] 我希望每个人都有美好的一天。

加噪是指在原有有限数据集的基础上通过对词的替换、删除等操作生成与原数据集相似的新数据。

Jason W. Wei 等 [4] 提出并验证了几种简便的加噪技巧，同义词替换（Synonyms Replace, SR）、随机插入 (Randomly Insert, RI)、随机交换 (Randomly Swap, RS)、随机删除 (Randomly Delete, RD)。该方法在训练数据较少的情况下取得了较好的效果。EDA 方法介绍简表如表 1-1所示。

同义词词典（Thesaurus）方法是加噪扩增的方式之一。Zhang Xiang 等人 [12] 提出了字符级卷积神经网络用于文本分类，在实验中发现将单词替换为它的同义词进行数据扩增可以在很短的时间内生成大量的数据。

S Kobayashi[6] 提出了一种基于上下文的情境增强数据扩增算法, 用于对文本进行分类和任务独立于一个领域的数据扩增。通过将单词替换为使用标签性条件架构的双向语言模型所要预测的其他单词, 扩增了在监督数据集中的文本。

Matt J. Kusner 等 [5] 提出了一种通过生成对抗网络的方法来生成和原数据同分布的数据，模型可以通过模仿输入的真实句子来产生真实的句子。扩句-缩句-扩句法是把原来的句子缩写，然后把它扩写。这种方法产生的句子结构与原始句子相似，但可能会造成语义信息的丢失。

刘挺 [13] 提出一种简单而有效的方法产生和利用大规模伪训练集完成零代词解析任务。侯宇泰等人 [14] 通过研究利用在训练语言数据中与一个语句有着相同的语音和话题选择的另一个句子, 提出了基于序列得到的数据强化框架。

Luque 等 [15] 提出了实例交叉增强扩增方法，这种方法是基于遗传学中染色体杂交的原理。这一情感分析方法的数据集分为两部分，用两个具有相同极性 (即正/负) 的随机推文交换。这一方法假设，即使扩展结果并非非句法或语义的，新文本也会保持情感极性。

回译和加噪两种数据扩增技术路线为探索新的解决深度神经网络任务中的文本缺失问题提供了思路。这些通用的文本数据扩增方法在文本分类、机器理解等领域提供了解决方案。

但是，在特定的专业领域如司法领域，这些通用的文本数据扩增方法在应用时会改变文本的语法结构，难以保证生成文本的数据质量，难以直接迁移到司法等具有固定语法结构和领域特征的领域。目前司法领域文本数据缺乏特定

的数据扩增规则，Guo[16] 提出了基于司法领域特征的 Blind 扩增和反事实扩增方法，并开发了 Taujud 司法裁判文书扩增工具，在均衡不同类别的数据分布、保持标签稳定性上取得了有效的结果，并基于司法文书的语法特性，使用变分编码技术应用于文本数据扩增领域，实现了司法文本数据自动化生成系统。

综上所述，目前通用的文本数据扩增方法提供了简单的数据扩增思路，但是在特定领域文本处理任务中，由于缺乏对领域特征的适配，扩增过程中对文本的语法、语义等信息造成了破坏，影响文本任务的处理结果。因此，本文将基于有限的文本数据集，充分挖掘文本中蕴含的领域特征，将领域特征应用到扩增方法中，得到基于领域特征的高质量数据。

1.2.2 文本领域特征提取研究现状

涉及不同领域的不同类型、不同格式信息是进行科学研究、促进经济发展的宝贵知识源泉，以自然语言为特征的文本数据是其主要内容，而计算机不能直接理解文本的语义信息。随着信息数据的大爆发，人们对于这些信息资源的处理日益迫切，如何处理和利用这些文本数据至关重要。文本分类、文本挖掘、语义分析、自然语言处理等文本处理任务推动着文本领域特征的研究。

领域术语是域专家用于描述和描述领域知识的基本信息载体，也是信息检索和领域特征提取的重要单元 [17]。领域术语一般只在一个或几个特定的领域文本出现，其一般是名词性短语，某一领域的众多术语可以构建领域词典，这是刻画领域特征的重要工具。对于领域术语的抽取，刘桃等 [18] 提出了一种基于信息熵的领域术语抽取方法，显著提高了文本分类的精度；岑咏华等 [19] 基于双层隐马尔科夫模型实现了中文泛术语识别和抽取的系统。构建领域词典有助于文本分类任务的实现；陈平等 [20] 针对电力审计领域的文本分类任务，提出了基于专业词典增强领域特征的文本分类方法。

用来表示文本的基本单位通常称为文本属性或特征，特征元素具有下列特征：(1) 特征元素能够区分目标文本和其他文本；(2) 特征元素必须能够识别 (1) 文本内容；(3) 特征元素应当相对易于分离；(4) 特征元素不能太多。文本特征提取是选取一部分能够直接反映文档内容信息的词语，并计算其权重。极大地减少了文本空间的规模和稀疏度，提高了文本分类的执行速度；少量主题可以被提取出来直接反映文本，以便于快速理解文本内容；这使得在计算文本之间的相似性时更加精确 [21]。在中文文本特征的表示方法中，常用的有五种：基

于字的特征表示、基于词的特征表示、基于词组的特征表示 [22]、基于 N-gram 的特征表示、基于概念的特征项表示 [23]。

特征提取就是根据一个特征评价函数计算每个特征的权值，然后将这些特征按照权值进行排序，选择出一些最高评分的特征 [21]。常用的特征估计函数构造方法有 TF-IDF，词频方法，文档频率方法，互信息法，信息增益等。特征词的特性会影响其权值，如其词频、词性、文档频次、标题、在句子中的位置、句法结构、依存句法、是否属于领域词库、词语长度等。

综上所述，目前对于文本领域特征的提取有较多的研究方向和较为成熟的研究内容，本文将吸取前人在领域特征方面的研究经验，在文本数据扩增中依据领域特征，实现符合文本所在领域特征、保持文本原始标签的扩增。

1.3 本文主要研究工作

随着大数据与人工智能技术的飞速发展，新兴技术逐渐深入应用于各行业领域的服务中，以促进各行业领域的转型升级和服务质量的提高。涉及不同领域的不同类型、不同格式的数据是促进经济发展和进行科学研究的宝贵数据资源，信息数据的大爆发与新兴技术的结合使人们更加深入地挖掘数据资源中的知识财富和领域信息。以自然语言为主要特征的文本数据是行业领域信息的主要内容，以深度学习技术为主的人工智能技术处理文本数据构建的符合人们现实需要的 AI 模型正在成为文本数据处理和利用的主要方式。

深度学习技术的实现和 AI 模型的构建是建立在大规模高质量领域数据的基础上的，对构建的模型性能的评估也依赖于广泛而具有代表性的大量数据。大规模高质量领域数据的获取在现实条件下不同领域中存在成本高、质量差的问题，阻碍着模型泛化能力的提升。如何利用有限的领域数据扩增大量的高质量数据推动模型性能的提升，对于减轻数据获取成本、提高模型泛化能力具有较为重要的意义。

目前，文本数据扩增方式有很多，而这些文本字符级别的扩增方法易抹去体现文本领域特征的词语以及体现领域特征的语义结构信息，导致扩增后的文本不能很好的体现所在领域的特征，对文本分类模型性能的提升效果不明显。

本文的主要研究内容为：针对当前文本数据扩增技术对于领域特征处理的不足，将已有的有限领域文本数据集作为数据扩增的基础数据集，使用自然语

言处理技术、主题模型等技术充分挖掘数据集中蕴含的领域特征，统一领域数据集的处理步骤，提出四种通用的基于领域特征的文本数据扩增方法：基于TF-IDF 权重的特征裁剪方法、基于主题模型的特征融合方法、基于依存句法的特征变换方法和基于词频词性的特征替换方法。最后，设计实验评估扩增文本的领域特征一致性和对于模型性能提升的效果。

本文的主要工作分为两大部分，一部分为基于领域特征的文本数据扩增技术的设计与实现步骤。论文参考现有的通用数据扩增方法，并结合领域特征挖掘技术，设计了基于已有有限数据集中特征进行文本扩增的方法。基于 TF-IDF 权重的领域特征裁剪方法是以文本词语在数据集中的 TF-IDF 值为依据，结合依存句法分析技术进行剪枝操作，保持数据集基本特征和语义一致性。基于主题模型的特征融合方法是使用主题模型技术聚类相似文本，将文本与相似目标文本进行内容交换，实现特征融合。基于依存句法的特征变换是使用依存句法分析技术解构文本，将句法树中依存关系相同文本进行交换，在不改变文本内容的情况下进行文本扩增。基于词频词性的特征替换方法是基于数据集分析构建领域集高频词表和词向量模型，提出了领域特征词语的词性表，将符合高频词和相关词性的词语集使用词向量模型推荐近义词进行替换。

另一部分是扩增文本的质量评估。构建高质量文本分类模型，将扩增后文本数据集作为测试集，与 EDA 技术对比，评估模型在两个数据集上的表现来评估文本是否保持原有的类别标签和领域特征。使用文本数据扩增技术对司法和媒体领域中的开源文本数据集进行扩增，观察扩增数据作为训练数据集对于文本分类模型性能提升的效果，评估扩增后文本的数据质量。

1.4 本文的组织结构

本文基于已有的数据扩增方法设计一种通用的基于领域特征的文本数据扩增技术，其中包含四种领域特征扩增方法，规定统一的数据集处理方式和扩增步骤，并设计实验评估扩增后文本的数据质量。本文的组织结构如下：

第一章绪论，介绍选题背景和意义、国内外文本数据扩增和文本领域特征提取的研究现状，以及本文的主要研究工作。

第二章技术综述，将介绍文本所使用的相关技术，包括自然语言处理相关技术、TF-IDF 算法、主题模型技术等。

第三章基于领域特征的扩增方法，通过分析处理原始领域数据集，挖掘领域特征，提出四种文本数据扩增方法，并对这四种方法进行详细介绍说明。

第四章文本扩增实验设计，将扩增技术应用在具体领域文本中进行数据扩增，司法领域和媒体领域。构建文本分类模型评估该技术扩增文本的质量。

第五章总结与展望，将总结本论文所做的工作，分析技术存在的不足，并对该技术的进一步研究和完善并提出展望。

第二章　技术综述

本章介绍了在研究基于领域特征的文本数据扩增技术时所用到的一些技术，包括自然语言处理技术、LDA 主题模型，TF-IDF 算法。这些技术对于分析文本数据，挖掘语料集中的领域特征具有重要作用。

2.1　自然语言处理技术

文本数据扩增技术是通过对文本的处理从而生成新的文本的技术。在处理文本数据时会用到自然语言处理技术对中文文本进行处理，包括对中文文本分词、词性标注、使用词向量计算词语的语义相似性、分析文本的依存句法结构等。本节介绍了在进行文本数据扩增时所用的相关自然语言处理技术。

2.1.1　中文文本分词

中文分词是在进行文本分类、信息检索等中文文本信息处理时的关键技术及难点。[24] 英文文本中的空格是英文分词的分割符, 因此英文分词操作要比中文分词简单得多, 只需用空格来分割英文文本即可。而在中文中，文本是汉字按照固定的语法规则组合而成的一个连续的字序列，没有明显的分割符号，因此，对中文进行分词要更困难一点。

词语是中文中表达语义信息的基本单位，在进行自然语言处理时一般需要将分词处理作为最基本的步骤，分词的质量将影响到是否可以准确提取到语义正确的词语。由于中文本身具有的多样性和其语言规则的复杂性，中文分词处理成为一项具有挑战性的技术。作为自然语言处理领域的一项重要技术，自然语言处理的研究者们对分词进行了深入研究，提出了许多分词方法，大致可以分类为词典分词法和统计分词法 [25]。

词典分词是将待分析的中文文本按照一定的规则与已建词典库中的词条进行匹配，如果与词典中相应的字符串匹配，则进行分词。这种方法对词典的

构造质量和字符串匹配规则要求较高，不能很好地解决分词时的歧义问题。例如：

[原句] 政府工作的永恒主题就是提高人民生活水平。

[恰当分词] 政府/工作/的/永恒/主题/就是/提高/人民/生活/水平。

[不恰当分词] 政府/工作/的/永恒/主题/就是/提/高人/民生/活水/平。

如果采用不同的词典库和字符串匹配规则会产生不同的分词结果，容易造成分词结果的歧义。

统计分词法是目前应用较为广泛的分词方法。首先，由语言学或领域专家给出大量已经分词的文本作为训练集，通过隐马尔可夫链模型（HMM）[26]、条件随机场模型（CRF）[27]、支持向量机（SVM）[28] 等机器学习算法训练，生成分词统计模型。其次，在进行分词任务时对分词结果进行预测，确定最优的分词结果。经过多年研究进步，统计分词法的语料集已较为完备，在中文分词任务上的性能有了很大提升。

目前，进行中文分词的开源工具不胜枚举，分词质量得到众多使用者的肯定并且在不断发展。使用较为广泛的分词工具有哈工大语言平台（LTP），中科院分词系统（ICTCLAS），清华中文词法分词工具（THU Lexical Analyzer for Chinese），结巴分词器（jieba），HanLP 分词器等。

2.1.2 词性标注

词性标注（Part-of-speech Tagging，POS ）[29] 是指在句子中分词后对每个词进行标注的技术，即判断一个词是属于名词、动词、形容词还是其他词汇。词性标注是文本预处理步骤之一，主要被应用于文本挖掘和自然语言处理领域。在中文词汇标注方法中，有基于统计的方法和基于规则的方法，其准确性与分词的准确性有着密切的关系。

词性标注在本质上是分类问题。目前，分词和词性标注一体化模型发展迅速，将分词和词性标注结合在消除歧义和提高整体效率上取得很好的效果[30]。部分词性标注编码如表2-1所示。

词性标注示例：

[例句] 政府工作的永恒主题就是提高人民生活水平

[词性标注] 政府 [n]/工作 [v]/的 [u]/永恒 [z]/主题 [n]/就 [d]/是 [v]/提高 [v]/人

民 [n]/生活 [v]/水平 [n]

表 2-1: 部分词性标注编码表

词性代码	描述	举例	词性代码	描述	举例
a	形容词	最	nr	人名	特朗普
b	区别词	副	ns	地名	纽约
c	连词	和	q	量词	次
d	副词	非常	r	代词	有些
e	叹词	唉	v	动词	举办
f	方位词	中	z	状态词	依然
...	...	...	...	...	...

2.1.3 依存句法分析

依存句法分析 (Dependency Parsing, DP)) 是一种通过分析语言单元中成分之间的依赖关系来确定其句法结构的技术 [31]。Tesniere 在 1959 年提出依存句法分析理论，该理论认为，句子中的主要词语可以支配其他成分，主要词语是动词且只有一个，所有被其支配的成分依存于该动词 [32]。

表 2-2: 依存句法分析部分标签说明表

标签	描述	关系类型
SBV	Subject-verb	主谓关系
VOB	Verb-object	动宾关系
FOB	Fronting-object	前置宾语
ATT	Attribute	定中关系
ADV	Adverbial	状中关系
HED	Head	核心关系
COO	Coordinate	并列关系
...	...	...

目前，依存句法分析方法主要有：基于规则的方法、基于统计的方法和基于深度学习的方法。衡量分析质量的指标主要有：依存正确率（dependency accuracy，DA）、根正确率（root accuracy，RA）、完全匹配率（complete match，CM）等。

当前有许多的开源自然语言处理工具可用于中文的句法分析，哈工大语言云平台 LTP 中包含了依存句法分析工具，该工具采用神经网络进行依存句法分析的算法，能够有效地分析中文依存句法，其在测试集的 LAS 值取得了 0.814 的成绩 [33]。依存句法关系的部分标签说明如表2-2所示：

依存句法结构以谓词为核心通过树形结构来表示句子中各部分之间的关系，体现句子中的逻辑关系。依存句法树的定义如下：

（1）依存句法描述词语之间的依存关系，这种依存关系可以表示为二元组（W, R）的形式，其中 **W** 为词语集合，**R** 为依存关系的集合，对于任意词语 $w_1, w_2 \in \mathbf{W}$，如果 w_2 依存于 w_1，则 $\langle w_1, w_2 \rangle \in \mathbf{R}$, 可记作 $w_2 <_{\mathrm{r}} w_1$，其中 r 是表示词语具体的依存关系，包括 ATT（定中关系），SBV（主谓关系），VOB（动宾关系）等。

（2）按照依存句法公理，依存关系存在最小上确界 Root，且每一个词的依存词不超过 1 个，符合：$\forall < w_1, w_2 > \in \mathbf{R} \wedge \forall w_3 \in \mathbf{W} \wedge w_3 \neq w_1 \Rightarrow \langle w_3, w_2 > \notin \mathbf{R}$。

（3）依存句法结构可以表示成树的形式，记为 **T=(W,E)**，其中 **W** 表示词语（结点）的集合，**E** 表示依存关系（边）的集合，对于任意词语 $w_1, w_2 \in \mathbf{W}$，如果 $w_2 <_{\mathrm{r}} w_1$，则一定存在一条有向边 $(w_1, w_2) \in \boldsymbol{E}$。其中 w_1 为 w_2 的父结点 [34]。

如图2-1所示，箭头指示了词与词之间的依赖关系，箭头指向了依赖当前词的成分，而箭头的标签则表示了相应的依存关系。一句话里有一个词不依存其他成分，如例句中的核心词“举办”，“Root”结点指向核心词。“周末”、“学校”、“在”、“运动会”与“举办”建立了依存关系，其中“学校”的句法标签是“SBV”（主谓关系），代表句子的主语是“学校”。因此，通过依存句法树可以图形化表示句子的依存关系。

依存句法分析例句：周末学校在操场举办秋季运动会。

[依存句法分析树]

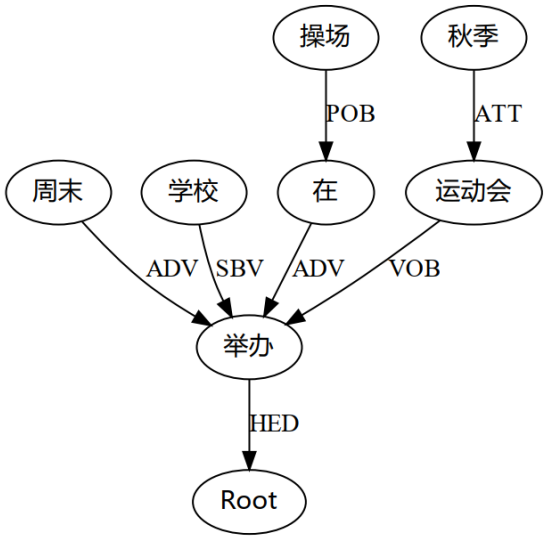


图 2-1: 依存句法树示例图

2.1.4 词向量

在自然语言处理技术中，文本作为非结构化的数据，如果要被计算机处理首先需要转换为可计算的数值型数据。实现文本到数值的转化，需要对文本进行分词、去停用词操作，而后将词语作为文本的基本单位进行数值化转化。词语的数值化表示有离散型和分布式型两种表示方法。

以词表的大小作为向量长度，每个词都表示为文本词汇表中的一个索引，或独热 (one-hot) 编码向量，其相应的索引位置为 1 其余部分为 0。例如，在某段文本中，“青岛”、“南京”词语可以表示为：

“青岛”表示为 [0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ...]

“南京”表示为 [0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 ...]

这种表示方式具有简单、健壮的优点，但词向量表示并无关联，不包含任何语法语义信息。如果文本词汇表很大，那么独热编码就是稀疏向量，向量模型计算量急剧增加，会导致维数灾难。

分布表示法是用定长、连续、稠密的向量来表示词。词向量 (wordembedding) 是一种常用的分布式表示方式。词向量是一种稠密矢量，其维度相对较小，将原始稀疏的大维向量强制嵌入到小维空间中，将 one-hot 编码向量中的每个元素从整数型转换为浮点型，表示整个实数域，它的每个维度都有实数，而不是大部分维度为 0，类似为：[0.764, −0.147, −0.907, 0.608, −0.486, ...]。

本文使用 Word2vec 作为训练词向量的工具。Word2vec 可以在数据集上进

行高效地训练，最终训练结果就是词向量。它是一个浅层的神经网络，分别为输入层，隐藏层和输出层，其中包含两个模型：跳字模型（Skip-Gram）和连续词袋模型（Continuous Bag of Words，CBOW）。CBOW 和 Skip-gram 模型结构图如 2-3。

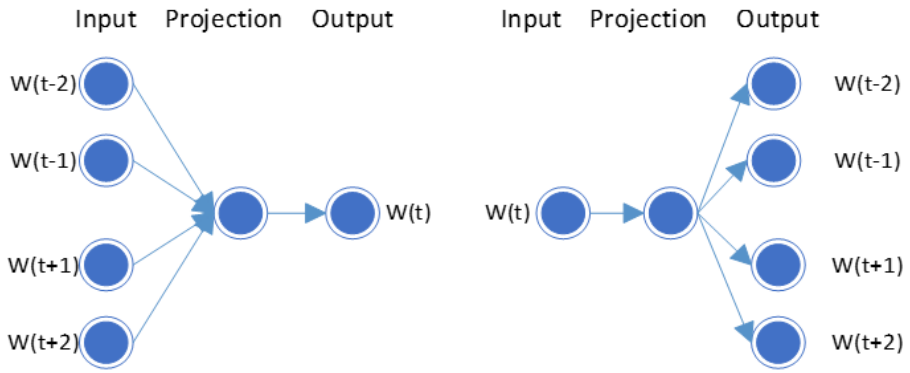


图 2-2: CBOW 和 Skip-gram 模型结构图

连续词袋模型的输入是上下文中某一词的词向量，输出是该词的词向量，而不考虑词间的次序。跳字法的思路与之相反，即输入是单词的词向量，而输出是单词的上下文词向量。通过对大量语料进行训练后，隐含层向量就可以用作预训练的词向量。对于小型数据集，CBOW 算法比较适合，而 Skip-Gram 则适合于大型数据集。

在词向量空间中，意思相近的词语会被映射到空间中相近的位置。根据这一特性，使用夹角余弦值可以计算词语间的相似度。夹角余弦的计算方法如公式 2-1所示，空间中的两个词的语义越相似，它们的夹角余弦值越接近 1，反之为 0。

$$\mathrm{sim}\,(W_1, W_2) = \frac{\sum_{i=1}^{n} W_{1i} * W_{2i}}{\sqrt{\sum_{i=1}^{n} (W_{1i})^2}\sqrt{\sum_{i=1}^{n} (W_{2i})^2}} \tag{2-1}$$

其中：W_1，W_2 为词向量，W_{1i}，W_{2i} 分别表示 W_1,W_2 的各分量。

2.2 LDA 主题模型

主题模型（topic model）是以非监督学习的方式对文集的隐含语义结构（latent semantic structure）进行聚类（clustering）的统计模型 [35]。主题模型

一般是指经典的 LDA 主题模型，也就是隐含狄里克雷分布 (Latent Dirichlet Allocation)[36]，该模型由 D.Blei 等人于 2003 年提出。

LDA 主题模型是一个贝叶斯概率模型，它包含词、文档和主题三层结构。这个模型将文本视为“词袋”，即将每段文本视为一组没有顺序的单词。

LDA 模型是一种无监督学习算法，认为每一段文字由多个符合概率分布的主题组成。每一主题都包含多个特征词，特征词也符合概率分布，只要确定语料集和主题个数 K，模型就可以自动生成主题类型。其中，每个主题下词语出现概率的计算公式如下：

$$p(\text{词语} \mid \text{文档}) = \sum_{\text{主题}} p(\text{词语} \mid \text{主题}) * p(\text{主题 文档})$$

该模型可以预测数据集中每个文本的主题并以概率的分布的形式给出，也可以给出数据集中每个主题包括的特征词，即文档-主题分布和词语-主题分布。

对于数据集中的文本，LDA 的生成过程如下：

1）对每篇文本，在主题分布中随机选取一个主题

2）在步骤 1 选取到的主题对应的词语分布中随机选取一个词语

3）重复上述过程直至遍历文本中每个词语，从而生成文档-主题分布和词语-主题分布。

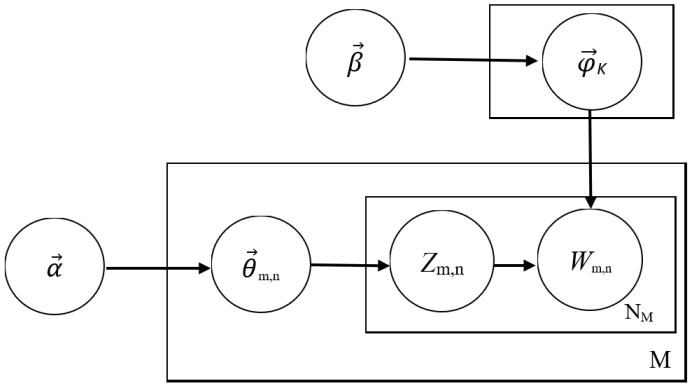


图 2-3: LDA 模型图

LDA 模型图如图2-3所示，图中方框表示在该模型中反复迭代的内容，$W_{m,n}$ 节点表示可观察的值，其他节点表示隐含的随机变量或参数，箭头表示依赖关系。其中 M 表示文档数目，N_M 表示文档 m 的总词数，$\overrightarrow{\theta_m}$ 表示文档 m 的

主题概率分布，$\overrightarrow{\phi_k}$ 表示第 K 个主题下的词分布，$\vec{\alpha},\vec{\beta}$ 分别为多项分布 $\overrightarrow{\theta_m},\overrightarrow{\phi_k}$ 的先验分布（即狄利克雷分布）的参数。

LDA 算法步骤如下：

1、对于数据集 M：

选取 K 个主题-词分布 $\overrightarrow{\phi_k} \sim \mathrm{Dir}(\vec{\beta})$，这里 $k \in \{1,2,\ldots,K\}, \mathrm{Dir}(\vec{\beta})$ 表示参数为 $\vec{\beta}$ 的狄利克雷分布。

2、对于数据集中每篇文档：

1）从文档-主题的分布 $\overrightarrow{\theta_m} \sim \mathrm{Dir}(\vec{\alpha})$ 选取一个主题，这里的 $m \in \{1,2,\ldots,M\}$。

2）文档中词 $W_{m,n}$ 重复以下过程，这里的 $n \in \{1,2,\ldots,N_m\}$。$W_{\mathrm{m,n}}$

a. 为词 $W_{m,n}$ 从主题分布中选取一个主题 $Z_{\mathrm{m,n}}$，这里 $Z_{m,n} \in \{1,2,\ldots,K\}$，共有 K 个主题。

b. 根据 $Z_{\mathrm{m,n}}$ 对应的第 K 个主题-词分布 $\overrightarrow{\phi_k}$，从这个多项分布中采样出当前词。

2.3 TF-IDF 算法

TF-IDF（term frequency－inverse document frequency，词频-逆向文本频率）[37] 是一种常用的针对数据集中词语的加权方法。

TF 是词频，用来衡量某个词语在数据集中出现的次数。有些常用词在表达主题特征方面作用不大，一些出现频率较小的词语可以清晰表达文档所属主题。因而，又提出 IDF 指数弥补词频统计的不足。IDF 是逆文本频率指数，在统计学上，一个词或一句话的重要性与它在文本中出现的次数成正比，而在语料库中则是相反的。所以 TF 和 IDF 的结合能够评估出单词在文本中的重要性，并可以有效进行文本分类和提取文本信息。

对于包含 n 个文本的语料集 D={d_1,d_2...d_n}, 文本 d_i 中词语 v 对于该文本的重要性由它在 d_i $i \in (1,n)$ 中出现的频率和在语料集中出现的频率共同决定。词语 v 的 TF-IDF 值计算公式如2-2所示：

$$tfidf(v,d,D) = tf(v,d) \times idf(v,D) \tag{2-2}$$

其中，词语 v 在 d_i 中的词频计算公式如2-3所示，它描述了文档中词语的

词频计算方法：

$$tf(v,d) = \frac{f_{d(v)}}{\sum_{\omega \in d} f_{d(v)}} \tag{2-3}$$

词语 v 在语料集 D 中的逆向文本频率 IDF 计算公式如2-4所示，它描述了词语 v 在其他文档中出现的频率：

$$\text{idf}\,(v,D) = \log_2\left(\frac{|D|}{|(d \in D, t \in d)| + 1}\right) \tag{2-4}$$

$|D|$ 代表语料集的文本数量，$|(d \in D, t \in d)|$ 代表词语 v 在语料集中出现的频率。考虑到分母上可能会为 0，因此取 $|(d \in D, t \in d)| + 1$ 作为分母。通过公式可以看出，词语在语料集中出现的次数越多，IDF 越小，TF 越大，也就是说，对于一个词语，如果它在一个文本中出现的次数越多，同时在语料集中出现的次数越少，说明它的 IF-IDF 越大，该词对文本的重要性越大。

在实践中，为避免一些常用词对关键词加权的影响，一般在计算 TF-IDF 前会对语料集进行去停用词操作。

2.4 本章小结

本章首先介绍了自然语言处理的相关技术，这是进行文本处理的关键技术，包括中文分词技术、依存句法分析技术、词向量和词性标注。其次介绍了 LDA 主题模型技术，这是领域特征抽取的技术，可以在数据集上构建文档-主题分布和特征词-主题分布。然后介绍了 TF-IDF 算法，该算法在数据集的层面针对数据集中词语进行加权，可以衡量某个词语在数据集中的重要程度。这些技术通过对文本的处理和领域特征的挖掘，为接下来基于领域特征的文本数据扩增技术研究提供技术支撑。

第三章　基于领域特征的扩增方法

本章首先介绍了在进行基于领域特征扩增之前对于文本所在数据集的一般处理步骤，其次详细介绍了四种基于领域特征的文本数据扩增方法：基于TF-IDF 权重的特征裁剪方法，基于主题模型的特征融合方法，基于依存句法的特征变换方法和基于词频词性的特征替换方法。为便于介绍这四种方法的实现过程，本章以在司法裁判文书数据集 CAIL2018_ALL_DATA 进行文本扩增为实验案例，在该数据集上实现这四种方法的扩增。

3.1　文本预处理

文本预处理基于领域特征的文本数据扩增的初始步骤，它包括文本结构化、文本分词、去停用词和文本词频统计。文本预处理的目的是在文字扩增之前将数据以结构化形式存储起来，同时保存文本的预处理结果 (文本分词结果和词频统计结果)，避免扩展过程中同一文本多次重复处理，造成计算资源的浪费。预处理结果存储在 json 格式中。

文本结构化处理是抽取数据集中的原始文本数据和其所属标签，并添加序号，进行初始的分词操作后以 json 格式存储，以便于后续的读取和处理。为便于不同领域数据集的处理和存储，该步骤将文本数据统一处理成标准结构化文本格式。标准结构化文本格式如表 3–1所示。

表 3–1: 文本预处理的标准结构化文本格式表

字段名称	格式	含义
id	int	此条文本的序号
lable	string	此条文本的标签
content	string	此条文本的内容
content_seg	list	此条文本的分词格式

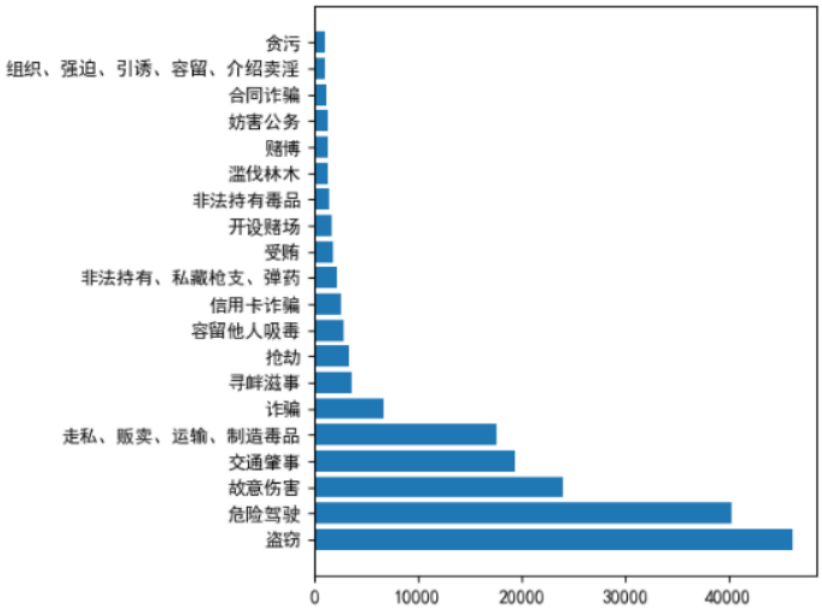


图 3-1: 司法裁判文书数据集所涉罪名类型分布图

以司法领域数据为例。本文选取司法裁判文书数据集CAIL2018_ALL_DATA，该数据集是 2018 中国“法研杯”法律智能挑战赛所用开源数据集，是大量司法裁判文书及其标签的集合，包含裁判文书所涉罪名、被告人、刑期、罚金等属性。

```
def wordProcess():
    ltp = LTP()
    ltp.init_dict(path="law.txt", max_window=4)
    f = open('judicial.json', 'r', encoding="utf8")
    stopwords = getStopwords()
    lines = json.loads(f.read())
    wordJSON = []
    for line in lines:
        line_segs = ltp.seg([line])
        wordprocess =''
        for line_seg in line_segs[0][0]:
            if line_seg not in stopwords and
                    is_all_chinese(line_seg):
                wordprocess += line_seg + "□"
        wordJSON.append(wordprocess)
    return wordJSON
```

图 3-2: 文本预处理分词操作核心代码

该数据集数据质量较高，可用作司法领域学术研究。对于数据集结构化处理时的内容选择，只选取实验所需的标签（司法裁判文书内容及其所涉罪名标签），将数据集内容选取限制在最小范围，然后添加文本序号和分词操作。

对于司法裁判文书数据集 CAIL2018_ALL_DATA，本文选取其中的 test 文件为扩增案例实验的基础数据集，在该文件中有 217016 条文书数据，涉及 20 种罪名，该数据集罪名类型数量分布如图 3-1所示。其中盗窃（46177 条），危险驾驶（40276 条），故意伤害（24017 条），交通肇事（19441 条），走私、贩卖、运输、制造毒品（17642 条）等五种类型数量较多，故选取这五类文书共 147553 条作为实验数据集。

在进行文本分词操作的过程中，实验使用了哈工大 LTP 自然语言处理工具进行分词操作，调用 LTP 分词器的 Python 的 API 接口。LTP 分词器是通用的文本分词工具，在对司法裁判文书的分词处理中，易将很多具有司法领域特征的专有名词进行拆分，使得分词的结果不准确，如：司法领域词语“驳回上诉”会被分为“驳回”、“上诉”两个词。为保持分词结果的领域特征性，避免一些领域特征词被“拆解”，在分词工具中添加开源的法律领域词典，如果在分词时遇到司法领域的特征词将不会被分割。文本分词操作核心代码如 3-2所示。

图 3-3: 司法裁判文书数据集高频词语词云图

去停用词也是文本预处理的一个重要步骤。停用词是指在文本中不包含具体含义的虚词，如：“的”、“了”、“即”等词。在计算 TF-IDF 值和词向量时，停用词的出现会影响计算结果，因此在文本预处理中一般将分词后结果

集中的停用词删除。在司法裁判文书数据集中，文书已经经过脱密处理，包含大量的数字和特殊字符，在去停用词时，同时将包含数字和特殊字符的文本进行删除。

词频统计是计算分词后的结果集中每个词出现的次数并排序。词频统计结果可以在一定程度上反映了词的重要性，高频词集也反映了所在领域的特征。为使词集真实反映司法领域特征，处理中将人名如：“张某某”等大量出现且非领域特征词删除。由此，生成该数据集词云图，如图3-3所示。可以看出，词云图具有鲜明的司法领域特征。

3.2 基于 TF-IDF 权重的特征裁剪

文本特征提取是指从文档中选取一部分能反映其内容信息的词语及其权重的计算。TF-IDF 算法可以对数据集中的词语进行加权，是一种能够有效提取文本特征的方法。基于 TF-IDF 权重的特征裁剪是在词语加权的基础上结合文本的依存句法分析，对文本进行裁剪从而实现数据扩增。

TF-IDF 算法是一种对数据集中词语加权的方法。TF 是指词频，IDF 是逆文本频率指数。一般而言，一个词语的重要性与其在文本中出现的次数成正比，而与其在数据集中出现的次数成反比。将 TF 和 IDF 相结合，可以评价单词对文本的重要性，并能有效地提取文本的特征信息。

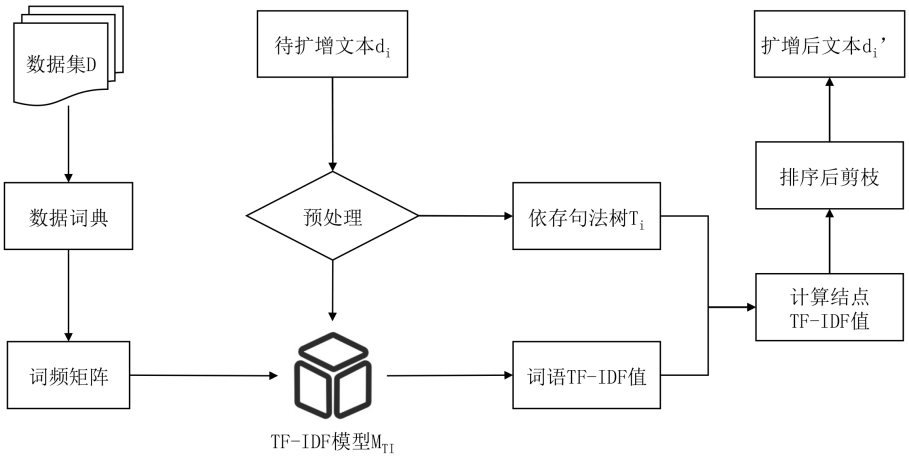


图 3-4: 基于 TF-IDF 权重的特征裁剪扩增方法流程图

使用 TF-IDF 方法对文本中的词语进行加权，将文本中反映文本特征的重要词语筛选出来，对于权重较低的词语进行删除，从而达到保留文本特征的

前提下，实现文本扩增的目的。对于传统的基于 TF-IDF 权重的扩增方法只对单个词语进行删除操作，难以保证语义的合法性和领域特征的一致性。基于 IF-IDF 权重的特征裁剪方法将对文本的操作粒度扩展到依存句法树的“树枝”。

特征裁剪（Feature Clipping，FC）扩增方法步骤为：对于包含 n 个文本的数据集 $D = \{d_1, d_2 \cdots d_n\}$，根据 TF-IDF 计算方法构建基于数据集 D 的 TF-IDF 模型 M_{TI}，对于待扩增文本 d_i ($d_i \in D$) 进行分词和依存句法分析，得到 d_i 中分词后的词集 $W = \{W_1, W_2 \ldots W_n\}$ 和依存句法树 T_i。根据模型 M_{TI} 计算词集 W 中每个词语的 TF-IDF 值。其中，停用词、数字和特殊符号的 TF-IDF 值取 0。进而计算出依存句法树 T 中每个父结点及其所属的全部子结点的 TF-IDF 值总和，并按数值降序排列，将排在后 p% 的树枝随机删除，最后得到扩增后文本 d_i’。其扩增方法流程图如3-4所示。

基于 TF-IDF 权重的特征裁剪（Feature Clipping，FC）文本扩增方法是在领域数据集的基础上构建 TF-IDF 模型，将待扩增文本数据进行分词和依存句法分析后，根据 TF-IDF 模型得到每个词的 TF-IDF 值，其中停用词、数字和特殊符号的 TF-IDF 值取 0，进而计算依存句法树中每个父结点及其子结点的 TF-IDF 值的和，以评估每个树枝的重要性，将 TF-IDF 总和较小的树枝删除，从而得到新的文本的一种扩增方法。

```
def buildtfidf():
    f = open('wordprocess.json', 'r', encoding="utf8")
    res = json.loads(f.read())
    corpora_documents = []
    for item_text in res:
        corpora_documents.append(item_text.strip()
                                          .split("□"))
    #生成语料词典
    dictionary = corpora.Dictionary(corpora_documents)
    dictionary.save('dict.txt')
    #得到语料中每一篇文档对应的稀疏向量
    corpus = [dictionary.doc2bow(text) for text
                                in corpora_documents]
    tfidf_model=models.TfidfModel(corpus)
    tfidf_model.save("data.tfidf")
```

图 3-5: TF-IDF 模型构建核心代码

以司法领域为例进行特征裁剪文本扩增实验。首先，在选取的司法裁判文书数据集中构建 TF-IDF 模型。本文使用 python 的 gensim 库进行模型构建，将分词处理后的文本转换为 JSON 格式。接下来是生成数据词典，即为文本中出现的所有词语标记 id 序号，以区分不同的词语并便于计算每个词语在文本中出现的频次，并将生成的数据词典保存。然后根据数据词典计算每一条文本中每个词在文档中出现的次数得到对应的稀疏向量。最后使用 TF-IDF 算法构建数据集的 TF-IDF 模型。模型构建代码如图 3-5所示。

以司法裁判文书“公诉机关指控：2017 年 4 月 30 日 18 时许，被告人吴某在广东省肇庆市端州区广百时代广场公交车站台，趁被害人梁某准备上公交车时，盗窃了梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机（经鉴定价值人民币 1700 元）”为待扩增文本，在该文本上进行特征裁剪扩增示范。

首先将该文本进行分词和依存句法分析，得到分词列表和依存句法树。依存句法树表示了文本中语法成分之间的依存关系，是文本特征的一种表现形式，根据依存句法树可以清晰看出词语之间的关联关系。该文本的依存句法树如图 3-6所示。

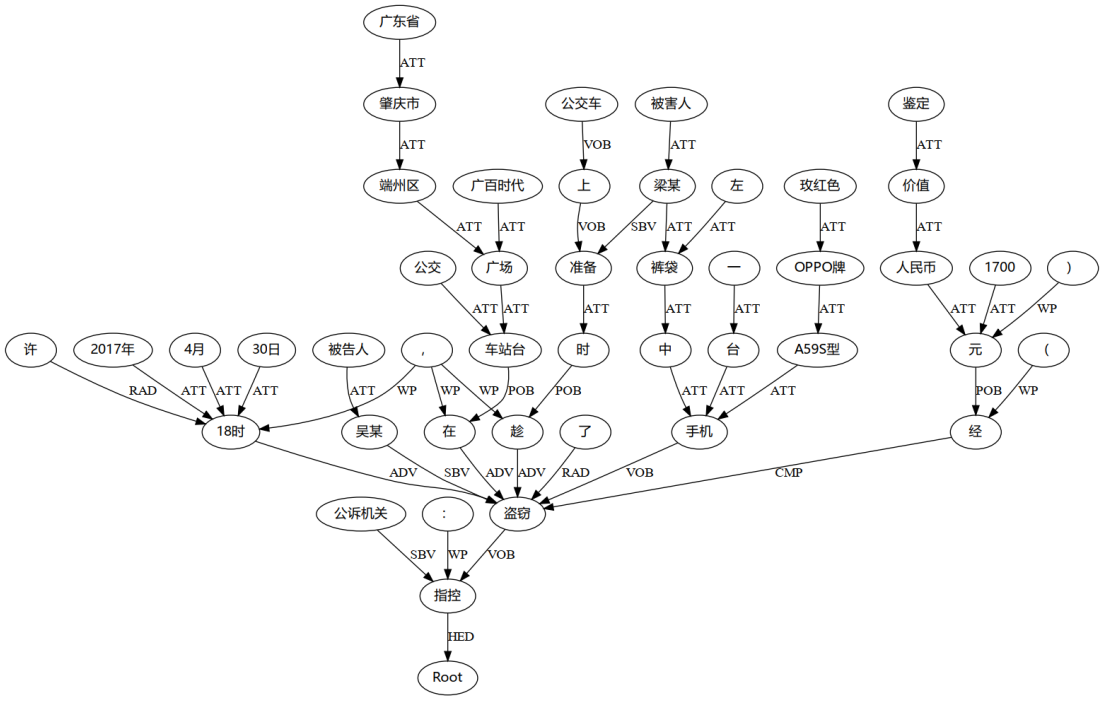


图 3-6: 特征裁剪扩增方法待扩增示例文本依存句法树

其次，根据生成的 TF-IDF 模型，计算得到文本中每个词语的 TF-IDF 值，TF-IDF 值大的词语重要性较大、领域特征强，反之亦然。在计算 TF-IDF 值

时，对于文本中出现的数字、停用词和特殊字符（标点符号等），它们对于文本的重要性极低，为了便于计算，将 TF-IDF 值设置为 0。词语赋值代码如图3-7所示。

```
#得到每个词语的TF-IDF值
'''
    line_seg: 分词列表
    string_tfidf : 模型中包含词语的TF-IDF值
    dictionry: 语料词典列表
'''
def tfidf(line_seg, string_tfidf, dictionry):
    line_dic = {}
    for i in line_seg:
        if i in dictionry.keys():
            line_dic[i] =
                string_tfidf[dictionry[i]][1]
        else:
        #非词典中的词语的TF-IDF值置为0
            line_dic[i] = 0
    return  line_dic
```

图 3-7: 待扩增文本分词词语 TF-IDF 赋值计算核心代码

如果在文本扩增时只以词语的 TF-IDF 值为依据进行删除，易导致文本语义的缺失和语法的不规范。依存句法树中词语之间相互联系，一个词结点与多个词结点相互关联，之间有依存关系，多个词结点可以看作文本中一个小整体。本方法选择以依存句法树的树枝为操作的基本粒度，计算每个父结点及其所有子结点的 TF-IDF 值和，以此为依据进行树枝裁剪操作，保证文本的语法的合法性和领域特征一致性。

在选取依存句法树枝时，存在树枝所包含结点数量大小的问题，如果将树枝所含结点设为 1，则裁剪的基本粒度就是一个词语，如果设置范围为文本的分词数目，则裁剪粒度就是整棵树。为此，设置参数 lengthWeight 来确定最终裁剪树枝范围，如设置 lengthWeight=0.5，代表所含结点数量小于文本结点数量的一半的树枝为裁剪树枝的范围。

在划定裁剪范围后得到若干根树枝，计算范围内所有树枝的 TF-IDF 值总和并降序排列。树枝的 TF-IDF 值代表了树枝对于文本的重要性，选取排在后面的树枝进行裁剪。在划定排序范围时，同样存在范围不清的问题，设置参数 rangeWeight 来确定选取排序范围，如设置 rangeWeight=0.5，代表将排序在后半

段的树枝作为待裁剪树枝列表。生成待裁剪树枝列表的代码如图3-8所示。

```
#lin_seg: 分词结果
#line_dic: 词语TF-IDF值词典
#dep: 依存句法关系
def countTree(line_seg,line_dic,dep,lengthWeight,
                        rangeWeight,quantityWeight):
    treelist = []       #依存句法树中所有的树枝
    length = len(line_seg)      #所有分词数量
    all = {}        #所有枝的tfidf集合
    cliplist = []       #待剪枝叶集和
    clipping = []       #待剪枝叶的序号
    #找到所有的枝叶
    for i in range(len(line_dic)):
        list = getAllson(dep, i)
        if len(list) <= length * lengthWeight
                                and len(list) > 1:
            treelist.append(list)
    #计算所有的枝叶tfidf
    for cl in range(len(treelist)):
        sumtfidf = 0
        for c in treelist[cl]:
            sumtfidf = sumtfidf +
                        line_dic[line_seg[c-1]]
        all[cl] = sumtfidf
    #给所有的枝叶根据tfidf值排序
    sortList = sorted(all.items(), key=lambda
                                item:item[1])
    #排序列表的裁剪范围
    rangelist = sortList[:int(len(treelist)*
                                rangeWeight)]
    #待裁剪树枝列表
    solist = random.sample(rangelist,
            int(len(treelist)* quantityWeight))
    for tu in solist:
        cliplist.append(treelist[tu[0]])
    for i in cliplist:
        for j in i:
            if j not in clipping:
                clipping.append(j)
    return clipping
```

图 3-8: 生成待裁剪树枝列表核心代码

在待裁剪树枝列表中，如果每次扩增都固定将 TF-IDF 值最小的部分树枝进行裁剪，会导致每次扩增后的文本都相同。为了使每次运行产生不同的结果，设置参数 quantityWeight，如果 quantityWeight=0.5，表示随机选取待裁剪列表中的一半树枝为最终裁剪对象。

表 3-2: 基于 TF-IDF 权重的特征裁剪扩增方法参数表

参数	默认值	含义
lengthWeight	0.4	计算 TF-IDF 时选取树枝的范围
rangeWeight	0.4	树枝 TF-IDF 排序列表的选择范围
quantityWeight	0.4	待裁剪树枝数量占待裁剪列表的比重

基于 TF-IDF 权重的特征裁剪扩增方法的裁剪参数表如3-2所示。该方法将三个参数值默认为 0.4，表示将特征裁剪幅度保持在一个适中程度。

最后，根据待裁剪列表根据参数 quantityWeight 随机选取树枝进行删除操作。待扩增文本在经过一次特征裁剪扩增后得到扩增后文本结果如表3-3所示。扩增后文本的依存句法树如图3-9所示。

表 3-3: 特征裁剪扩增方法待扩增示例文本扩增结果对照表

初始文本	扩增文本
公诉机关指控：2017 年 4 月 30 日 18 时许，被告人吴某在广东省肇庆市端州区广百时代广场公交车站台，趁被害人梁某准备上公交车时，盗窃了梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机（经鉴定价值人民币 1700 元）。	公诉机关指控：被告人吴某在端州区广百时代广场公交车站台，趁被害人梁某准备时，盗窃了梁某左裤袋中手机。

由表3-3和图3-9可以观察出，扩增后文本与初始文本相比，内容精简，但是核心语义信息不变，仍然可以看出扩增后文本仍然是盗窃案件。初始文本中的“2017 年 4 月 30 日 18 时许”、“广东省肇庆市”、“上公交车”、“一台玫红色 OPPO 牌 A59S 型”、“（经鉴定价值人民币 1700 元）”等短语被裁

减，也可以直观的看出，这些短语对于反映领域特征的作用较小，该方法有效的将一些权重较小的树枝删减且保证了领域特征一致性。

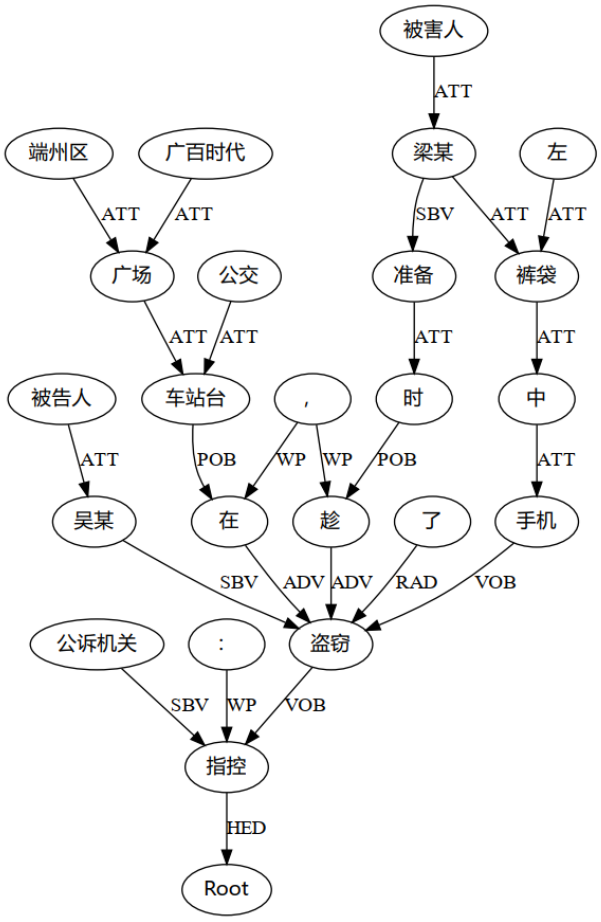


图 3-9: 特征裁剪扩增方法待扩增示例文本扩增结果依存句法树

3.3 基于主题模型的特征融合

基于主题模型的特征融合扩增方法是一种从数据集中选取与待扩增文本特征相似度较高的目标文本，抽取文本中的特征相互替换，从而实现扩增的方法。进行特征融合的关键是根据文本相似度进行筛选推荐和文本特征抽取。

在根据文本相似度进行相似文本筛选时，使用 LDA 主题模型技术。LDA 主题模型是隐含狄利克雷分布模型，以非监督学习的方式对文本进行聚类，是一种包含词、文档和主题三层结构的贝叶斯概率模型。该模型可以预测数据集中每个文本的主题、也可以给出每个主题包含的特征词。使用 LDA 主题模型进行文本筛选推荐是属于基于内容的推荐方法，可以从数据集中发掘并提取主

题，进而在待扩增文本所属主题中选取与待扩增文本相似度较高的文本，实现较高质量的筛选推荐。文本特征抽取使用依存句法树对文本中的依存关系进行分析，从而获取文本的基本特征。

特征融合（Feature Fusion，FF）扩增方法步骤为：对于包含 n 个文本的数据集 $D = \{d_1, d_2 \dots d_n\}$，根据困惑度 p 确定最优主题数 K，构建 LDA 主题模型 M_{1da} 并得到主题-文档表。对于待扩增文本 d_i ($d_i \in D$) 进行预处理，根据模型 M_{1da} 得到 d_i 所属可能性最大的主题 q($q \in 1, 2 \dots K$)，计算主题 q 中所有文本与 d_i 的余弦相似度并排序，得到相似度较高的目标文本 d_j。对 d_i 和 d_j 进行依存句法分析，得到对应的依存句法树 T_i 和 T_j，选取 T_i 和 T_j 上具有相同依存关系的树枝进行替换，得到扩增后文本 d_i'。扩增流程图如图3-10所示。

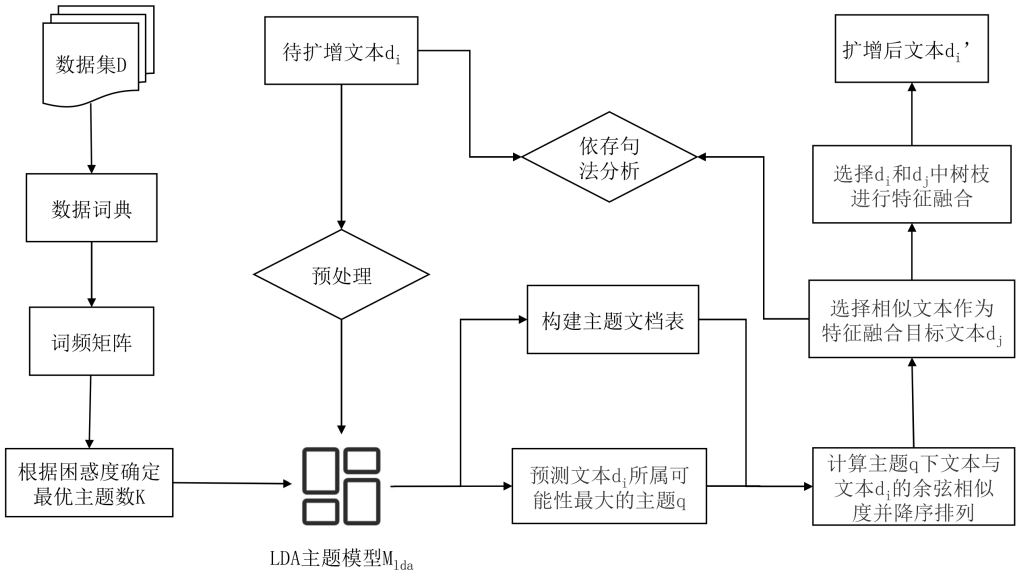


图 3-10: 基于主题模型的特征融合扩增方法流程图

基于主题模型的特征融合文本扩增方法是在已有数据集上训练 LDA 主题模型。在训练主题模型的过程中很重要的一个过程是确定模型的主题数，主题数的大小影响着主题模型性能的好坏，而主题数的确定不是随意的。

主题数的确定一直是一个复杂的问题。目前有基于经验调试法、基于困惑度法、贝叶斯统计标准方法 [38] 等。LDA 主题模型本质上是一种贝叶斯概率模型，而困惑度是一种评估概率模型的方式，为了客观性确定最优主题数目，本方法选取困惑度作为评估模型好坏的指标。困惑度的计算需要提供一个测试集，其中包含 M 个测试文本。对于每个文本，它的单词序列为 $W = \{\vec{\omega}_1 \cdots \vec{\omega}_m\}$，而 v 则是训练好的概率模型。测试集中的句子都是具有正常

语义的句子，因此训练好的模型应该在测试集上的概率越高越好，概率公式如3-1所示：

$$P(\widetilde{W} \mid v)=\prod_{m=1}^{M} p\left(\vec{\omega}_{\widetilde{m}} \mid v\right)^{-\frac{1}{N}} \tag{3-1}$$

接下来给出困惑度的表达式，其中 Nm 文本中的单词数量：

$$\text{perplexity}(\widetilde{W} \mid v)=\prod_{m=1}^{M} p\left(\widetilde{\omega_{\widetilde{m}}} \mid v\right)^{-\frac{1}{N}}=\exp -\frac{\sum_{m=1}^{M} \log p\left(\widetilde{\omega_{\widetilde{m}}} \mid v\right)}{\sum_{m=1}^{M} N_{m}} \tag{3-2}$$

显然，如公式3-2所示，句子概率越大，则语言模型越好，此时困惑度应该越小。

以司法裁判文书数据集为例，在初始化设置中，设主题数目为 K $\in$ [10, 150]，间隔数 step=10，困惑度的测试集随机取数据集中 10% 的数据，对于不同的 K 值训练出的模型，其困惑度结果如图3-11所示。

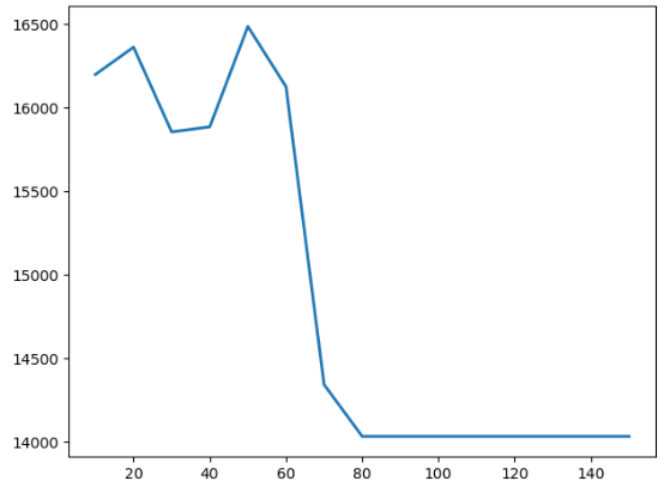


图 3-11: 司法数据集主题模型在不同 K 值下的模型困惑度曲线图

可以看到，困惑度的总体趋势是随着主题数目 K 值的增加而减少，但是 K 值超过 40 时，困惑度又向上浮动，在超过 70 时急剧减少。由图可得，在困惑度大于 80 时可以取得最小值。因此，在此数据集上可取主题数目 K 为 100，生成最优模型。LDA 模型生成代码如3-12所示。

```
def getLDAmodel():
    #获取语料集
    corpora_documents = getCorpora()
    #加载词典文件
    dictionary = corpora.Dictionary.load('dict.txt')
    # 向量的每个元素代表一个词语在文档中出现的次数
    dict_corpora = [dictionary.doc2bow(i) for i
                                    in corpora_documents]
    #将生成的语料保存成MM文件
    MmCorpus.serialize('data_corpora.mm', dict_corpora)
    np.random.seed(SOME_FIXED_SEED)
    lda = models.LdaModel(dict_corpora, num_topics=100,
    id2word = dictionary, iterations=1000)
    #将生成的主题模型保存为文件
    lda.save(r'data_mylda')
```

图 3-12: LDA 主题模型构建核心代码

在代码实现中，首先要读取数据集，加载在基于 TF-IDF 权重的特征裁剪扩增处理时生成的词典文件，该文件为每个词语标注了序号，便于对词语进行定位处理。然后统计每个词语在数据集中出现的次数，将计算结果文件保存，以便于后续读取和应用。最后，使用 gensim 库的 LDA 模型构建方法构建 LDA 主题模型。

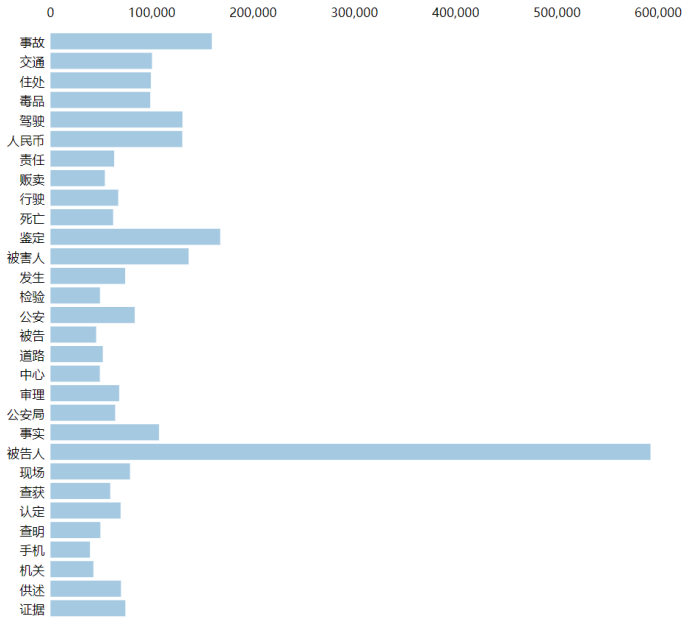


图 3-13: 主题模型的关键词统计图

在模型构建完成后，通过主题模型分析工具 pyLDAvis 库来对模型进行分析，用图像化的方式对建模结果进行展示。本文采用斯坦福大学提出的术语显著性 term saliency 公式 [39] 来计算模型中单词的重要程度。其中，w 为特定单词，T 为潜在主题，公式定义如 3-3所示。

$$\text{saliency (w)} = \text{P(w)} \times \sum_{T} P(T \mid w) \log \frac{P(T \mid w)}{P(T)} \tag{3-3}$$

主题模型的概率分布在高维空间，并不利于直观展示，BUJA 等人首先使用多维标度法 (Multidimensional Scaling) 对主题进行可视化处理 [40]。多维标度法简称 MDS，MDS 可以对高纬度数据进行降维，并在二维空间或者三维空间中进行展示。MDS 是对象集之间距离或差异的视觉表示，“对象”可以是颜色、面孔、地图坐标、政治立场，或任何真实的或概念的激励因素 [41]。图中更相似（或距离较短）的对象比不太相似（或距离较长）的对象直观上显示更加靠近。除了将集合的不同之处解释为图形上的距离之外，MDS 还可以用作高维数据的维度缩减技术。主题分布作为高维数据，很适合使用 MDS 转化成二维坐标轴上的可视化图像。通过 MDS 生成了 100 个主题之间的距离图，如图 3-14所示。

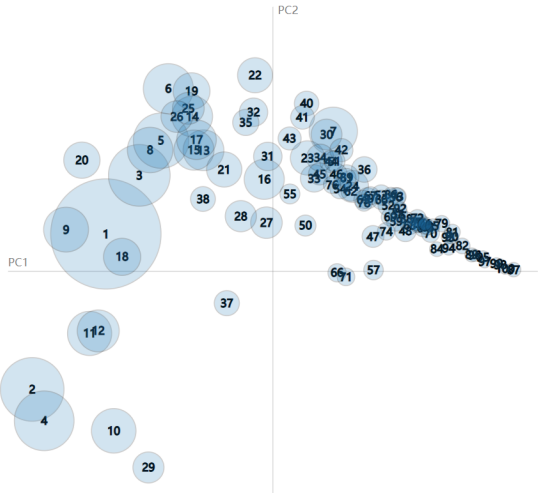


图 3-14: 基于 MDS 的主题距离分布图

通过计算字典中每个单词的 term saliency，模型中词重要性前 30 位的单词如图 3-13所示。由图可以看出，“事故”、“交通”、“毒品”等词语表示了司法裁判文书中的一些案件类别，与本文选取的数据样本集相同。这些词语具有鲜明的司法领域特征。

可以看到，100 个主题由直径大小不同的圆形来表示，直径大小和主题内单词分布的总词频成正比。也就是说，圆越大，说明这个主题越“大众”。主题按照规模半径由大到小标注为 1-100。其中，第一象限和第二象限有大量区域紧密重叠在一起，说明主题之间有关键词的重复。事实上，根据研究的经验判断，主题之间的耦合度越低，越能反映文本中的某些特定性质。具体而言，主题 6 展示如图3-15所示。

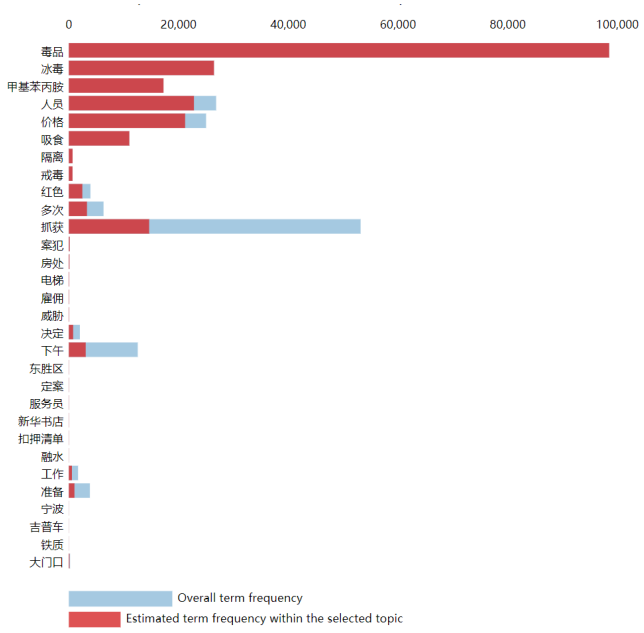


图 3-15: 主题模型 6 号主题的关键词分布图

主题 6 中，关键词按照与该主题相关度从高到低排序，“毒品”、“冰毒”、“甲基苯丙胺”等词可以将该主题初步定义为毒品案件相关。通过数据可视化，对主题表述的内容可以大体上进行理解与判断。主题模型对于大量文本的抽象归纳能力有了初步的体现。通过主题模型，可以将具有相同特征的文本进行聚类，从而筛选出特征相似的案件文本，为特征融合奠定基础。至此，已经构建完成 LDA 主题模型。

下一步，对于数据集中的 147553 条文本，依照主题模型对每个文本的主题分布进行分类，构建主题-文档表。对于数据集中每条文本，计算其文档-主题分布。

举例如下，以司法裁判文书“公诉机关指控：2017 年 4 月 30 日 18 时许，被告人吴某在广东省肇庆市端州区广百时代广场公交车站台，趁被害人梁某准备上公交车时，盗窃了梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机（经

鉴定价值人民币 1700 元）。”为待扩增文本。该文书的模型预测结果为：[(” 主题 4”, 0.038645916), (” 主题 7”, 0.09708382), (” 主题 19”, 0.07094217), (” 主题 23”, 0.17881365), (” 主题 36”, 0.038298436), (” 主题 65”, 0.037962645), (” 主题 73”, 0.07958109), (” 主题 80”, 0.04309544), (” 主题 85”, 0.047075737), (” 主题 91”, 0.038341522), (” 主题 92”, 0.21098988), (” 主题 95”, 0.08657002)]。每个元组组成为（主题，分布概率），预测结果的条形图如图 3-16所示。

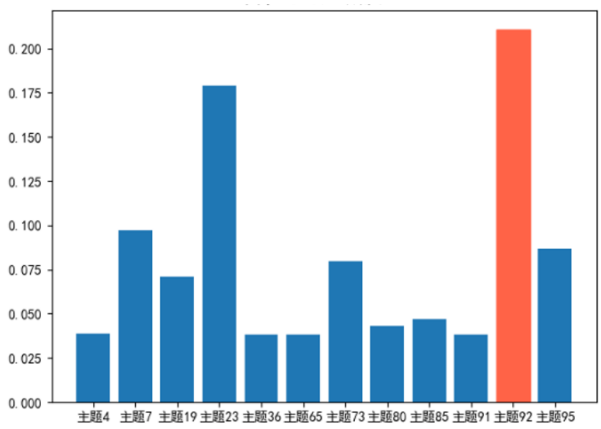


图 3-16: 待扩增文本的主题模型预测结果条形图

据图可知，待扩增文本属于主题 92 的概率最大。因此，该文本 id 的主题分布为主题 92。按照该方法，依次计算所有文本，得到每个文本最大可能的主题分布，构建主题-文档 JSON 表，将每个文本的 ID 放到其主题下，最终得到该数据集的主题-文档表。

对于待扩增文本，首先对文本进行预处理，去停用词和分词操作并加载字典文件，得到该文本的词频向量。接下来，用训练好的主题模型计算词频向量的主题分布，由图 3-16可得，该文本分布在 12 个主题之间，主题 92 的概率最大为 0.211。选取主题-文档表中主题 92 下的所有文本计算与待扩增文本的相似度。在相似度的计算中，选择余弦相似度为衡量标准，以文本的词频向量作为向量。余弦相似度就是通过计算两个向量的夹角的余弦值来评估二者的相似度。对于两个同维度的向量 A 和 B，余弦相似度计算公式如 3-4。

$$\text{similarity}(A,B) = \cos\theta = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3-4}$$

其中，A_i 和 B_i 代表 A、B 的分量。在计算过程中，对文本中的单词使用

TF-IDF 模型进行加权，以提高准确性。由于文本中每个单词分量的 TF-IDF 值都在 [0,1] 之间，所以两个文本的余弦相似度恒为正。最后，为提高计算效率，将待扩增文本与主题中所有文本的余弦相似度排序结果保存为索引文件。

根据排序结果，选择相似度前三的文本作为特征融合的目标文本。为了保证特征融合扩增时的多样性，随机选取相似度排名前三位中的某一文本。推荐文本表如3-4所示。

表 3-4: 基于主题模型的相似文本推荐表

编号	文本内容	相似度	排序
20082	江西省新建县人民检察院指控，2014 年 4 月 1 日 9 时许，被告人朱 XX 在新建县开关厂对面的 219 路公交车站台，趁失主周 X 上 219 路公交车时尾随其后，用右手将周 X 上衣右边口袋的白色步步高手机偷走。被告人朱某得手后准备逃离时被巡逻民警追上抓获，当场查获被盗窃的 VIVO20T 手机一部，经价格认证部门价格鉴定，该手机价值人民币 2231 元。案发后，公安机关将当场收缴的手机发还给了失主周 X。	0.862241	1
29580	江油市人民检察院指控：2015 年 12 月 14 日 17 时许，被告人王某某行至江油市太平路东段建兴市场公交站，趁正在上公交车的被害人李某某不备之际，盗窃了其放置于上衣右边包内的白色红米 NOTE 手机一部。后在逃离过程中，王某某被民警抓获，民警当场从其身上搜出被盗的红米 NOTE 手机一部。经鉴定，被盗手机价值 520 元。案发后，被盗手机已发还被害人李某某。	0.851054	2
15078	公诉机关指控，2015 年 11 月 29 日 17 时许，被告人曾某乘坐 208 路公交车，当车行驶至沙坪坝区新桥正街附近时，其趁乘客唐某某不备之机，扒窃唐某某裤子口袋内的现金 310 元，后被当场抓获。公安机关已追回被盗现金发还给唐某某。	0.836014	3

至此，根据主题模型已经得到待扩增文本的相似推荐文本。待扩增文本为盗窃案件，地点发生在公交站台，由推荐文本表可知，这三个司法裁判文书均为盗窃案件且地点均在公交车或公交站台，第一个和第二个案件盗窃物品与待扩增文本相同，均为手机。因此，根据主题模型可以筛选出与待扩增文本特征

相似的目标文本，且准确性很高。文本筛选代码如图3-17所示。

```
def getRecommend(lda, dict, tfidf, dict_corpora, text):
    vector = lda[dict.doc2bow(text)]
    #文本所在主题
    topic = sorted(vector, key=lambda item: -item[1])
                                                    [0:1][0]
    #获取主题文档对应表
    dict = json.loads(getTopicDocument())
    #需要对比的文档
    search_corpus = []
    index_num = 0
    index_dict = {}
    for i in dict[str(topic)]:
        search_corpus.append(tfidf[dict_corpora][i-1])
        index_dict[index_num] = i
        index_num += 1
    index = similarities.MatrixSimilarity
                                (lda[search_corpus])
    #计算相似度
    sims = index[vector]
    #相似度排序
    sorted_sims = sorted(enumerate(sims), key=lambda
                                        item: -item[1])
    #获取文档编号-文本表
    df = getDocument()
    count = 0
    recomm = []     #返回的相似文本表
    for doc in sorted_sims:
    recomm.append(index_dict[doc[0]])
        count += 1
        if count == 3:
            break
    return recomm
```

图 3-17: 基于主题模型的相似文本筛选核心代码

本次扩增选择相似度最高的 ID 为 20082 的文本作为目标文本。扩增时的特征融合以两个文本的依存句法关系为依据。

首先使用 LTP 工具得到待扩增文本和目标文本的依存句法关系，找到两个文本中具有相同依存关系的树枝，如：待扩增文本中“公诉机关”与“指控”之间是 SBV（主谓）关系，目标文本中“江西省新建县人民检察院”与“指

控”也是 SBV（主谓）关系；待扩增文本中“2017 年 4 月 30 日 18 时许”与“盗窃”是 ADV(壮中) 关系，目标文本中“2014 年 4 月 1 日 9 时许”与“尾随”也是 ADV（壮中）关系。

在特征融合中，将目标文本中与待扩增文本中依存关系相同的树枝进行替换，实现在保证文本语义语法不变的情况下，文本内容的改变，得到新的扩增后的文本。

在特征融合扩增时，融合原则是以树枝为特征融合的基本粒度，以依存关系作为融合的依据，因此，需要找到两个文本的具有相同依存关系的树枝进行交换。

依存句法树是树形结构，第一层是一个核心词指向 Root 结点，两者是 HED 关系，第二层的结点数相对于树的全部结点数来说较少，部分文本第二层只有两三个结点，如果以此为特征选取粒度，则将深度改变待扩增文本的结构，实现不了特征融合扩增的目的。如果将两个文本中所有具有相同依存关系的树枝（结点）任意搭配，则会出现一个长句被一个词替换的情况，如待扩增文本中“梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机”与“盗窃”是 VOB（动宾）关系，而目标文本中“其后”与“尾随”也是 VOB（动宾）关系，如待扩增文本中“梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机”被“其后”替换，将会破坏文本的语义信息，关键特征被消除。

为保证特征选取的效果而又不会影响文本的语义和结构，在两个文本的固定层数第三层中进行依存关系配对。得到两个文本的依存句法树中第三层的每个树枝与第二层的依存关系，把具有相同依存关系的树枝配对，作为特征融合的备选。

在司法数据集中，司法裁判文书内容较长，所含信息较为丰富，依存句法树结构较复杂且结构相似，为保证扩增前后文本内容长度尽可能不发生较大改变，因此在特征融合时选择在固定层数的树枝上进行树枝配对。如在其他领域的数据集上进行特征融合扩增，可以根据其领域文本的内容、结构、长度等选择在固定层数上配对或者在句法树范围内配对。

在进行依存关系配对的过程中，首先是对目标文本进行分词和依存句法分析操作，然后依次遍历 ROOT 结点，第一层结点、第二层结点和第三层结点，最后得到第三层结点与第二层结点的依存关系。其中，“WP”关系代表标点，在文本中标点的存在与文本特征关系甚微，因此，在遍历依存关系时将去除“WP”关系。遍历第二层和第三层结点代码如图3-18所示。

```
#dep: 依存关系表
def getDep(dep):
    pretuple = dep1[0]
    first = 0
    second = []
    threes = []
    for tu in pretuple:
        if tu[2] == "HED":
            first1 = tu[0]
    for tu in pretuple:
        if tu[1] == first and tu[2] != "WP":
            second1.append((tu[0], tu[2]))
    for i in second:
        three = []
        for j in pretuple:
            if j[1] == i[0] and j[2] != "WP":
                three.append((j[0],j[2]))
        threes.append(three)
    return second,threes
```

图 3-18: 依存句法树树枝结点层次遍历核心代码

遍历出待扩增文本和目标文本的第三层结点后，将两个文本所有具有相同依存关系的结点依次配对，得到特征融合的备选元组集（待扩增文本结点，目标文本结点）。

在元组的选择中存在待扩增文本中结点选择重复的问题。例如，对于元组集 [(1, 13), (1,18), (6,26),(8,30)]，随机选择 2 个元组，如果选到元组集中 (1, 13) 和 (1,18) 两个元组，在进行树枝结点替换时，1 号树枝结点依次被 13 号和 18 号树枝替换，但该树枝最终只能被一个树枝结点替换，最终只相当于被替换一次，而且在交换过程中容易出现冲突。因此，在随机选择时先将元组集按照待扩增文本的结点编号进行聚类，随机选择类别，如果某一类中有多个元组，则再随机选取类别中某一元组。

表 3-5: 基于主题模型的特征融合扩增方法参数表

参数	默认值	含义
quantityWeight	0.4	选择元组类别数量占备选元组集类别数量的比重

在特征融合扩增时，按待扩增文本树枝编号分类后的元组集数量较多，在扩增时设定权重参数，随机选择备选元组集类别数量中 length*quantityWeight 个类别，如果选中的类别中存在多个原则，则再随机选取一个元组。quantity-Weight 默认初始值为 0.4。参数表如表 3-8所示。

最后，根据选择的元组替换相应的树枝结点，最终扩增后的结果如表 3-8所示。扩增后文本的依存句法树如图 3-19所示。可以看到，待扩增文本中添加了目标文本的特征。

表 3-6: 特征融合扩增方法待扩增示例文本扩增结果对照表

初始文本	目标文本	扩增后
公诉机关指控：2017 年 4 月 30 日 18 时许，被告人吴某在广东省肇庆市端州区广百时代广场公交车站台，趁被害人梁某准备上公交车时，盗窃了梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机（经鉴定价值人民币 1700 元）	江西省新建县人民检察院指控，2014 年 4 月 1 日 9 时许，被告人朱 XX 在新建县开关厂对面的 219 路公交车站台，趁失主周 X 上 219 路公交车时尾随其后，用右手将周 X 上衣右边口袋的白色步步高手机偷走。被告人朱某得手后准备逃离时被巡逻民警追上抓获，当场查获被盗窃的 VIVO20T 手机一部，经价格认证部门价格鉴定，该手机价值人民币 2231 元。案发后，公安机关将当场收缴的手机发还给了失主周 X。	公诉机关指控：2017 年 4 月 30 日 18 时许，被告人吴某在新建县开关厂对面的 219 路公交车站台，趁失主周 X 上 219 路公交车时盗窃了梁某左裤袋中一台玫红色 OPPO 牌 A59S 型手机（经鉴定价值人民币 1700 元）

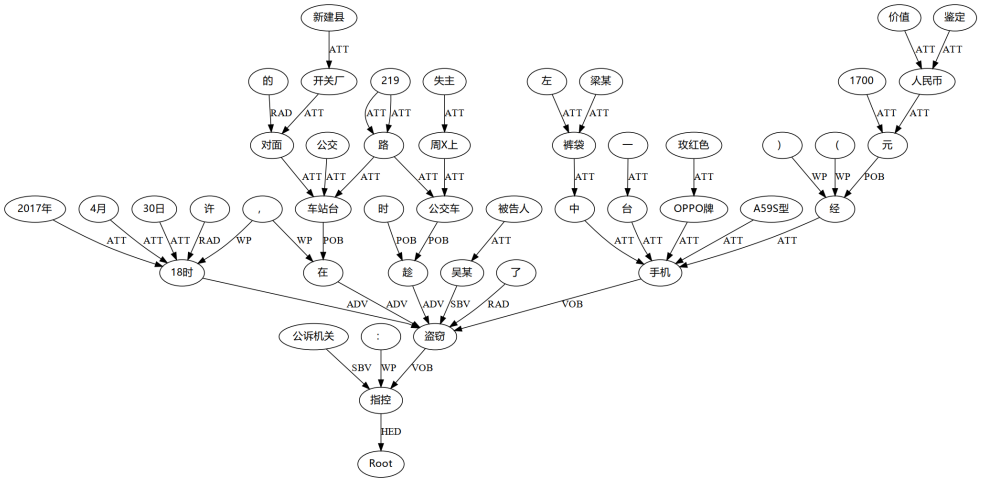


图 3-19: 特征融合扩增方法待扩增示例文本扩增结果依存句法树

3.4 基于依存句法的特征变换

基于依存句法的特征变换文本扩增是一种不依赖于文本所在领域数据集的扩增方法，该方法以文本的依存句法关系为特征变换依据，在文本内容不变的情况下，改变文本的语序结构进行扩增，以保持其领域特征不变。

特征变换（Feature Transaction，FT）扩增方法步骤为：对于包含 n 个文本的数据集 $D = \{d_1, d_2 \cdots d_n\}$，对于待扩增文本 d_i ($d_i \in D$) 进行依存句法分析，得到文本 d_i 的依存句法树 T_i，将 T_i 中符合长度要求的树枝筛选出来并得到树枝根结点与其父结点的依存关系，将存在包含关系的树枝进行合并，对于筛选出来的树枝按照依存关系进行匹配，组成待选树枝对集 TT_i，在扩增时，随机选择 TT_i 中的树枝对进行交换，得到扩增后文本 d_i'。扩增流程图3-20如下所示。

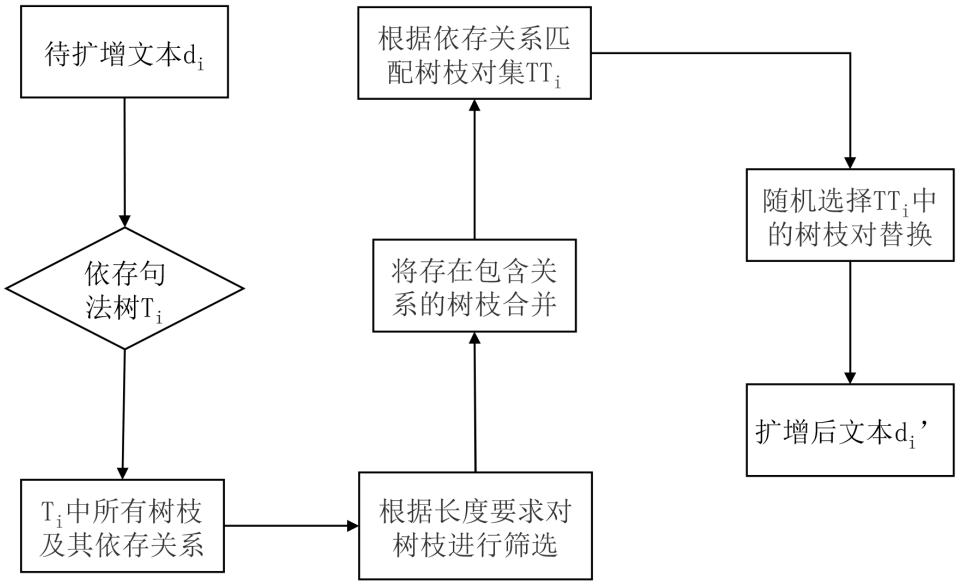


图 3-20: 基于依存句法的特征变换扩增方法流程图

基于依存句法的特征变换扩增方法与特征裁剪和特征融合扩增方法不同，其不依赖于文本所在的数据集，不在数据集的尺度进行特征挖掘，而是在文本的尺度中不改变句子依存关系的情况下进行语序结构的调整，保持文本的基本特征和语义信息。

以司法裁判文书数据集为例。设“上海市静安区人民检察院指控被告人闫某于 2014 年 11 月 9 日 2 时许，在本市静安区延安西路 XXX 号上海戏剧学院门口，采用撬锁的方式，窃得被害人尼古某某停放于此的白色莫曼顿牌 iRide600FS 型自行车一辆，在逃逸过程中被公安人员人赃俱获。”作为为待扩

增文本。在进行扩增的第一步是对文本进行依存句法分析，得到文本的依存句法树。如图3-21所示。

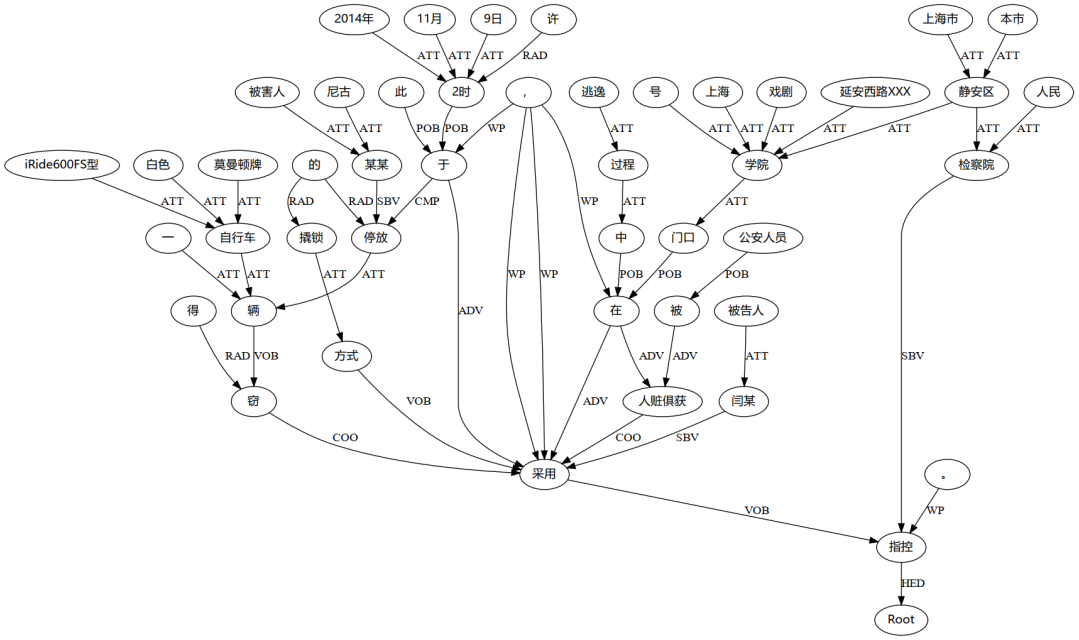


图 3-21: 基于依存句法的特征变换待扩增示例文本依存句法树

接下来是筛选出符合长度要求的树枝。首先是得到依存句法树中所有树枝（包括父结点及其所有子结点），在这里设置长度选择权重 lengthweight，将树枝中分词数量小于等于分词个数 *lengthweight 的所有树枝筛选出来。在树枝筛选时，对于叶子结点来说，相比于其所在的树枝，体现的文本特征内容更少，因此，将树枝筛选长度设为大于 1。

```
#筛选树枝
def getAllSon(seg, dep, lenArg):
    treelist = []  # 所有的枝
    length = len(seg)  # 所有分词数量
    for i in range(len(seg)):
        list = getSon(dep, i) # 获取其所有子节点
        if len(list) <= length * lenArg and len(list) > 1:
            treelist.append(list)
    treelist = getCollection(treelist)
    return treelist
```

图 3-22: 依存句法树树枝筛选核心代码

筛选符合长度要求的树枝后得到备选的树枝集合，但是存在结点重复的问题。例如，得到树枝集合：[[1,2],[3,1,2],[8,9],[13,11,12,10]……]，该集合中，存在 [1,2] 和 [3,1,2] 两个树枝，两者都包含 1、2 结点且存在包含关系。如果在特征变换时选中该组，则两个树枝无法进行交换。因此，在筛选出树枝后将存在包含关系的树枝进行合并。例如，将 [1,2] 和 [3,1,2] 合并为 [3,1,2]。筛选和合并树枝代码如图3-22和3-23所示。

```
#将存在包含关系的树枝合并
def getCollection(lists):
    list = lists[:]
    for m in lists:
        for n in lists:
            if set(m).issubset(set(n)) and m != n:
                list.remove(m)
                break
    return list
```

图 3-23: 依存句法树包含关系树枝合并核心代码

对于合并后的树枝集合，得到每个树枝与其父结点的依存关系，将具有相同依存关系的树枝配对，作为下一步进行特征变换的树枝配对待选集。在待选集中选择需要进行交换的树枝时，确定配对数量时设定权重参数 selectweight，将配对集长度 length*selectweight 作为选择的配对元组数量，在配对集合中随机选择。参数表如3-8所示。

表 3-7: 基于依存句法的特征变换扩增方法参数表

参数	默认值	含义
lengthweight	0.2	树枝长度范围参数
selectweight	0.4	配对集数量选择参数

在选择树枝配对集时，和特征融合扩增一样，同样存在树枝重复选择的问题。例如，树枝配对元组集合 [(0, 1), (0, 5),(1,5),(3,7),(8,9)]，将 selectweight 参数设为 0.4，也就是选择两个配对元组。如果在随机选择的情况下，选择了 (0, 5) 和 (1, 5) 两个元组，则 0 号树枝与 5 号树枝交换，1 号树枝也与 5 号树枝交换，导致 5 号树枝出现的位置依次被 0 号和 1 号替换，5 号树枝在文本中出现

两次，不符合特征变换时结构变换内容不变的要求。如果选择了 (0,1) 和 (0,5) 两个元组结果一样。为避免这个问题，在元组选择时，已经被选择的树枝在接下来的选择中避免再次出现。选择代码如图3-24所示。

```
#slice: 待选择配对集合
#length: 选择配对元组数量
def getDifTree(slice, length):
    sli = []
    num = []
    for i in range(len(slice)):
        s = random.choice(slice)
        if s[0] not in num and s[1] not in num:
            sli.append(s)
            num.append(s[0])
            num.append(s[1])
        if len(sli) == length:
            break
    return sli
```

图 3-24: 获取包含不同结点树枝配对元组核心代码

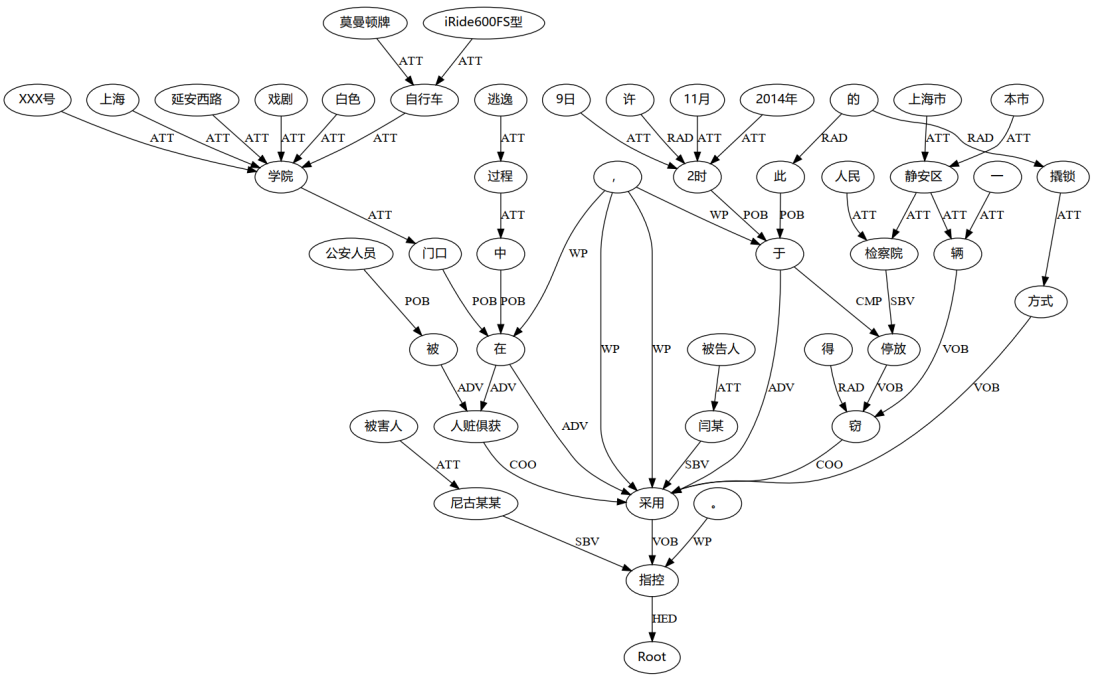


图 3-25: 特征变换扩增方法待扩增示例文本扩增结果依存句法树

随机选择树枝配对元组后，将待扩增文本中相应的树枝进行替换即可。对于待扩增文本，经过特征变换扩增后。其扩增后文本表如表3-8所示，其依存句法树如图3-25所示。

表 3-8: 特征变换扩增方法待扩增示例文本扩增结果对照表

初始文本	扩增文本
上海市静安区人民检察院指控被告人闫某于 2014 年 11 月 9 日 2 时许，在本市静安区延安西路 XXX 号上海戏剧学院门口，采用撬锁的方式，窃得被害人尼古某某停放于此的白色莫曼顿牌 iRide600FS 型自行车一辆，在逃逸过程中被公安人员人赃俱获。	被害人尼古某某指控被告人闫某于 2014 年 11 月 9 日 2 时许，在白色莫曼顿牌 iRide600FS 型自行车延安西路 XXX 号上海戏剧学院门口，采用撬锁的方式，窃得上海市静安区人民检察院停放于此的本市静安区一辆，被公安人员在逃逸过程中人赃俱获。

3.5 基于词频词性的特征替换

基于词频词性的特征替换扩增方法是一种将文本分词后的词性标注结果和数据集中词频统计后的高频词表作为特征替换的依据，并根据在数据集中训练出的词向量模型来替换相应的词语，实现文本扩增的方法。

特征替换（Feature Replacement，FR）扩增方法步骤为：对于包含 n 个文本的数据集 $D = \{d_1, d_2 \cdots d_n\}$，分词后统计得出词频记录 WF，并在数据集 D 上训练词向量模型 WM。对于待扩增文本 d_i ($d_i \in D$) 进行分词和词性标注，如果一个词在 WF 中属于高频词语且词性能够体现领域特征则被选为待替换词语，构建文本待替换词语集 WL。在 WL 中随机选择若干词语使用模型 WM 得到其近似词语进行替换，最终得到扩增后文本 d_i'。扩增流程图如3-26所示。

基于词频词性的特征替换扩增方法依赖于文本所在的数据集，需要使用数据集计算词频和训练词向量。以司法裁判文书数据集为例。在文本预处理阶段，已经得到该数据集的词频统计结果，依照词云图可以看出，词频较高的词语可以很好的反映出文本的领域特征，相对而言，词频较低的词语重要性更低，不能很好反映文本的领域特征。

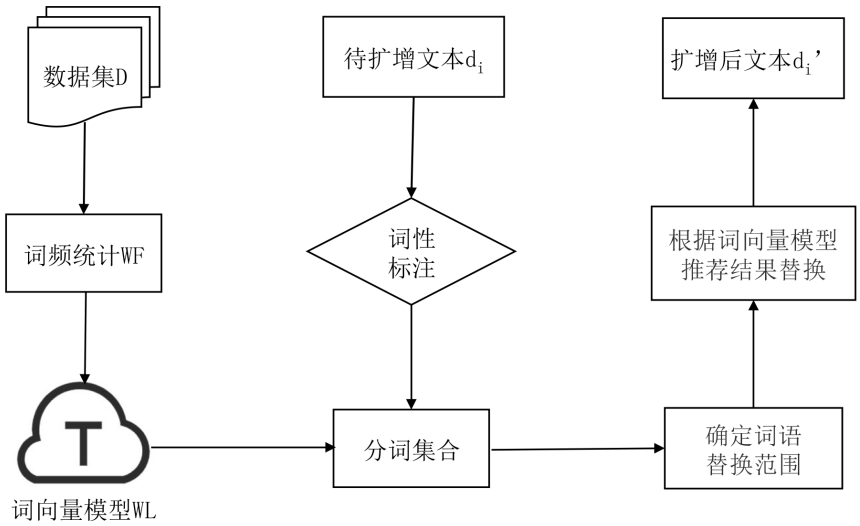


图 3-26: 基于词频词性的特征替换扩增方法流程图

经统计，司法裁判文书数据集中词语个数接近 26 万，词语总频率超过 1400 万。由数据集词频统计图3-27可得。词频大于 1000 的词语的总频率占所有词语频率的 82%，而词语个数只有不到 1500 个。观察可得，词频小于 100 的词语一般是地名等. 因此，本次实验将词频大于 1000 的词语设为高频词语，构建高频词表。

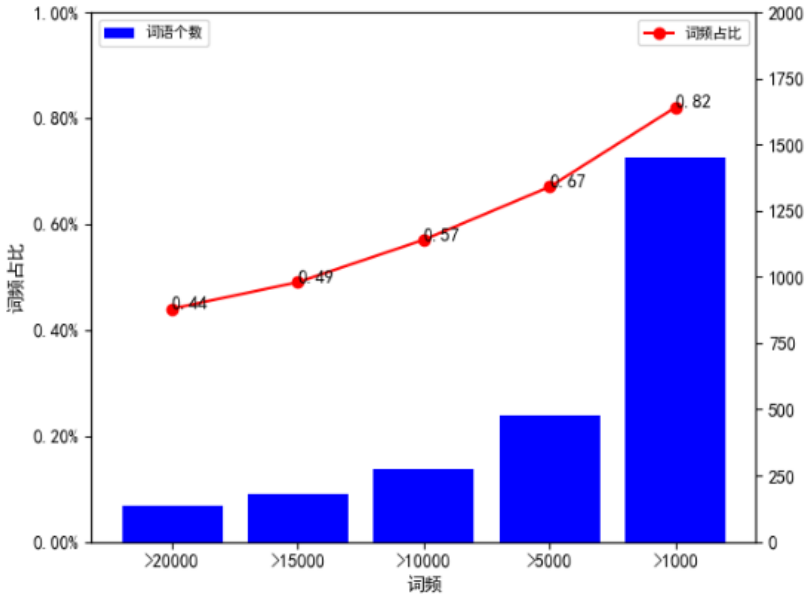


图 3-27: 司法裁判文书数据集词频统计图

构建高频词表后需要训练词向量模型。词向量模型可以计算词语之间的相

似度，并且可以给出某个词语的相似词语。为保持领域特征，词向量模型须在文本所在的领域数据集上进行构建。这样，词向量推荐的相似词语会符合领域特征要求。对于数据量较小的数据集，难以训练出的高质量的领域词向量模型，对此，可以采用开源的近义词工具来代替词向量模型进行近义词推荐。

词向量模型的构建使用 gensim 库的 Word2vec 方法，在构建之前，要对数据集进行去停用词和分词处理，这些操作在文本预处理阶段已经完成。传递参数时设定 sg=0，表示使用 CBOW 算法进行构建，CBOW 算法在小型数据集上表现较好。size=200 表示输出的词的向量维数为 200，维数越多，刻画词语特征越精准。window=5 表示训练窗口大小，默认为 5，最后选择训练并发数，保存模型文件。训练代码如图 3-28所示。

```
#训练词向量
def buildWord2Vector():
    model = Word2Vec(LineSentence("stop_seg.txt"),
                     sg=0, size=200, window=5, min_count=5)
    model.save("top_seg.model")
    model.wv.save_word2vec_format("stop_seg.vector",
                                  binary=False)
```

图 3-28: 词向量训练核心代码

完成高频词表和词向量模型的前期准备后，接下来需要对文本进行处理。以司法裁判文书”经审理查明：被告人向某于 2014 年 2 月 8 日下午 15 时 30 分许在本市江干区七堡中心路某饭店上班期间，趁人不备，窃得店主姜某放在吧台抽屉内的钱包里的现金人民币 4956 元，后在逃离途中被抓获。赃款已追回并发还给被害人。”为待扩增文本，进行特征替换扩增。

首先对该文本进行分词和词性标注。词性标注是将文本中的词语按照其含义和上下文内容进行标记的文本处理技术。文本由不同词性的词语组成，不同词性的词语对于文本作用也不同，有形容词、连词、标点、副词等。对于标点、数字、助词、叹词等没有具体含义的词性词语，其不能鲜明反映文本的领域特征。

根据 863 词性标注集中 28 种词性标签进行梳理，重点将能够鲜明体现领域特征的形容词、副词、名词等词性标签的词语筛选出来。词性标签表格如表 3-9所示。在文本处理中，如果词语的词性为表中的某个词性，则作为特征替换备选词。

表 3-9: 特征替换扩增方法词性选择表

词性标签	描述	举例	词性标签	描述	举例
a	形容词	最	nz	专有名词	汉语
b	区别词	副	ns	地名	杭州
d	副词	非常	nh	人名	张某
i	成语	义无反顾	v	动词	举办
n	一般名词	城市	j	缩写	文教

依据特征替换扩增的规则，只有某个词语属于高频词并且其词性在特征替换词性表中才可以被选入特征替换的备选词列表，对备选词列表中的词语使用训练好的词向量模型进行相似词语推荐替换。如图3-29所示，依存句法关系图中红色字体词语为待扩增文本中的所有备选词语。可以看到，备选词语大都是具有实际意义的词语，是文本中能够反映领域特征的主要词语。

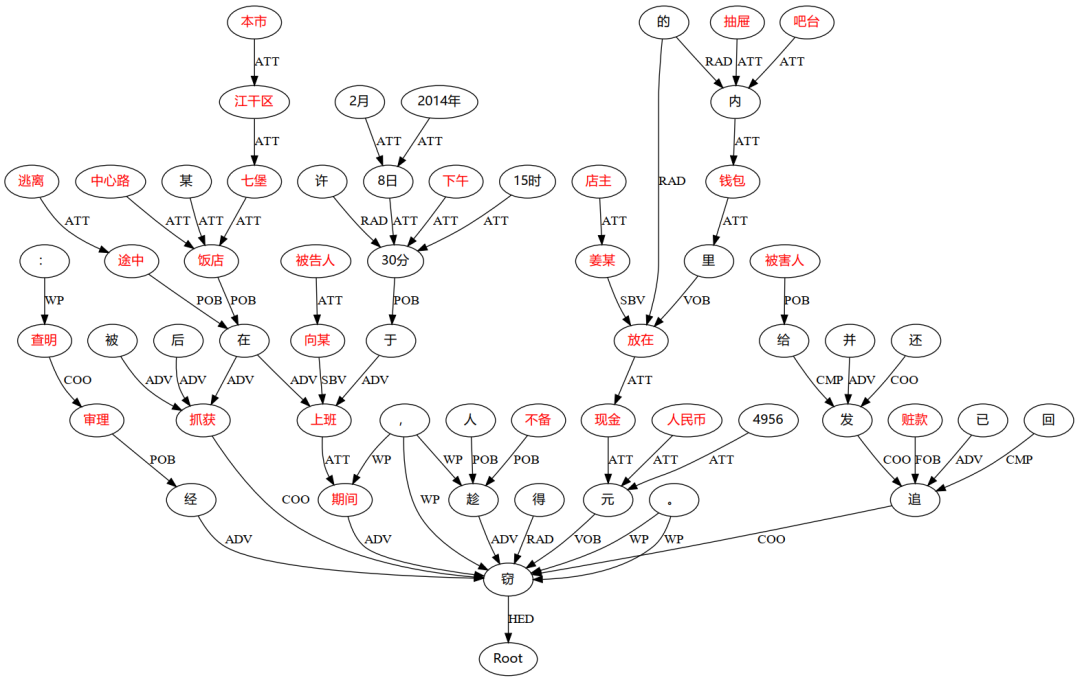


图 3-29: 基于词频词性的特征替换方法待扩增示例文本依存句法树

在选择替换词语时，设权重参数 replaceweight，默认值为 0.4，则选取的替换词语数量为备选词列表长度 length*replaceweight。参数表如 3-10所示。

表 3-10: 基于词频词性的特征替换扩增方法参数表

参数	默认值	含义
replaceweight	0.4	备选词数量选择参数

选取替换备选词语后，根据词向量模型的相似词语推荐结果，在前 5 个近似词语结果中随机选取并替换。替换结果如表 3-11所示。

表 3-11: 特征替换扩增方法待扩增示例文本扩增结果对照表

初始文本	扩增文本
经审理**查明**：被告人**向某**于 2014 年 2 月 8 日下午 15 时 30 分许在本市江干区七堡中心路某饭店上班期间，趁人不备，窃得店主**姜某**放在**吧台**抽屉内的钱包里的现金**人民币** 4956 元，后在逃离**途中**被抓获。赃款已追回并发还给**被害人**。	经审理**审明**：被告人**饶某**于 2014 年 2 月 8 日晚上 15 时 30 分许在本市江干区七堡园林路某饭店上班期间，趁人不备，窃得店主**孟某**放在**收银**台抽屉内的钱包里的现金**余** 4956 元，后在逃离**随即**被抓获。赃款已追回并发还给**被害**。

3.6 本章小结

本章详细介绍了基于领域特征的文本数据扩增技术的四种方法和基本步骤，分别为基于 TF-IDF 权重的特征裁剪方法、基于主题模型的特征融合方法、基于依存句法的特征变换方法和基于词频词性的特征替换方法。基于领域特征的文本数据扩增技术的四个扩增方法中有三个需要基于数据集生成的模型，挖掘数据集中的领域特征为数据扩增奠定基础。

文本预处理是为了在文本扩增时节约处理时间和提高计算效率，对数据集数据进行结构化、分词、去停用词处理，并进行词频统计。在后续进行训练模型时需要多次进行分词和依存句法分析处理，这会耗费大量的空间和时间资

源，因此预先对数据集的重复耗时操作进行处理，将结果保存下来，节省后续处理时间。

基于 TF-IDF 权重的特征裁剪方法是根据数据集训练出 TF-IDF 模型，将文本中的词语用 TF-IDF 值区分开来，TF-IDF 值大的词语可以较好反映文本的特征，也反映了词语对于文本的重要程度，TF-IDF 值相对较小的词语对于代表文本中的特征无足轻重，一般是停用词等。该加权方法结合文本的依存句法树，计算树枝总的 TF-IDF 值，将文本的筛选粒度从词语扩大到树枝代表的短语、短句，将 IF-IDF 值较小的树枝随机裁剪，裁剪树枝后的文本在保留其基本特征的前提下更加精炼，生成新的扩增文本。

基于主题模型的特征融合方法是根据数据集训练出主题模型，筛选出数据集中与待扩增文本相似的文本，依据两个文本中的依存关系，替换树枝，达到特征融合扩增的目的。主题模型一般用于文本分类，在构建主题模型时根据困惑度确定主题数目，得到质量较高的主题模型，根据主题模型预测文本所在主题，再计算主题下的文本与待扩增文本余弦相似度，筛选出相似度较高的文本，依据文本中树枝的依存关系进行相应替换。该方法使用主题模型可以保证在扩增时文本中树枝来源于与文本特征相似的文本，保证扩增后文本特征不发生较大变化。

表 3-12: 基于领域特征扩增方法介绍简表

扩增方法	特征裁剪	特征融合	特征变换	特征替换
英文缩写	FC(Feature Clipping)	FF(Feature fusion)	FT(Feature Transaction)	FR(Feature Replacement)
技术理论	TF-IDF 权重	主题模型	依存句法分析	词性标注
是否依赖数据集	是	是	否	是
是否依赖模型	是	是	否	是
是否使用依存句法	是	是	是	否
是否使用词性标注	否	否	否	是
参数个数	3	1	2	1

基于依存句法的特征变换方法不依赖于数据集，在文本中依存关系相同的树枝之间进行变换，文本中的内容不发生改变，而在语序上进行变化。

基于词频词性的特征替换方法是依据数据集中的高频词和能够反映文本特征的词性，筛选出文本中的特征词语，使用在数据集中训练出的词向量模型推荐的相似词语对特征词语进行替换，达到不改变文本语义的情况下，实现文本的扩增。四种方法的介绍表如3-12所示。

上述四种方法基于文本所在领域的特征使用不同的技术、从不同维度进行扩增，扩增步骤简单、扩增思路清晰，而且在扩增时可以选择不同的参数提高扩增文本的多样性和扩增粒度。

第四章　基于领域特征的扩增数据生成及实验评估

本章主要通过基于不同领域数据集进行文本数据扩增，设计实验验证基于领域特征的文本数据扩增技术的有效性。在选择领域数据集时，本文选择了司法领域和媒体领域的开源数据集，通过训练高质量的文本分类模型来评估扩增后文本是否保持原始类别标签和领域特征，最后通过扩增后的文本数据集训练文本分类模型，评估该技术对模型性能提升的效果。

4.1　扩增数据标签评估

在进行文本数据扩增时，一个基本的问题是扩增后的文本是否还保留原有的标签。文本的原始标签表明了该文本的领域类别和领域特征性。扩增后文本的标签是否与其原始句子的标签一致，是检验扩增方法是否保留其原有领域特征、扩增结果是否有效的基本标志。

然而，在扩增时如果句子有明显的变化，那么原始的标签可能不再有效，扩增后文本不符合原有的领域特征，其类别标签可能就发生改变，出现这种情况意味着扩增文本的质量不高。为检验基于领域特征的文本数据扩增技术的四种扩增方法在进行文本数据扩增时是否保留其原有的标签，在本节设计文本分类实验，构建文本分类模型，来验证扩增后的文本是否还保留其原有的类别特征。

本次标签评估实验使用司法裁判文书数据集。首先，从盗窃、危险驾驶、故意伤害、交通肇事和走私、贩卖、运输、制造毒品五个类别数据集中分别抽取 15000 条文本共 75000 条数据作为训练集，分别抽取 1000 条文本共 5000 条数据作为验证集，分别抽取 1500 条文本共 7500 条数据作为测试集，在此基础上训练一个文本分类任务的 CNN 模型，在该数据集上为罪名预测模型。

该模型的任务是对这五种案件类型的司法裁判文书进行分类，模型训练

完成后在测试集上测试该 CNN 模型的性能表现。经过在测试集上的运行，该模型在测试集上表现优异，整体准确率达到 98.54%。模型评价指标精确率（precision）、召回率（recall）和 F1 值（f1-score）矩阵如下表 4-1所示。

由此表可知，该模型性能优异、质量较高，在类别标签预测上误差较小。将扩增后的文本作为测试集，使用该模型在测试集上进行文本分类实验，可以在检验扩增后的文本标签是否发生改变，是否仍然保持着文本的领域特征。

表 4-1: 司法领域分类模型评估指标结果表

案件类型	precision	recall	f1-score
盗窃	0.98	0.99	0.98
危险驾驶	0.99	0.98	0.98
故意伤害	0.99	0.99	0.99
交通肇事	0.99	0.98	0.98
走私、贩卖、运输、制造毒品	0.99	0.98	0.98

模型构建完成后，选取部分数据进行标签评估实验。本次实验所用的数据是从司法裁判文书数据集中随机抽取未被用做训练集和测试集的文本数据，每种案件类型各 100 个，共 500 个原始文本数据。这些文本分别经过特征裁剪（FC）、特征融合 (FM)、特征变换（FT）和特征替换（FR）扩增后得到 2000 个文本作为测试集 testFeature。为更加直观的检验模型在测试集上的表现，实验设置对照组，将这 500 个原始文本数据分别使用 EDA 技术中同义词替换（SR）、随机插入（RI）、随机交换（RS）和随机删除（RD）对文本进行扩增后得到的 2000 个文本作为测试集 testEDA。在进行特征扩增和 EDA 扩增时均使用默认参数，两个测试集中的文本数据全部是原始文本数据扩增后得到的，不包含原始文本数据。

模型在这两个测试集 testFeature 和 testEDA 上运行后，得到模型在测试集 testFeature 上的准确率为 93.38%，在测试集 testEDA 上的准确率为 91.70%。由此可知，模型在基于领域特征扩增的测试集上表现更好，这在一定程度上反映了基于领域特征扩增技术得到的文本能够保持文本类别特征，并且比通用的 EDA 技术在保持文本标签方面表现更好。

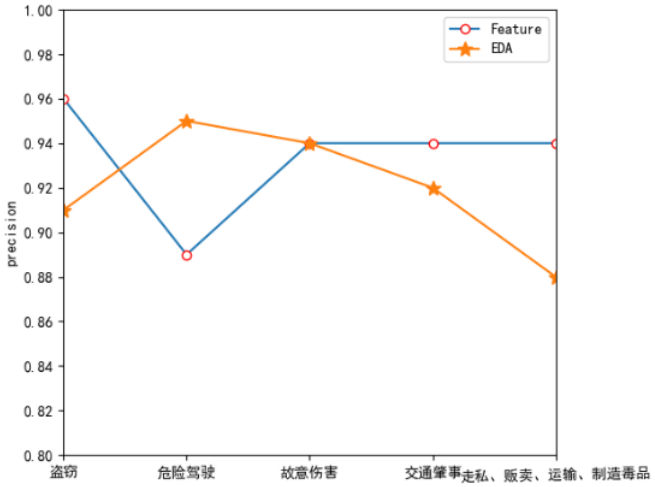


图 4-1: 模型在测试集 testFeature 和测试集 testEDA 上的精确率

准确率（Accuracy）指标并不总能有效评价一个分类模型的性能。精确率（Precision）是衡量模型正确预测为某个类别的比例占全部预测为该类别的比例。召回率（Recall）是衡量模型正确预测为某个类别的比例占全部实际为该类别的比例。在两个测试集上的精确率和召回率结果如图4-1和4-2所示。

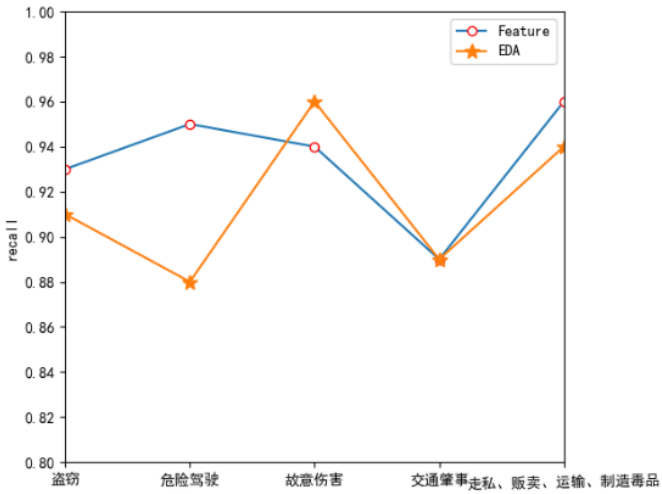


图 4-2: 分类模型在测试集 testFeature 和测试集 testEDA 上的召回率

分类模型在测试集 testFeature 和 testEDA 的精确率和召回率结果由图可得，模型在测试集 testFeature 上盗窃、交通肇事和毒品类案件的精确率大于测试集 testEDA，“故意伤害”类别的精确率两者相等。模型在测试集 testFeature 上盗窃、危险驾驶、毒品类案件的召回率大于测试集 testEDA，“交通肇事”

类别的召回率两者相等。测试集 testFeature 中有三个类别的精确率和召回率比测试集 testFeature 高，有一个类别相同。

这可以反映出测试集 testFeature 的质量更高。精确率和召回率是矛盾关系，两个指标是此消彼长的，在衡量时应该综合考虑，也就有了 F 值。F 值是精确率和召回率的加权调和平均，如公式 4-1所示。当 a=1 时，就是最常见的 F1 值。F1 综合考虑了精确率和召回率的结果，当 F1 较高时则能说明模型比较有效，数据质量也较高。

$$\mathrm{F} = \frac{\left(\alpha^2 + 1\right) P * R}{\alpha^2 (P + R)} \tag{4-1}$$

综合加权计算后，得到该模型在两个数据集上的 F1 值，如图 4-3所示。

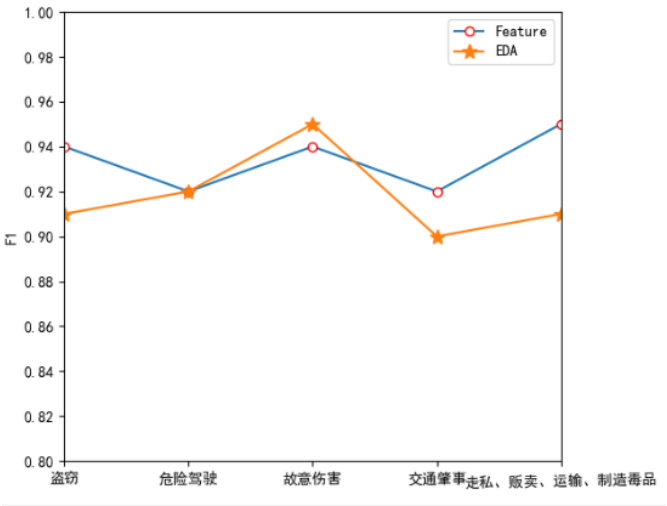


图 4-3: 模型在测试集 testFeature 和测试集 testEDA 上的 F1 值

由图可得，分类模型在测试集 testFeature 五个类别中有三个类别的 F1 值明显大于测试集 testEDA，“危险驾驶”类别两者的 F1 值相同，只有“故意伤害”类别测试集 testEDA 的 F1 值微大于测试集 testFeature。

综上，模型的各项评估指标在测试集 testFeature 上更优。这也说明了基于领域特征扩增方法得到文本数据比通用的 EDA 技术扩增得到的文本数据更能保持其领域特征和类别标签。因此，基于领域特征的文本数据扩增方法在保持扩增文本标签和领域特征方面是有效的，使用该方法扩增得到的数据质量较高。

4.2 扩增实验数据准备

在第三章已经选取了司法领域中的司法裁判文书开源数据集进行扩增，为验证基于领域特征的文本数据扩增技术的有效性，本文选取其他领域的文本数据进行扩增实验。本节使用 THUCNews 数据集 [42] 进行扩增。THUCNews 中文文本分类数据集是由新浪新闻 RSS 订阅频道在 2005-2011 年间对历史数据进行过滤后生成的，采用 UTF-8 纯文本格式。这一数据集包含 74 万份新闻文档，共有 14 个候选分类，例如：金融、房地产、证券、教育等等。本文选取其中五个类别（体育、财经、教育、科技、时政）各 6500 条共 32500 条数据作为本次媒体领域扩增的基础数据集。

为让数据扩增时符合其所在领域基本特征，按照基于领域特征文本扩增技术的数据集处理一般流程，首先对其数据集进行预处理，进行分词和高频词语统计等，然后预训练出扩增所需的 TF-IDF 模型、LDA 主题模型和词向量模型，并在分词操作中注意分词的精准性，处理时添加开源的领域词典和停用词表。对于数据集的处理遵循以下步骤，如图4-4所示的基于领域特征文本数据扩增方法一般步骤。

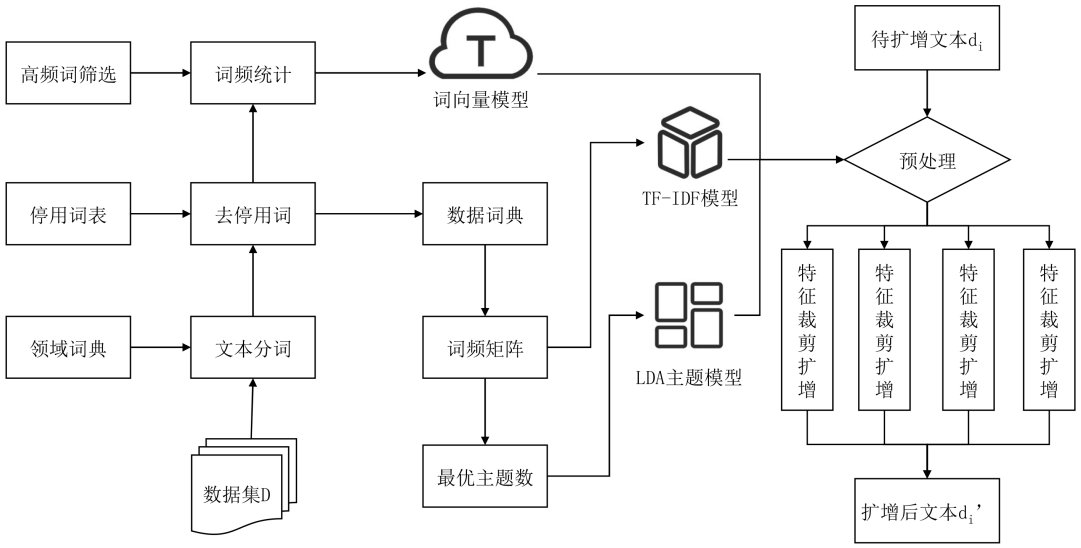


图 4-4: 基于领域特征的文本数据扩增方法一般步骤

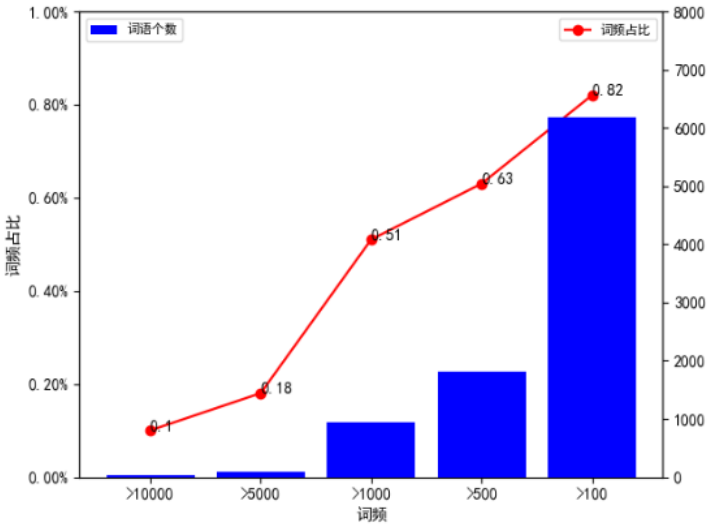


图 4-5: 媒体领域数据集词频统计图

在该数据集处理中，使用了使用哈工大 LTP 自然语言处理工具对文本操作，使用了领域词典提高分词的精准性，其中领域词典使用搜狗开源领域词库，停用词表使用哈工大的中文停用词表。根据词频统计表可以得到，语料集总词数 16 万，词频总数 50 万，如图4-5所示的词频统计图，在图中可以观察到，词频大于 100 的词语总词频占比 82%，词语个数 6000，因此，选取词频大于 100 的词语作为高频词。

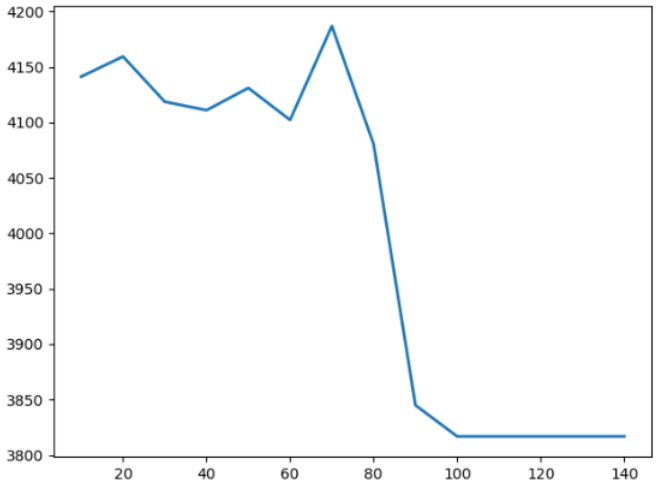


图 4-6: 不同 K 值下的主题模型困惑度曲线图

在构建主题模型时，根据困惑度确定主题个数。由图4-5可知，当主题数目大于等于 100 时困惑度最小，因此，可以确定主题数为 100。这是在数据扩

增前需要确定的两个核心指标。该数据集的文本结构和长度与司法领域数据集相似，在特征融合扩增选择相同依存关系树枝时，同样选择固定层数第三层。

完成扩增前语料集处理和模型的构建，下一步就是根据实验需要使用特征裁剪、特征融合、特征变换和特征替换方法实现扩增。

4.3 训练数据实验评估

为验证基于领域特征的文本数据扩增技术的可用性，设计对比实验通过使用基于领域的特征扩增方法和目前常用的 EDA 数据扩增方法扩增训练数据，检验领域特征扩增技术是否可以使文本分类的深度学习模型的性能得到提升以及提升的效果。实验设计如下。

本实验首先在司法领域数据集数据集中的盗窃、危险驾驶、故意伤害等五个类别中的每个类别随机选取了 5000 条共 25000 条司法裁判文书数据，作为实验的基础领域数据集，并用其来构建高频词表、训练扩增所需的 TF-IDF 模型、主题模型和词向量模型等以使扩增数据符合司法数据集的领域特征。

表 4-2: 司法领域数据集基于领域特征扩增结果表

	基础数据	抽取数据	特征裁剪	特征融合	特征变换	特征替换
盗窃	5000	1000	1000	1000	1000	1000
危险驾驶	5000	1000	1000	1000	1000	1000
故意伤害	5000	1000	1000	1000	1000	1000
交通肇事	5000	1000	1000	1000	1000	1000
毒品犯罪	5000	1000	1000	1000	1000	1000
总数	25000	5000	5000	5000	5000	5000

在文本分类模型训练中，模型训练数据的输入数据为司法裁判文书，标签为案情所涉及的罪名。训练集数据的生成过程为：从已经选取的 25000 条文本数据中再次随机选择每个类别中的 1000 条共 5000 条司法裁判文书数据作为原始训练数据。将这 5000 条司法裁判文书数据通过基于领域特征的文本数据扩增技术的四种方法进行扩增，生成经过扩增的训练数据集共 20000 条，包括基于

TF-IDF 权重的特征裁剪方法生成的 5000 条数据、基于主题模型的特征融合扩增方法生成的 5000 条数据、基于依存句法的特征变换方法生成的 5000 条数据和基于词频词性的特征替换方法生成的 5000 条数据。司法领域数据集基于领域特征扩增结果表如4-2所示。

表 4-3: 司法领域数据集 EDA 扩增结果表

	原始数据	随机删除	随机插入	随机交换	同义词替换
盗窃	1000	1000	1000	1000	1000
危险驾驶	1000	1000	1000	1000	1000
故意伤害	1000	1000	1000	1000	1000
交通肇事	1000	1000	1000	1000	1000
毒品犯罪	1000	1000	1000	1000	1000
总数	5000	5000	5000	5000	5000

作为对比，司法领域数据集同时使用 EDA 技术进行扩增。将已经选取的 5000 条司法裁判文书原始数据集使用 EDA 技术的四种方法进行扩增，生成经过扩增的训练数据集共 20000 条，包括使用随机删除（RD）扩增方法生成的 5000 条数据、使用随机插入（RI）扩增方法生成的 5000 条数据、使用随机交换（RS）扩增方法生成的 5000 条数据。司法领域数据集使用 EDA 技术进行扩增结果表如表4-3所示。

其次，对于媒体领域数据集，将已经选择的媒体领域数据集的 32500 条文本中作为实验的基础领域数据，构建相应的模型。在该数据集中，体育、财经、教育等五个类别中每个类别也随机抽取 1000 条共 5000 条数据，作为实验的原始训练数据。同司法领域的数据集一样，将 5000 条训练数据通过基于领域特征的文本数据扩增技术的四种方法进行扩增，生成经过扩增的训练数据共 20000 条，包括基于 TF-IDF 权重的特征裁剪方法生成的 5000 条数据、基于主题模型的特征融合扩增方法生成的 5000 条数据、基于依存句法的特征变换方法生成的 5000 条数据和基于词频词性的特征替换方法生成的 5000 条数据。媒体领域数据集基于领域特征扩增结果如表4-4所示。

表 4-4: 媒体领域数据集基于领域特征扩增结果表

	基础数据	抽取数据	特征裁剪	特征融合	特征变换	特征替换
体育	6500	1000	1000	1000	1000	1000
财经	6500	1000	1000	1000	1000	1000
教育	6500	1000	1000	1000	1000	1000
科技	6500	1000	1000	1000	1000	1000
时政	6500	1000	1000	1000	1000	1000
总数	32500	5000	5000	5000	5000	5000

同时，媒体领域数据集使用 EDA 技术进行扩增。将已经选取的 5000 条原始数据集使用 EDA 技术的四种方法进行扩增，生成经过扩增的训练数据集共 20000 条，包括使用随机删除（RD）扩增方法生成的 5000 条数据、使用随机插入（RI）扩增方法生成的 5000 条数据、使用随机交换（RS）扩增方法生成的 5000 条数据。媒体领域数据集使用 EDA 技术进行扩增结果表如表 4-5所示。

表 4-5: 媒体领域数据集 EDA 扩增结果表

	原始数据	随机删除	随机插入	随机交换	同义词替换
体育	1000	1000	1000	1000	1000
财经	1000	1000	1000	1000	1000
教育	1000	1000	1000	1000	1000
科技	1000	1000	1000	1000	1000
时政	1000	1000	1000	1000	1000
总数	5000	5000	5000	5000	5000

实验涉及两个领域，司法领域和媒体领域，每个领域设置 6 个训练集，利用每个训练集分别训练文本分类模型，司法领域模型为罪名预测模型，媒体领域模型为新闻类别预测模型。两个领域数据集使用基于领域特征的扩增方法和 EDA 扩增方法分别进行扩增，并将两者扩增结果进行比较，将 EDA 扩增数据作为对照组。

在基于领域特征扩增方法中特征裁剪方法与 EDA 中随机删除方法都是对文本内容进行相应删除实现扩增，因此将这两种方法进行对比，同样的，特征融合扩增方法与随机插入扩增方法都是在文本中插入新的内容实现扩增，因此将这两个方法进行比较；特征变换扩增方法与随机交换扩增方法都是对文本中的内容进行交换实现扩增，因此对这两个方法进行比较；特征替换扩增方法与同义词替换扩增方法都是对文本中的词语进行同义词替换操作实现扩增，因此将这两个方法进行比较。

表 4-6: 模型训练集数据表

训练集	领域特征扩增数据	EDA 扩增数据
训练集 1	5000 条原始数据	5000 条原始数据
训练集 2	训练集 1+5000 条特征裁剪扩增数据	训练集 1+5000 条随机删除扩增数据
训练集 3	训练集 1+5000 条特征融合扩增数据	训练集 1+5000 条随机插入扩增数据
训练集 4	训练集 1+5000 条特征交换扩增数据	训练集 1+5000 条随机交换扩增数据
训练集 5	训练集 1+5000 条特征替换扩增数据	训练集 1+5000 条同义词替换扩增数据
训练集 6	训练集 1+20000 条领域特征扩增数据	训练集 1+20000 条 EDA 扩增数据

由此，对于司法领域和媒体领域训练集的构建，具体为：训练集 1 为 5000 条原始数据，训练集 2 在训练集 1 的基础上加入 5000 条基于 TF-IDF 权重的特征裁剪扩增方法得到的数据，共 1 万条，对照组为在训练集 1 的基础上加入 5000 条使用随机删除扩增方法得到的数据，同样为 1 万条；训练集 3 是在训练集 1 的基础上加入 5000 条基于主题模型的特征融合扩增方法得到的数据，共 1 万条，对照组为在训练集 1 的基础上加入 5000 条使用随机插入扩增方法得到的数据，共 1 万条；训练集 4 是在训练集 1 的基础上加入 5000 条基于依存句法的特征变换扩增方法得到的数据，共 1 万条，对照组为在训练集 1 的基础上加入 5000 条使用随机交换扩增方法得到的数据，共 1 万条；训练集 5 是在训练集 1 的基础上加入 5000 条基于词频词性的特征替换扩增方法得到的数据，共 1 万条，对照组为在训练集 1 的基础上加入 5000 条使用同义词替换扩增方法得到的数据，共 1 万条；训练集 6 是在训练集 1 的基础上加入使用基于领域特征的四种扩增方法得到的 2 万条数据，共 25000 条，同样，对照组为在训练集 1 的基础上加入使用 EDA 的四种扩增方法得到的 2 万条数据，共 25000 条数据。模

型训练集数据表如表4-6所示。在原始数据中非用作训练集的数据中随机选取1000 条作为验证集，随机选取 2000 条作为测试集，损失函数为交叉熵损失函数，用在测试集上的准确率评估模型性能。

模型结构选取两种常用的文本分类模型卷积神经网络（CNN）和循环神经网络（RNN）。将司法领域训练集和媒体领域训练集的数据分别输入这两种文本分类模型里进行训练，观察模型在测试集上的表现。对于在训练过程中司法领域各组训练集生成的模型在测试集上的准确率表现情况如图4-7和4-8所示。

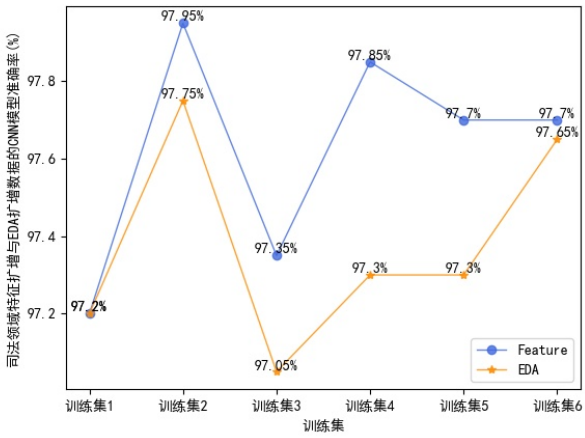


图 4-7: 司法领域特征扩增与 EDA 扩增数据的 CNN 模型准确率结果图

据图4-7和4-8，总体而言，相比于原始数据集，领域特征扩增和 EDA 扩增数据集的加入使模型相比在原始训练集上构建的模型在测试集上的准确率增加，模型的泛化能力提高，同时，使用领域特征扩增方法得到的数据比 EDA 方法扩增的数据对于使模型准确率提高幅度更大，可见基于领域特征的文本数据扩增的四种方法是可行有效的，扩增得到的数据质量相对较高。

具体来看，CNN 文本分类模型比 RNN 模型在相同数据集上的表现更好，其在测试集上的准确率领先 RNN。

司法领域的司法裁判文书数据集的质量较高，每个类别的文本数据具有鲜明的类别特征。由图4-7可知, 包含原始数据的训练集 1 训练得到的 CNN 模型准确率达到了 97.2%，在这样较高的准确率的前提下，训练集 2、3、4 和 5 加入相同规模的扩增数据，训练得到的模型准确率依然得到了提高，特别地，加入特征扩增数据训练的模型都比加入 EDA 扩增数据的模型准确率高。其中，训练集 2 的模型准确率值最高，使用领域特征扩增数据训练的模型准确率达到了 97.95%，EDA 扩增数据训练的模型准确率达到了 97.75%。对于在训练集 6

来说，加入了四组共 20000 条扩增数据，训练得到的模型准确率并不是训练集中的最大值。总的来说，五组在加入了特征扩增数据的训练集上得到的模型相比于在训练集 1 上生成的模型准确率平均改善率有 0.51%，高于加入 EDA 扩增数据训练集的 0.21%。

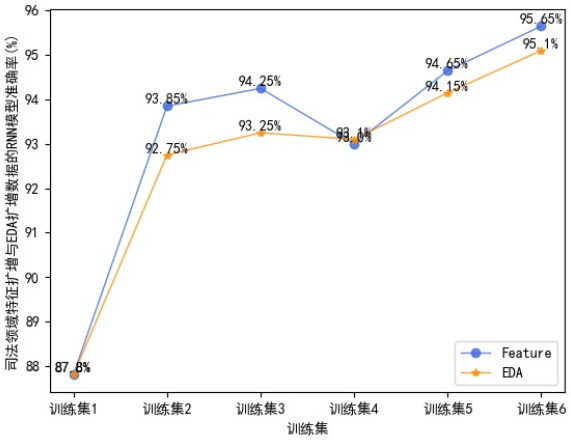


图 4-8: 司法领域特征扩增与 EDA 扩增数据的 RNN 模型准确率结果图

由图4-8可知，由包含原始数据集的训练集 1 训练得到的 RNN 模型在测试集上的准确率只有 87.8%，添加经过特征扩增和 EDA 扩增得到的数据后，模型的准确率大幅提高，均在 92% 以上。五个加入扩增数据的训练集中，除训练集 4 外，加入特征扩增数据训练的模型都比加入 EDA 扩增数据的模型准确率高，对于训练集 4，两个模型的准确率差别为 0.1%。相比于训练集 1，五组训练集生成的模型平均改善率分别为：6.48% 和 5.87%。总而言之，在司法领域数据集上，基于领域特征的扩增方法在提高模型准确率方面比 EDA 方法更有效。

同时，在媒体领域数据集中，加入扩增数据的训练集训练得到模型的准确率结果如图4-9和4-10所示，同司法领域相同，领域特征扩增数据的加入使模型相比于原始训练集上构建的模型测试集上的准确率提高，表示分类模型的泛化能力提高，总体而言，加入特征扩增数据的训练集比加入 EDA 扩增数据的训练集生成的模型的准确率高。

具体来看，如图4-9所示，包含原始数据的训练集 1 训练得到的 CNN 模型的准确率有 92.2%，在加入扩增数据的训练集中，训练集 2 生成的模型准确率最高，训练集 3 生成的模型准确率最低。值得注意的是，加入特征扩增数据的训练集生成的模型都比训练集 1 生成的模型准确率要高，而加入 EDA 扩增数据的训练集 3、4、5 生成的模型比原始数据训练集 1 生成的模型准确率更低。对

于加入扩增数据后模型准确率改善情况，五组在加入特征扩增数据的训练集上得到的模型相比于训练集 1 上生成的模型准确率平均改善率为 0.66%，而在加入 EDA 扩增数据上训练集生成的模型平均改善率只有 0.01%。显然，基于领域特征的扩增方法相比于 EDA 方法更能有效提高模型准确率。

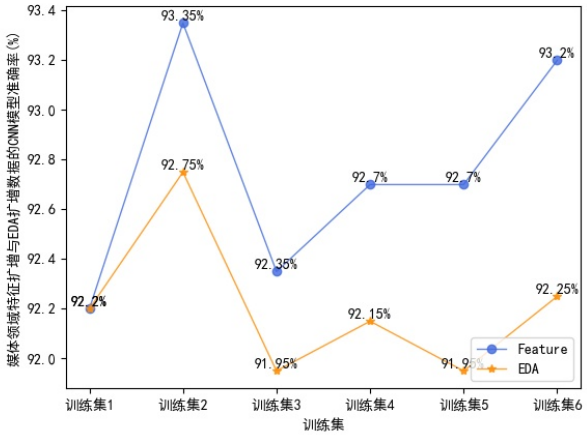


图 4-9: 媒体领域特征扩增与 EDA 扩增数据的 CNN 模型准确率结果图

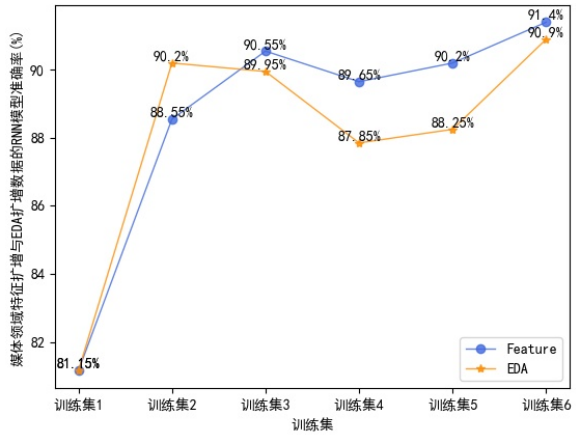


图 4-10: 媒体领域特征扩增与 EDA 扩增数据的 RNN 模型准确率结果图

媒体领域数据集使用两种扩增方法得到的扩增数据集在 RNN 模型上的准确率结果图如4-10所示。由实验结果图可知，原始训练集加入领域扩增数据和 EDA 扩增数据后，由此训练生成的模型在测试集上的准确率相比于训练集 1 生成的模型大幅提高，准确率提高幅度均在 6% 以上。其中，在训练集 2 上，加入特征扩增数据构成的训练集生成的模型在测试集上的准确率低于加入 EDA

扩增数据的生成的模型。总体来看，模型准确率平均改善率为 8.92%，在加入 EDA 扩增数据的训练集上，模型准确率平均改善率为 8.28%，基于领域特征的扩增方法在提高模型准确率方面稍高于 EDA 方法。

综上，通过司法领域数据集和媒体领域数据集及两种扩增方法得到的数据构建的训练集生成的文本分类模型的表现可知，基于领域特征的文本数据扩增技术的四种方法（FC、FF、FT 和 FR）对于提高模型的泛化能力，增强其在测试集上的表现是有效的，且比目前通用的 EDA 方法在提高模型泛化能力方面的效果要好。在实践中，可以根据实际情况，选择一种方法进行扩增或使用所有方法扩增文本数据，达到提高模型性能的目标。

领域特征扩增是在全部领域基础数据集上构建的 TF-IDF 模型、主题模型、词向量等为基础进行文本扩增的，然后选取部分数据集使用四种领域特征扩增方法进行扩增。据实验结果，加入扩增数据后训练集生成的模型相比于原始数据集训练的模型，在测试集上的准确率有了提高。如果将已有的全部数据用于模型训练，与部分数据进行扩增得到的数据相比，模型表现如何？为此，最后将司法和媒体原始数据集中除测试集和验证集外的所有数据作为训练集 7 分别使用 CNN 和 RNN 模型进行训练。其中，司法领域训练集有 22000 条文本，媒体领域训练集有 29500 条文本。领域原始数据集模型准确率结果如表 4-7所示。

表 4-7: 领域原始基础数据集模型预测准确率结果表

		训练集 7：领域原始数据集（除测试集和验证集部分）
CNN	司法	98.15%
	媒体	93.00%
RNN	司法	96.50%
	媒体	90.2%

由表 4-7可得，约五倍于训练集 1 数据量的训练集 7 生成的模型普遍比训练集 1 生成的模型在测试集上的准确率要高，尤其在 RNN 模型上，准确率提高了接近 9 个百分点。虽然训练集 7 数据量和数据质量较高，但生成的模型并不是所有训练集生成的模型中准确率最高的，例如，在媒体领域使用 CNN 构建的模型中，训练集 2 和 6 生成的模型比训练集 7 生成的模型准确率更高，在

媒体领域使用 RNN 构建的模型中，训练集 3 和 6 生成的模型比训练集 7 的准确率更高，训练集 5 与训练集 7 生成的模型准确率相同。在司法领域中，虽然训练集 1 至 6 构建的所有模型在测试集上的表现均没有超过训练集 7，但是在 CNN 模型上训练集 7 比训练集 1 的改善率与训练 2 至 6 与训练集 1 的平均改善率差别是 0.44%，在 RNN 模型上是 2.22%，基于原始数据集五分之一的数据量进行扩增得到了几乎等同于全部数据的准确率表现，可以说明基于领域特征的扩增方法在挖掘数据集领域特征、保证扩增数据质量方面效果显著。

最后，基于领域特征的四种扩增方法在扩增数据时的速度有所差异，特征融合扩增方法使用主题模型筛选相似文本进行扩增，该方法的扩增速度较慢，特征裁剪和特征替换扩增分别使用了 TF-IDF 模型和词向量模型，扩增速度次之，特征变换基于依存句法进行变换，扩增速度相对来说较快。实践中，为充分挖掘领域特征，将全部领域数据集利用起来构建相关模型，扩增时根据计算资源和时间资源要求综合选择合适的领域特征扩增方法。

4.4 本章小结

本章设计实验验证了基于领域特征扩增方法的有效性。首先设计扩增文本标签评估实验，验证扩增后的文本是否还保留其原有的领域特征，表现是在一定程度上是否保留其原有的标签。利用司法裁判文书数据集训练得到一个在测试集上准确率表现较好的 CNN 文本分类模型，使用领域特征扩增方法扩增的数据集 testFeature 作为测试集让模型来测试，并使用 EDA 技术生成的扩增数据集 testEDA 作为测试对照组，实验表明模型在测试集 testFeature 上表现更好，且准确率处于较高水平。可以说明，扩增后的文本仍然保留其原有的标签。

其次，在司法数据集和媒体数据集上使用领域扩增方法进行扩增，得到的数据集作为训练集在 CNN 和 RNN 模型上进行训练。

最后，将特征扩增方法与目前通用的 EDA 方法进行对比分析，把在司法和媒体数据集上使用这两种扩增方法扩增得到的数据作为训练集，通过文本分类实验数据对比分析得出，原始数据加扩增数据组成的训练集生成的模型比原始数据训练集上生成的模型在测试集上的准确率更高，且加入特征扩增数据的训练集生成的模型比加入 EDA 扩增数据的训练集生成的模型泛化能力更强。因此，基于领域特征的文本数据扩增技术是可行有效的。

第五章　总结与展望

5.1　总结

本文基于深度学习模型在开发中对于文本数据质量和数量的要求，在现有文本数据扩增技术和自然语言处理技术的基础上，充分挖掘数据集中的领域特征，研究并实现基于领域特征的文本数据扩增技术。为文本分类模型等对于文本数据有大量需求的开发者提供一种生成大规模、高质量训练数据的方法。

本文提出数据集预处理的步骤和四种基于领域特征的扩增方法。数据集预处理的步骤包括文本结构化处理、文本分词处理、去停用词和文本词频统计等，通过预处理节约计算资源，便于后续扩增。四种基于领域特征的扩增方法包括基于 TF-IDF 权重的特征裁剪方法，基于主题模型的特征融合方法，基于依存句法的特征变换方法和基于词频词性的特征替换方法。

其中，基于 TF-IDF 权重的领域特征裁剪方法是以文本分词在数据集中的 TF-IDF 值为依据，结合依存句法分析技术进行剪枝操作，实现文本数据扩增，以保持数据集基本特征和语义一致性；基于主题模型的特征融合方法是使用主题模型技术聚类数据集中的相似文本，将待扩增文本与相似目标文本进行内容交换，实现特征融合的文本扩增；基于依存句法的特征变换是使用依存句法分析技术解构文本，将句法树中依存关系相同文本进行交换，在不改变文本内容的情况下改变文本结构进行扩增；基于词频词性的特征替换方法是基于数据集分析构建领域集高频词表和词向量模型，提出领域特征词语的词性表，将符合高频词和相关词性的词语集使用词向量模型推荐近义词进行替换，实现文本数据扩增。扩增方法以司法领域中的司法裁判文书数据集为案例进行介绍。

本文通过设计实验对领域特征扩增文本进行质量评估，并与当前通用的 EDA 方法进行对比分析。首先，构建高质量文本分类模型，将特征扩增数据集作为测试集，与 EDA 技术对比，评估模型在两个测试集上的表现，得出领域特征扩增文本能够保持原有的类别标签和领域特征。其次，使用基于领域特征的文本数据扩增技术在司法和媒体两个领域的开源数据集进行扩增，观察扩增数据作为训练集对于文本分类模型性能提升的效果，并于 EDA 技术做对比，

评估扩增后文本的数据质量。经过实验验证，该扩增技术对于模型泛化能力的提高起到积极的作用，模型在测试集上的准确率相比于原始数据集生成的模型得到提高，且比 EDA 方法效果更好。

5.2 进一步工作

本文在已有数据扩增技术的基础上，挖掘有限数据集中的领域特征，提出四种领域特征扩增方法，并在司法和媒体领域数据集中设计扩增实验扩增训练数据，构建文本分类模型，验证扩增技术的有效性。在本文的基础上，该技术还有以下方面可以做进一步的改进。

文本的扩增速度还有待提升。该技术在特征裁剪、特征融合、特征替换方法中需要调用预训练的模型进行相应的处理，特征融合扩增时，需要筛选数据集中相似文本进行扩增，在此步操作中速度较慢。若提升扩增速度，在后续工作中可以扩充计算资源，在分布式系统中进行计算处理。

本文中每个文本数据扩增方式对应的最佳扩增参数还有待研究。四种扩增方法的参数在未来的实践应用时可以根扩增需要灵活设置。参数表明文本内容变化的幅度 [4]，在设置默认参数时遵循内容变换适中原则，让内容的变化保持在合理区间。后续将研究不同扩增方法的参数变化对扩增后文本质量乃至对模型性能的影响。

该技术作为一种通用的领域特征文本数据扩增方式，在后续可以针对不同领域的文本数据集在领域特征挖掘上更加精细。在分词时，更加细致地构建领域词典，增强分词的精准性。在构建词向量时，可以不局限于已有的数据集，获取更多的领域语料集构建词向量模型。在特征替换的词性表中，根据领域特征灵活增减不同的词性。在筛选相似文本时，研究设计更加高效、更加精准的相似文本筛选方式。

参考文献

[1] SHORTEN C, KHOSHGOFTAAR T M. A survey on Image Data Augmentation for Deep Learning[J/OL]. J. Big Data, 2019, 6 : 60. http://dx.doi.org/10.1186/s40537-019-0197-0.

[2] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6(1) : 1 – 48.

[3] LI Z, SPECIA L. Improving Neural Machine Translation Robustness via Data Augmentation: Beyond Back Translation[J]. CoRR, 2019, abs/1910.03009.

[4] WEI J W, ZOU K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[C/OL] // INUI K, JIANG J, NG V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. [S.l.] : Association for Computational Linguistics, 2019 : 6381 – 6387. http://dx.doi.org/10.18653/v1/D19-1670.

[5] KUSNER M J, HERNÁNDEZ-LOBATO J M. GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution[J]. CoRR, 2016, abs/1611.04051.

[6] KOBAYASHI S. Contextual augmentation: Data augmentation by words with paradigmatic relations[J]. arXiv preprint arXiv:1805.06201, 2018.

[7] XIE Q, DAI Z, HOVY E H, et al. Unsupervised Data Augmentation for Consistency Training[C] // LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020.

[8] TUITE C, AGAPITOS A, O'NEILL M, et al. Early stopping criteria to counteract overfitting in genetic programming[C/OL] // KRASNOGOR N, LANZI P L. 13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, Companion Material Proceedings, Dublin, Ireland, July 12-16, 2011. [S.l.]: ACM, 2011: 203 – 204.
http://dx.doi.org/10.1145/2001858.2001971.

[9] SAHA B N, KUNAPULI G, RAY N, et al. AR-Boost: Reducing Overfitting by a Robust Data-Driven Regularization Strategy[C/OL] // BLOCKEEL H, KERSTING K, NIJSSEN S, et al. Lecture Notes in Computer Science, Vol 8190: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III. [S.l.]: Springer, 2013: 1 – 16.
http://dx.doi.org/10.1007/978-3-642-40994-3_1.

[10] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. J. Mach. Learn. Res., 2014, 15(1): 1929 – 1958.

[11] MA J, LI L. Data Augmentation For Chinese Text Classification Using Back-Translation[J/OL]. Journal of Physics: Conference Series, 2020, 1651: 012039.
http://dx.doi.org/10.1088/1742-6596/1651/1/012039.

[12] ZHANG X, ZHAO J J, LECUN Y. Character-level Convolutional Networks for Text Classification[C] // CORTES C, LAWRENCE N D, LEE D D, et al. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015: 649 – 657.

[13] LIU T, CUI Y, YIN Q, et al. Generating and Exploiting Large-scale Pseudo Training Data for Zero Pronoun Resolution[C/OL] // BARZILAY R, KAN M. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. [S.l.]: Association for Computational Linguistics, 2017: 102 – 111.
http://dx.doi.org/10.18653/v1/P17-1010.

[14] HOU Y, LIU Y, CHE W, et al. Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding[C] // BENDER E M, DERCZYNSKI L, ISABELLE P. Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. [S.l.] : Association for Computational Linguistics, 2018 : 1234 – 1245.

[15] LUQUE F M. Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis[C] // CUMBRERAS M Á G, GONZALO J, CÁMARA E M, et al. CEUR Workshop Proceedings, Vol 2421 : Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019. [S.l.] : CEUR-WS.org, 2019 : 561 – 570.

[16] GUO Z, LIU J, HE T, et al. TauJud: test augmentation of machine learning in judicial documents[C/OL] // KHURSHID S, PASAREANU C S. ISSTA '20: 29th ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual Event, USA, July 18-22, 2020. [S.l.] : ACM, 2020 : 549 – 552. http://dx.doi.org/10.1145/3395363.3404364.

[17] 季培培, 鄢小燕, 岑咏华. 面向领域中文文本信息处理的术语识别与抽取研究综述 [J]. 图书情报工作, 2010(16) : 124 – 129.

[18] 刘桃, 刘秉权, 徐志明, et al. 领域术语自动抽取及其在文本分类中的应用 [J]. 电子学报, 2007, 35(002) : 328 – 332.

[19] 岑咏华, 季培培, 韩哲. 基于隐马尔科夫模型的中文术语识别研究 [J]. 现代图书情报技术, 2008, 24(12).

[20] 陈平, 匡尧, 胡景懿, et al. 增强领域特征的电力审计文本分类方法 [J]. 计算机应用, 2020 : 0 – 0.

[21] 田文颖. 面向专业领域的文本特征提取技术研究 [D]. [S.l.] : 国防科学技术大学, .

[22] 韩洁. 大规模 WWW 文档分类与特征词抽取方法研究 [D]. [S.l.] : [s.n.], 2002.

[23] KIM H, HOWLAND P, PARK H. Dimension Reduction in Text Classification with Support Vector Machines[J]. Journal of Machine Learning Research, 2005, 6(1): 37–53.

[24] 奉国和, 郑伟. 国内中文自动分词技术研究综述 [J]. 图书情报工作, 2011, 24(55).

[25] ZHANG M, DENG Z, CHE W, et al. Combining statistical model and dictionary for domain adaption of Chinese word segmentation[J]. Journal of Chinese Information Processing, 2012, 26(2): 8–12.

[26] VOINA A. Statistical estimation in hierarchical hidden markov model[J]. Cybernetics and Systems Analysis, 2014, 50(6): 898–912.

[27] ROGERS J G, CHRISTENSEN H I. A conditional random field model for place and object classification[C] // 2012 IEEE International Conference on Robotics and Automation. 2012: 1766–1772.

[28] ZHANG Y. Support vector machine classification algorithm and its application[C] // International Conference on Information Computing and Applications. 2012: 179–186.

[29] EKBAL A, SAHA S. Simulated annealing based classifier ensemble techniques: Application to part of speech tagging[J]. Information Fusion, 2013, 14(3): 288–300.

[30] 梁喜涛, 顾磊, OTHERS. 中文分词与词性标注研究 [J]. 计算机技术与发展, 2015(2015 年 02): 175–180.

[31] LI Z, ZHANG M, CHE W, et al. Joint models for Chinese POS tagging and dependency parsing[C] // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 1180–1191.

[32] 刘挺, 马金山. 汉语自动句法分析的理论与方法 [J]. 当代语言学, 2009, 11(2): 100–112.

[33] CHE W, LI Z, LIU T. Ltp: A chinese language technology platform[C] // Coling 2010: Demonstrations. 2010: 13–16.

[34] 周明, 黄昌宁. 面向语料库标注的汉语依存体系的探讨 [J]. 中文信息学报, 1994, 8(3) : 35 – 52.

[35] PAPADIMITRIOU C H, RAGHAVAN P, TAMAKI H, et al. Latent semantic indexing: A probabilistic analysis[J]. Journal of Computer and System Sciences, 2000, 61(2) : 217 – 235.

[36] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3 : 993 – 1022.

[37] RAMOS J, OTHERS. Using tf-idf to determine word relevance in document queries[C] // Proceedings of the first instructional conference on machine learning : Vol 242. 2003 : 29 – 48.

[38] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the National academy of Sciences, 2004, 101(suppl 1) : 5228 – 5235.

[39] GOLDSMITH-PINKHAM P, HIRTLE B, LUCCA D O. Parsing the Content of Bank Supervision[J]. Social Science Electronic Publishing, .

[40] BUJA A, SWAYNE D F, LITTMAN M L, et al. Data visualization with multidimensional scaling[J]. Journal of Computational and Graphical Statistics, 2008, 17(2) : 444 – 472.

[41] BAI S, BAI X, LATECKI L J, et al. Multidimensional scaling on multiple input distance matrices[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 31. 2017.

[42] 孙茂松, 李景阳, 郭志芃, et al. THUCTC：一个高效的中文文本分类工具包 [J], 2016.

致　　谢

在本篇论文完成之际，我要向所有关心我、帮助过我的人致以最诚挚的感谢。

首先感谢我的研究生导师刘嘉老师和陈振宇老师。刘嘉老师的耐心指导让我能够顺利完成论文。陈振宇老师让我认识到作为一名学者严谨与勤奋的态度，认识到对生活需要保有的热情与乐观。陈老师在整个研究生阶段给予了我很大的帮助，从初入实验室时帮助我确定研究方向，到科研过程中遇到问题时对我的建议与指引，再到论文写作过程中的悉心指导，我的毕业成果离不开陈老师的辛勤指导。

同时，我要感谢南京大学软件学院智能软件工程实验室 (Intelligent Software Engineering, iSE) 的所有老师与同学。感谢实验室老师们对我的指导与帮助，感谢学长学姐给予我学习与工作的经验，感谢实验室同学与我一起度过两年时光。

谢谢我的家人，在生活中给予了我莫大的帮助，在精神上给予我可以依靠的港湾，让我可以专心学业。

最后感谢为本文答辩评审的各位老师们，谢谢诸位的辛勤工作！

《学位论文出版授权书》

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：李翠阳

2021 年 5 月 20 日

论文题名	基于领域特征的文本数据扩增技术				
研究生学号	MF1932107	所在院系	软件学院	学位年度	2021
论文级别	☐ 硕士 ☑ 硕士专业学位 ☐ 博士 ☐ 博士专业学位 （请在方框内画勾）				
作者 Email	592918908@qq.com				
导师姓名	刘嘉 副教授				

论文涉密情况：

☑ 不保密

☐ 保密，保密期 (_____年_____月_____日 至 _____年_____月_____日)

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。