第八届“泰迪杯”数据挖掘挑战赛

基于卷积神经网络及集成学习的网络问政平台留言文本挖掘与分析

摘 要

互联网的快速发展为政府服务带来了极大的便利，网络问政平台积累了大量反映社情民意的文本数据，对这些数据应用自然语言处理技术与文本挖掘能大大提升政府的管理水平与施政效率。

针对任务一，由于原始数据集含有大量噪声，本文首先对原始数据进行预处理，包括去除特殊字符、文本去重、设计首尾冗余识别算法去除文本冗余信息，以及对文本进行分词并去除停用词。然后，本文通过数据增强方式，用卷积神经网络模型对文本按照一级标签分类，结果显示，本文建立的卷积神经网络模型在测试集上表现较好，得到的准确率为 91.4%，F1 值为 90.4%。最后，为了验证卷积神经网络模型的优越性，本文对比了多种模型的实验结果，结果证明，本文建立的卷积神经网络模型效果最佳。

针对任务二，对数据进行预处理后，首先用 TextRank 算法提取每条留言的关键词，依据关键词的词频去除噪声留言。对去噪后的数据集使用 Doc2Vec 训练句向量，依据语义相似性进行 K-means 聚类，分为 7 个大类，每类留言具有相同话题。随后，在每类中使用命名实体识别技术、模糊匹配算法以及高频词提取特定地点和特定人群对应的词汇，通过精确匹配得到 16 个留言数大于 10 的热点问题。最后，使用本文提出的异常时间点留言识别算法剔除每个热点问题中时间异常的留言，并定义了问题相关留言数量比、问题相关用户数量比、问题相关留言关注度与问题相关留言集中度四个热度评价指标，利用乘法合成法来组合熵值法与变异系数法所得的两种权重，再使用 TOPSIS 法计算热点问题的热度值，取排名前 5 的热点问题进行事件提取。

针对任务三，对数据进行预处理后，本文首先从相关性、完整性、可解释性和及时性四个方面建立了问答对统计相似度、问答对语义相似度、问答对主题相似度、是否引用法律条文、是否包含联系方式、答复意见句子长度、答复意见分词后词语个数以及问答对时间差 8 个指标。随后，以这 8 个指标为特征，使用 K-means 算法将答复意见聚为高质量、中等质量以及低质量三类。考虑到数据类别的不均衡性，本文建立基于集成学习的二阶段分类器对答复意见进行分类，即在第一阶段对数据进行“高”和“非高”分类，在第二阶段对“非高”类别的数据进行“中”和“非中”分类，使用两个基于集成学习的模型来共同完成多分类任务。最后，该二阶段分类器在测试集上的准确率达 98%，F1 值达 98.09%，Kappa 值达 96.93%。可见本文构建的二阶段分类器分类效果很好，能够对回复意见进行质量分类和有效评价。

关键词：卷积神经网络；TOPSIS；热点问题；K-means；集成学习

目 录

表 录

图 录

第一章 引言

1.1 挖掘背景

21 世纪的信息化发展日新月异，网络成了人们生活中必不可少的一部分。中国互联网络信息中心报告显示，截至 2019 年，我国网民规模达 8.29 亿，互联网普及率达 59.6%。互联网的强渗透力也为政府服务带来极大的便利，摆脱过去单调束缚的“人—信箱—政府”的民意收集模型，越来越多的网络问政平台的出现在人们的视野中。这种新的民意收集模式不仅简化人民群众表达心声的方式，还能快速及时的掌握社会热点问题，以提高服务效率。

然而，随着数据量的激增，对大数据量留言文本的快速处理成为一个亟待解决的问题。如今大数据技术发展迅速，应用于各行各业都取得不错的成就，将大数据技术引入政府服务，优化服务效率已然是大势所趋。自然语言处理技术是一门融合多学科的技术，近年来逐渐成为发展热潮，在实现人机交互方面取得巨大的进展。构建基于自然语言处理技术的智慧政务处理系统能避免人工处理留言的低效率，提高政府管理水平。基于此，本文运用时下流行的文本挖掘技术，对留言进行划分，并构建热点问题指标，推送人民群众关心的五大热点问题，并对相关答复设计一套评价指标，为政府工作提供参考。

1.2 挖掘意义

网络问政平台上的群众留言数量庞大且内容杂乱，依据人工经验处理留言分类工作会导致分类效率低、差错率高等问题。而建立关于留言内容的一级标签分类模型，可以帮助工作人员快速准确地处理留言分类工作，提高分类效率，降低差错率，减少工作量，大大提升工作效率，对相应职能部门快速获取其职责范围内的留言内容并及时处理具有重要意义。

热点问题是指某一时段内群众集中反映的某一问题。从一定意义上说，热点问题是公众利益和情绪的集中体现与表达，如果应对和处理不及时、不妥当，就很容易引起群众的不满和对立情绪，甚至演变成群体性事件。因此，定义合理的热度评价指标，找出热点问题及每个热点问题所对应的留言文本，可以帮助相关部门及时掌握问题动向，进行有针对性地处理，快速高效地解决问题，对提升民众满意度及相关部门服务效率具有重要意义。

相关部门对留言的答复意见，很大程度上反映了相关部门的服务效率、对留言问题的重视程度以及问题解决效果等。因此，建立一套完整的答复意见质量评

价方案是很有必要的，本文从相关性、完整性、可解释性与及时性等角度对相关部门答复意见的质量进行评价，对评价相关部门的工作效率、解决问题能力等具有重要参考意义。

1.3 问题描述

任务一分析：在处理网络问政平台的群众留言时，工作人员需要首先按照一定的划分体系将留言进行分类，以便后续将群众留言分门别类地划分给相应职能部门处理。但目前大部分留言分类工作还是依靠人工处理，由于网络问政平台留言数量大且内容杂，导致分类工作存在效率低、差错率高等问题。因此，对于任务一，需根据附件 2 所给留言数据，建立关于留言内容的一级标签分类模型，并使用 F1 值评价指标对所建立的分类模型进行评价。

任务二分析：热点问题是指某一时段内群众集中反映的某一问题。及时发现热点问题，有利于相关部门快速掌握问题动向，及时进行有针对性地处理，从而提升服务效率。因此，对于任务二，需根据附件 3 所给留言数据，将各留言进行归类，并定义合理的热度评价指标，给出各留言类别热度评价结果，按表 1 格式给出排名前 5 的热点问题，保存为文件“热点问题表.xls”；按表 2 格式给出相应热点问题类别下的留言信息，保存为文件“热点问题留言明细表.xls”。

任务三分析：相关部门对留言的答复意见，对于评价相关部门对群众所反映问题的解决效率、反馈情况等具有重要参考意义。因此，针对任务三，需根据附件 4 中相关部门对留言的答复意见，对答复意见从答复的相关性、完整性、可解释性等角度给出一套质量评价方案，并尝试实现对答复意见的评价。

第二章 群众留言分类

2.1 数据准备

2.1.1 数据描述

本文共有三个留言数据集，每一个任务对应一个数据集，另有一个数据集提供了留言的分类标签。针对任务一的数据集共有 9210 条留言，分为 6 个部分，包括“留言编号”、“留言用户”、“留言主题”、“留言时间”、“留言详情”与“一级分类”。“留言编号”是每条留言的编号，一条留言有一个编号；“留言用户”是每位网友的账号，一位网友一个账号；“留言主题”与“留言详情”记录用户所反映的现象；“一级分类”是对留言按标签分类，数据集内共有 7 种一级标签，其中城乡建设共 2009 条，环境保护共 938 条，交通运输共 613 条，商贸旅游共 1215 条，卫生计生共 877 条，教育文体共 1589 条，劳动和社会保障共 1969 条。图 2-1 显示各个分类标签的占比情况。

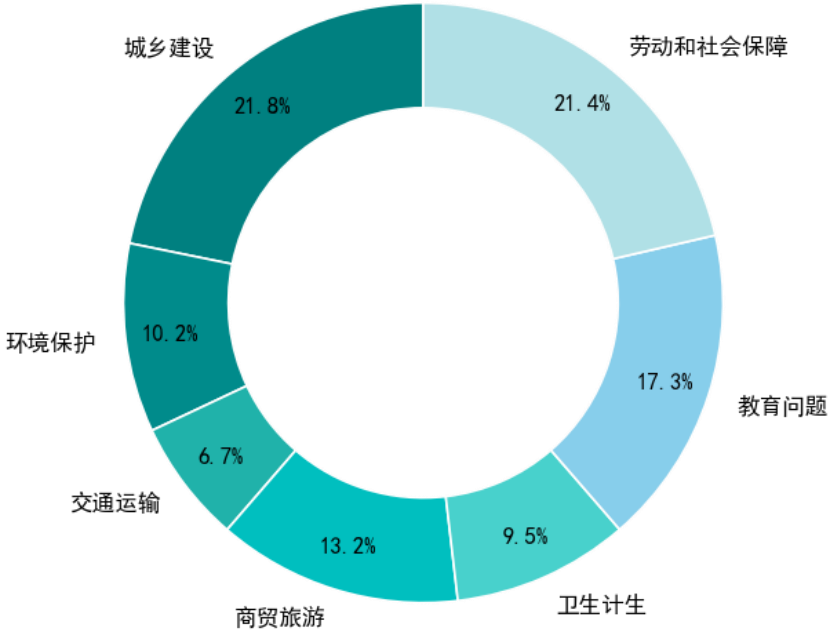

图 2-1 一级标签类别分布

2.1.2 数据预处理

对任务一的留言进行分类前，首先要对数据集进行数据清洗。留言文本语句多样、语言随意的特点导致其中存在大量不规范的内容，如噪声、文本重复、文

本冗余等情况，若不对留言文本内容进行处理，会影响分类器的分类效果。在数据集中，主要对“留言主题”与“留言详情”进行数据预处理。

1．去除特殊字符

⑴去除原因

第一类：空格。提取数据集文本时发现，提取出的信息中带有大量空格、“\xa0”、“\t”、“\n”等特殊字符，使得留言文本中存在很多空白内容，破坏了文本的规范性；

第二类：网址。原始数据集中存在大量网址，如：“https://baidu.com/”及“https://baidu.com/.05”这些网址信息夹杂在留言文本中，影响分词效果；

第三类：日期。留言文本中的日期数据并不包含分类的特征，属于无价值信息，因此在数据预处理阶段，本文去除数据集中结构为“XXXX 年 X 月 X 日”、“XXXX 年 X 月”的日期数据；

第四类：地名。原始数据集中带有“A 市”、“A8 县”，“B4 区”等英文单词为开头的地名，会影响分词效果，如：对“A1 区 A2 区华庭小区高层”分词时，若不删除地名，分词结果是“A1/区/A2/区华庭小区高层”，而删除地名后的分词结果是“华庭小区高层”，可见删除地名后分词结果更佳。且从语义角度来说，这些地名具有语义上的相似性，也会影响分类器的效果。

标点符号、常规英文字母与数字作为非文字符号的一部分将在分词结束后与停用词一同去除。因为本文发现先进行分词处理，后去除标点符号、常规英文字母与数字的分词效果优于先去除上述三项后分词的效果。

⑵处理方法

在 Python 中利用正则表达式去除原始数据集中的非文字符号。

2．去除重复留言

⑴去除原因

在数据预览时发现，部分“留言详情”内容存在高度一致性，因此需要去除重复文本，即筛选“留言详情”中完全重复与相似的留言，本文去除的对象有以下两类：

第一类：完全重复的留言。原始数据集中可能存在同一用户短时期内上传 2 条完全重复的“留言详情”以及后一用户复制前一用户“留言详情”的情况；

第二类：相似留言。同一用户两次上传的“留言详情”做了部分修改，大部分内容相同，以及后一用户复制前一用户留言并做部分修改，这些留言属于相似留言，反映的是同一现象，因此本文仅保留一条作为有效留言。

数据集中存在较多的完全重复与相似的留言文本，表 2-1 给出了完全重复与相似留言文本的示例。

表 2-1 重复与相似留言

类型	留言编号	留言主题	留言详情
完全重复	530	A3 区魏家坡小区脏乱差	尊敬的 A 市政府领导：您们好!我是 A 市 A3 区魏家坡…
	532	A 市魏家坡小区脏乱差	尊敬的 A 市政府领导：您们好!我是 A 市 A3 区魏家坡…
	303	A1 区蔡锷南路 A2 区华庭楼顶水箱长年不洗	A1 区 A2 区华庭小区高层为二次供水，楼顶水箱长年…
	319	A1 区 A2 区华庭自来水好大一股霉味	A1 区 A2 区华庭小区高层为二次供水，楼顶水箱长年…
相似重复	318635	A7 县供销社原 A 市泰阳商城丢失员工党员档案	尊敬的省供销社领导：您好！我是吴晓艳，1979 年 9 月 2…
	318735	A7 县供销社原 A 市泰阳商城丢失员工党员档案	您好！我是吴晓艳，1979 年 9 月 23 日生，身份证号…
	118115	对 K11 县沱江镇第七小学学位建设工程的质疑	K11 县沱江镇第七小学学位建设工程质疑事实依…
	118119	对 K11 县县为人小学综合楼工程项目招标办法的质疑	K11 县为人小学综合楼工程项目采用楚建建[2013]28…

（2）处理方法

对于完全重复的留言文本，处理的原理非常简单，利用 Python 程序判断语料库中是否存在完全重复的文本，若存在，则保留一条完全重复文本；

对于相似文本的处理，本文采用的是 Simhash 算法，该算法将文本映射成指纹，通过对比指纹来识别相似文本。其优点是检索速度快，适用于海量文本，因此本文选用 Simhash 算法解决相似文本问题。其原理是将在超大集合内查找数据的问题利用哈希函数转换映射的方法转化为在较小集合内查找数据的问题，大大减少计算量，主要包括：分词、计算 hash、加权、合并、降维。其中 hash 值的计算是通过 hash 函数对每一个词向量进行映射，产生一个 n 位二进制串，使字符串数字化。

相比于短文本，Simhash 算法更适用于长文本，由于数据中的留言详情经过去除特殊字符的处理后，其平均长度约为 200，文本内容较长，所以使用该算法取得较好的效果。求得文本之间的相似度后，保留一条为有效留言，其余作为无效留言去除。

经上述处理后，本文共去除 260 条完全重复留言，155 条相似留言，剩余 8795 条有效留言文本。

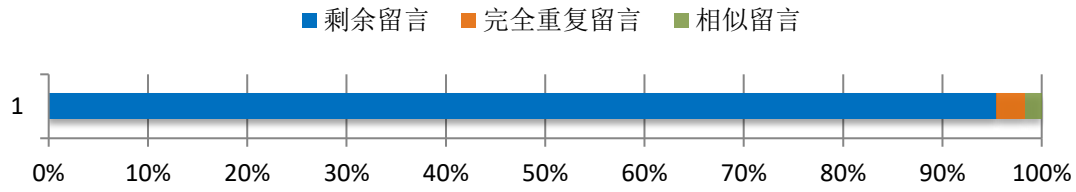

图 2-2 重复与相似留言占比

3．去除首尾冗余

（1）去除原因

本文对原始数据集进行预览时发现，一些“留言详情”的首部会出现“尊敬的胡书记：您好！”、“尊敬的领导：”等字样，在尾部会出现“谢谢！”、“希望能得到重视”等字样。这些内容是出于中文语言习惯，并不包含有价值的信息，可以视作文本的冗余部分。

（2）处理方法

为了解决首尾冗余对模型效果的影响，本文提出了一种首尾冗余识别算法，先对文本进行分句处理，再按照首尾句句子特点制定规则来过滤无关句子。

规则一：通过浏览发现，留言内容的首句以向领导问好的形式居多，也有直接言明遇到的实际问题，这时，就需要过滤向领导问好的无关句，观察发现，这类句子一般较短，即使略长也带有“好”字，所以以此为规则删除无关首句；

规则二：相比于首句，尾句的表达更加多样化，有表示感谢、请求领导解决问题、落款等形式，这些句子也无关问题本身，同样需要过滤，观察发现，这类句子也一般较短，但部分短句中也带有实际问题的表达，这种句子一般带有问号，所以以此为规则删除无关尾句。

该算法具体步骤为：

Step1：按标点符号对文本进行分句；

Step2：索引文本首句，若字段长度小于 7，则直接删除，否则进入下一步；

Step3：若字段长度在 7 至 15 之间，同时包含“好”字，则将其删除，其余保留；

Step4：索引文本尾句，若字段长度小于 4，则直接删除，否则进入下一步；

Step5：若字段长度在 4 至 10 之间，且不含问号（？），则直接删除，其余保留。

算法流程如下图：

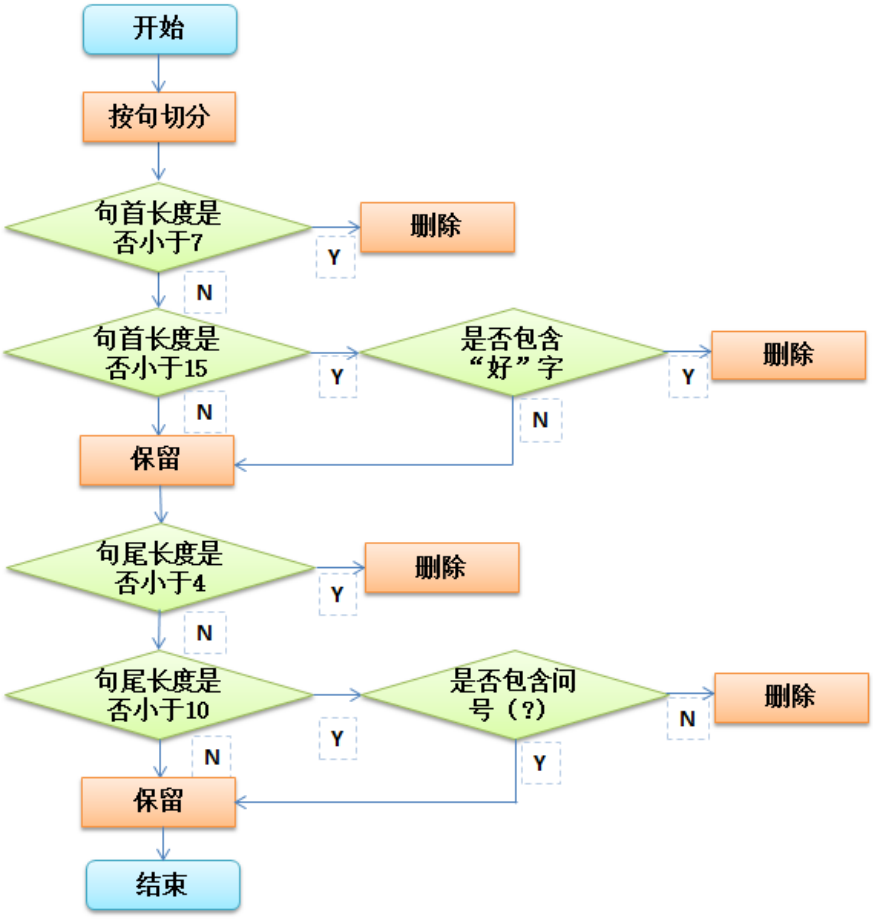

图 2-3 首尾冗余识别算法流程图

利用首尾冗余识别算法能较好的识别出以下内容（示例），且从后续的操作中可以证明，这种算法能提高分类器效果。

尊敬的省计划生育委员会领导：你们好！	希望一视同仁
张主任：您好！	马鞍镇马田村
各位老爷们好！	谢谢—！
尊敬的省领导，您们好！	有谁可以解答一下吗
尊敬的张健主任您好！	谢谢
尊敬的张健主任:您好！	谢谢。
邹主任，您好！	也请各位领导调查。
张厅长，您好！	此致！
张厅长，您好！	希望政府能给解决。
尊敬的省卫计主任：您好！	此致敬礼！
敬爱的领导你们好。	致此农民工谨上
尊敬各级领导好！	星期六
食药监局领导您好！	！！
尊敬的张主任：您好！	请领导们引起重视。

图 2-4 识别出的首部（左）与尾部（右）冗余

4．分词技术

中文以字为单位，多个字连在一起构成一个表达具体含义的词。在文本中，

句段之间通过标点符号进行简单分割，而词与词之间没有明显的分割符号，因此首先需要在语句中划分出具有独立意义的词，即通过分词工具将连续的字序列按照一定规则重新组合成词序列的过程。

分词结果对后续建模的效果具有至关重要的影响，目前使用较多的分词工具有 Jieba 分词工具、NLPIR 分词系统以及 HanLp 分词工具。

（1）Jieba 分词工具

Jieba 分词属于概率语言模型分词。在全切分所得的所有结果中求某个切分方案 S，使得 P(S)最大。Jieba 分词支持三种分词模式，包括全模式、精确模式、搜索引擎模式。

（2）NLPIR 分词系统

NLPIR 是中科院计算所研制的基于多层隐马尔科夫模型的中文词法分析系统。该系统提供了中文分词、词频统计、词性标注、命名实体识别、新词识别等功能。

（3）HanLp 分词工具

HanLp 是一系列模型与算法组成的工具包，具备功能完善、预料时新、性能高效的特点，提供了中文分词、词性标注、命名实体识别、句法分析、文本分类和情感分析等。

数据集中“留言主题”和“留言详情”均带有用户所要传达的信息，且“留言主题”中也含有“留言详情”中没有的关键词，所以为了丰富用户所要传达的语义信息，本文将数据集中的“留言主题”与“留言详情”合并作为分类的依据，随后使用 HanLp 分词工具对合并后的留言进行分词，所得分词效果较好。

5．过滤停用词

停用词是指那些包含信息少且在文本中大量出现的词语。比如“啊”、“和”、“的”、“他”等。这些词语和符号在所有文本中都会大量出现，并且在表示文本语义上具有较弱的作用。因此，过滤停用词能有效提高文本的检索效率和效果，使分类结果更加精确，同时过滤标点符号、常规英文字母与数字以及长度为 1 的词。本文使用的停用词表示例内容见下表 2-2。

表 2-2 停用词表

停用词	停用词	停用词
且	一旦	一方面
与	两者	为什么
个	中小	乃至于
乃	主要	且不说
么	之一	于是乎

经过上述预处理、分词及过滤停用词的操作后，本文得到较准确的分词结果，表 2-3 显示分词后结果示例，从中可以看出，本文的分词结果效果不错。

表 2-3 分词结果

留言详情分词结果
西湖建筑集团/占道/施工/安全隐患/大道/便道/未管所/路口/加油站/路段/人行道/包括/路灯/西湖建筑集团/燕子/安置房/项目/施工/围墙/上下班/期间/路上/人流/车流/极多/安全隐患/请求/文明城市/整改/极不/文明/路段
在水一方/大厦/烂尾/多年/安全隐患/位于/书院路/主干道/在水一方/大厦/一楼/四楼/人为/拆除/设施/烂尾/多年/护栏/围着/占用/人行/道路/护栏/锈迹斑斑/倒塌/危机/过往行人/车辆/请求/部门/牵头
蔡锷南路华庭楼顶/水箱/长年/华庭/小区/高层/供水/楼顶/水箱/长年/自来水/龙头/霉味/日常生活/必不可少/用品/致癌物/住在/健康/保障/政府/街道/领导/重视/环保部门/检测/健康/生活/环境
投诉/盛世/耀凯/小区/物业/无故/停水/购买/盛世/耀凯/小区/两层/共计/平方/足额/缴纳/物业费/费用/小区/入住/成立/小区/业委会/物业/公司/小区/为所欲为/物业/业主/服务/管理/业主/小区/水电费/供电/供水公司/想问/政府职能/部门/业主/投诉/物业/公司/部门

2.2 特征提取

将文本输入分类器之前，需要将文本转化为计算机所识别的符号数学形式，NLP 中最直观的表示方式就是 One-hot Representaion，它将每个词表示为一个向量，长度为词表大小，只有一个维度值为 1，但是这种表示方式没有考虑语义，存在“词汇鸿沟”的缺陷。因此一般采用 Distributed Representaion，它表示低维实数向量，常见维度为 50 维和 100 维。这种方法可以让相关或相似的词距离上更接近。常用的训练词向量方法有 Word2Vec 与 FastText。

1．原理

（1）Word2Vec 原理

Word2Vec 模型包含三层神经网络算法，训练词向量的同时达到降维的目的，它能从海量的文档数据中训练高质量的词向量，本文从语料库中训练词向量，得到的结果更加准确[1]。Word2Vec 的底层包括两种训练模型，分别是 CBOW 和 Skip-gram。CBOW 模型的训练输入是某特征词上下文相关的词对应的词向量，而输出就是这特定词的词向量。Skip-gram 的思路与 CBOW 相反，其输入特定词的词向量，而输出是特定词对应的上下文词向量，如下图 2-5 所示。本文使用

Skip-gram 模型。

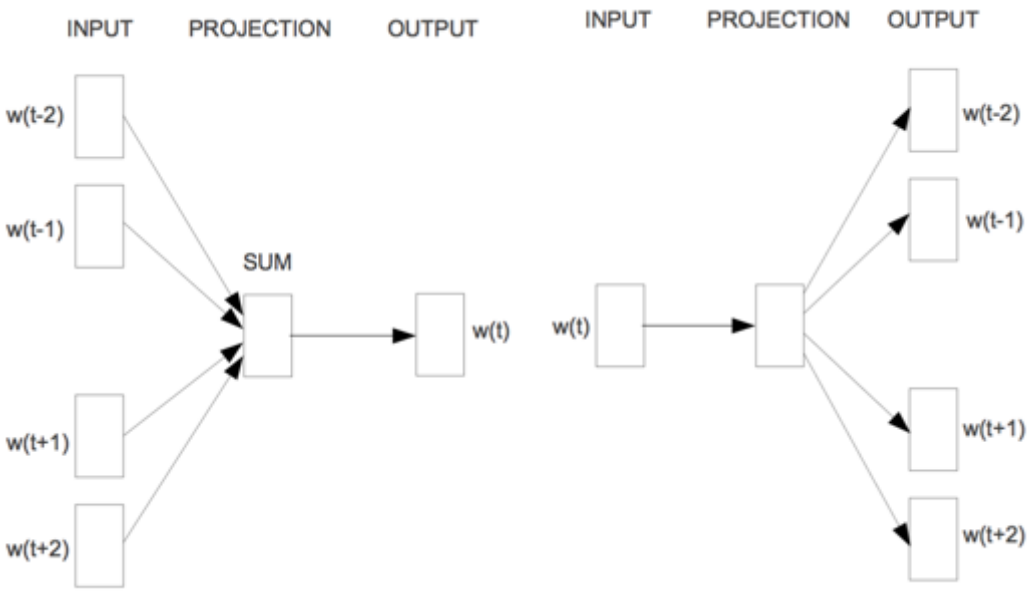

图 2-5 CBOW(左)与 Skip-gram(右)模型示意图

（2）FastText 原理

FastText 是 2016 年发布的一个开源文本分类器[2]。它的显著特点是速度非常快，其在保持分类效果的同时，大大缩短了训练时间。且 FastText 能够自己训练词向量，不需要预先训练词向量。整体流程为：输入一段文本至 FastText 模型，输出这个词序列属于不同类别的概率。序列中的词和词组组成特征向量，特征向量通过线性变换映射到中间层，中间层再映射到标签。

①模型架构

FastText 的模型架构基于 Hierarchical Softmax，有三层架构：输入层、 隐藏层、输出层。FastText 将整个文本作为特征去预测文本的类别。

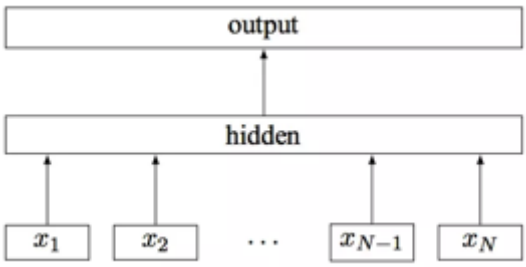

图 2-6 FastText 模型架构

②层次映射

将输入层中的词和词组构成特征向量，再将特征向量通过线性变换映射到隐藏层，隐藏层通过求解最大似然函数，然后根据每个类别的权重和模型参数构建 Huffman 树，将 Huffman 树作为输出。

③N-gram 特征

N-gram 是基于语言模型的算法，基本思想是将文本内容按照子节顺序进行大小为 N 的窗口滑动操作，最终形成窗口为 N 的字节片段序列。同时，FastText 过滤低频的 N-gram 以提高效率。

FastText 为计算单词表示也提供了两种模型：Skip-gram 和 CBOW，这和 Word2Vec 内容一样，本文使用 Skip-gram 模型。

2．训练词向量

本文采用融合 Word2Vec 模型与 FastText 模型训练词向量的结果，即利用这两种模型分别训练出词向量后，对同一个词的词向量按顺序拼接为一个新的词向量来表示该词，如：“语言”在 Word2Vec 中可能被表示成[0.1,0,1,0.1,0.2,0.12,…]，在 FastText 中可能被表示成[0.2,0.2,0.2,0.11,0.21,…]，那么“语言”这个词的最终表示结果为[0.1,0,1,0.1,0.2,0.12,…, 0.2,0.2,0.2,0.11,0.21,…]。利用这种方式得到的词向量能体现更丰富的语义。本文分别用 Word2Vec 与 FastText 训练词向量，维数都为 100 维，拼接后的词向量维度为 200 维。

2.3 建立模型

2.3.1 卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）是一种深度前馈人工神经网络，已在多领域持续快速发展[3、4]。下图 2-7 是一个传统的用于文本分类任务的卷积神经网络结构，也是本文用于说明的一个简化例子。由此可以看出卷积神经网络主要包括嵌入层（Input Layer）、卷积层（Convolution Layer）、池化层（Pooling Layer）、全连接层（Fully Connected Layer）、激活函数（Activation Function）与损失函数损失函数。

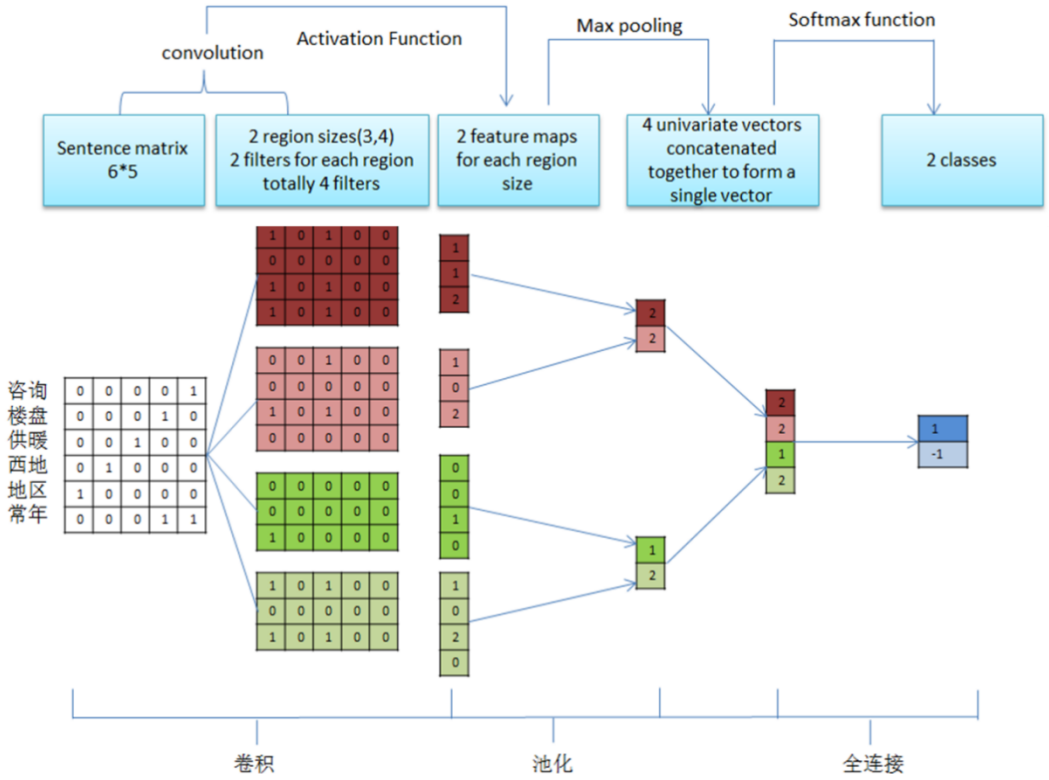

图 2-7 卷积神经网络结构图

1．输入层（Input Layer）

输入层是一个个词向量堆叠形成的 $n \times k$ 矩阵，其中 n 为一个句子中的词数，k 是每个词对应的词向量的维度，图 2-7 中的输入数据是一个 6×5 矩阵；

2．卷积层（Convolution Layer）

卷积层进行卷积运算，一般包括输入，核函数和输出特征映射。卷积是一种局部操作，通过一定大小的卷积核作用于输入数据的局部区域来获取数据的局部特征信息，图 2-7 中的第二列构成一个卷积核。卷积过程如下图 2-8 所示，Feature map 是卷积运算的输出，它是由对应的矩阵与卷积核做逐点乘的求和。通过卷积操作将输入的 6×5 矩阵映射成一个 3×1 的矩阵

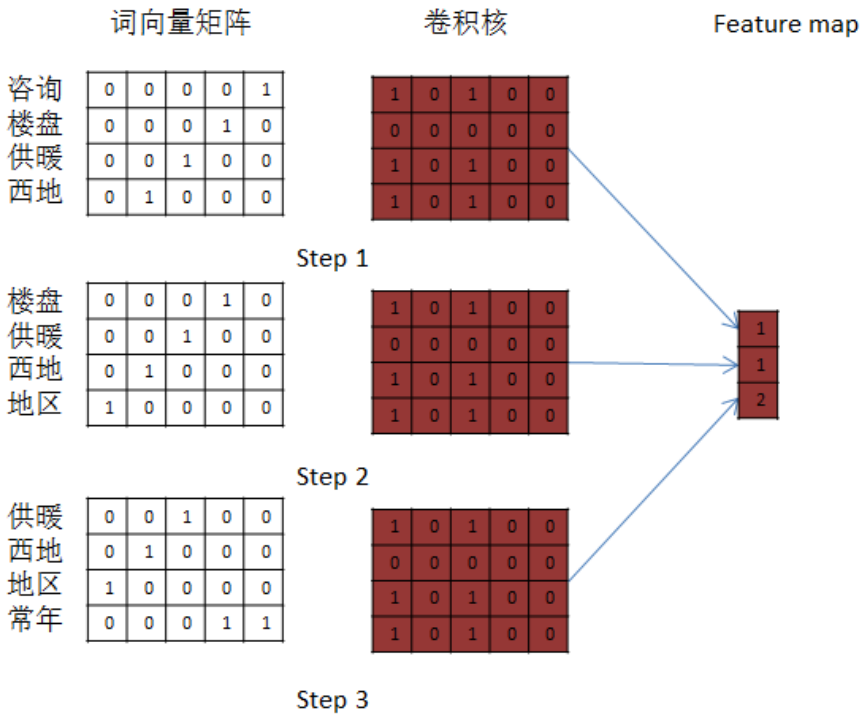

图 2-8 卷积过程

3．池化层（Pooling Layer）

池化是卷积神经网络重要的操作，它是一种非线性的下采样算法。它在保持主要特征的情况下，降低了维度，加快计算，在一定程度上可以防止过拟合，有助于模型的优化。常用到的池化函数有最大值池化函数与平均值池化函数，本文采用最大值池化函数。从图 2-9 可以看出上例中的 Feature map 经过池化从三维变成了一维。

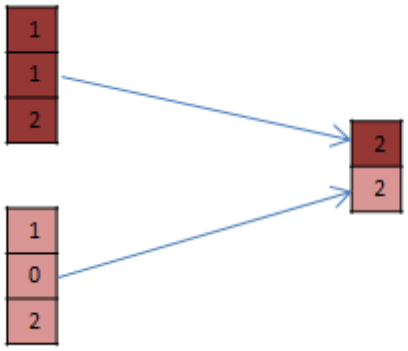

图 2-9 池化结果示意图

4．全连接层（Fully Connected Layer）

全连接层是卷积神经网络的最后一层，起到分类器的作用。卷积神经网络中的卷积层、池化层和激活函数层等操作将原始数据映射到抽象特征空间，全连接层是将特征表示映射到样本的标记空间。

5．激活函数（Activation Function）

激活函数用于对卷积的结果进行一个非线性函数的映射。把被激活的神经元的一些特征通过函数保留并且映射出来，将原本的线性函数转换成非线性函数，从而解决一些非线性问题。常用的激活函数有很多，本文使用的是 ReLu（Rectified Linear Unit）激活函数，其表达式为：

$$\sigma(x)=\max\{0,x\}=\begin{cases}x, x\geq 0\\ 0, x<0\end{cases}$$

6．损失函数

本文使用交叉熵损失函数，其公式表示为：

$$H(p,q)=-\sum_{i=1}^{n}p(x_i)\log(q(x_i))$$

2.3.2 模型设计

本文针对任务一的一级标签分类任务，应用基于卷积神经网络的分类模型。通过分析语料集样本的长度，指定一个输入序列的长度 seq_length，比 seq_length 短的样本序列需要填充，比 seq_length 长的序列需要截取，并利用预训练得到的词向量输入到词嵌入层，又由于宽度不同的卷积核可以捕捉到不同个数的相邻词之间的相关性，所以本文定义多个卷积核，并使用这些卷积核对输入分别做卷积计算，然后经过时序最大池化，再将向量连结起来，通过全连接层将连结后的向量变换为各类别的输出。该模型的详细介绍如下：

1. 词嵌入层：将文本编码成数字格式并填充到相同长度，再初始化 Word2Vec 和 FastText 预训练好的词向量。

2．卷积层：对输入的文本序列进行卷积核大小为 3、4、5 的卷积操作，提取文本特征。

3．池化层：采用 1-Max Pooling，经过最大值池化层对数据降维，并采用 ReLU 作为激活函数来防止方向传播中出现的梯度问题。

4．全连接层：将池化的向量进行拼接。

5．Dropout 层：用于缓和过拟合化。

6．Softmax 层：将连接后的向量变换为有关各类别的输出。

7．网络优化：采用梯度下降以优化卷积神经网络，优化函数为交叉熵损失函数，并在损失函数上添加 L2 正则化项。

由于类别不平衡，本文通过数据增强的方法提高模型的性能，即对数量较少的类别进行同义词替换、随机插入、随机交换、随即删除的操作。

具体参数见下表 2-4。

表 2-4 卷积神经网络参数表

解释	值
词向量维度	200
序列长度	200
卷积核数目	128
卷积核尺寸	[3,4,5]
全连接层神经元	32
Dropout 保留比率	0.8
学习速率	0.001
学习速率减缓因子	0.9
最大梯度范数	6.0
L2 正则化参数	0.01
每批训练个数	16

2.3.3 模型效果评价

1．评价指标

本文对模型效果进行评价时，采用准确率（Accuracy，简记为 A）和 F1 值（F-Score）作为性能评价指标。

混淆矩阵是对分类问题所得的预测结果的总结，它能显示模型在预测结果时产生的误差。混淆矩阵的关键在于计算预测结果正确和不正确的数量，并汇总在一个矩阵中，多分类问题的混淆矩阵如下表 2-4 所示，其中 a_{ij} 指预测结果中真实类别为 i 类，预测类别为 j 类的样本数。

表 2-5 混淆矩阵

混淆矩阵		预测值			
		1	2	…	n
真实值	1	a_{11}	a_{12}	…	a_{1n}
	2	a_{21}	a_{22}	…	a_{2n}
	…	…	…	…	…
	n	a_{n1}	a_{n2}	…	a_{nn}

（1）准确率

准确率是分类模型中最常见的性能评价指标，准确率越高，分类器效果越好。对于给定的测试集数据，准确率是被模型正确预测的样本数的累加与总样本数之比。

$$A=\frac{\sum_{i=1}^{n}a_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}}$$

(2)F1 值

F1 值是查全率（Recall，简记为 R）和查准率（Precision，简记 P）的调和平均值。在多分类问题中，每个类别都有一个 R_i、P_i 和 F_i 值，总体 F1 值一般取每个类别 F_i 值的平均值。F1 值综合考虑了 P 和 R 的结果，当它较高时则能说明试验方法比较有效。

$$P_i=\frac{a_{ii}}{\sum_{i=1}^{n}a_{ij}},j=i,\quad R_i=\frac{a_{ii}}{\sum_{j=1}^{n}a_{ij}},\quad F1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

2．模型效果

将预处理后的数据集划分为训练集、验证集和测试集。下面给出卷积神经网络模型在测试集上得到的分类效果：

表 2-6 卷积神经网络各类别 F1 值

类别	类别名	F_i 值
1	城乡建设	0.90
2	环境保护	0.95
3	交通运输	0.82
4	教育文体	0.95
5	劳动和社会保障	0.93
6	商贸旅游	0.88
7	卫生计生	0.90

表 2-7 卷积神经网络效果评价指标值

评价指标	值
F1 值	0.904
准确率	0.914

从上表可以看出，7 个类别的 F1 值均在 82%以上，整体准确率达到了 91.4%，F1 值达到 90.4%，，可见本文所采用的基于卷积神经网络的分类模型效果较好。该卷积神经网络对训练集与验证集的指标如图 2-10 所示。

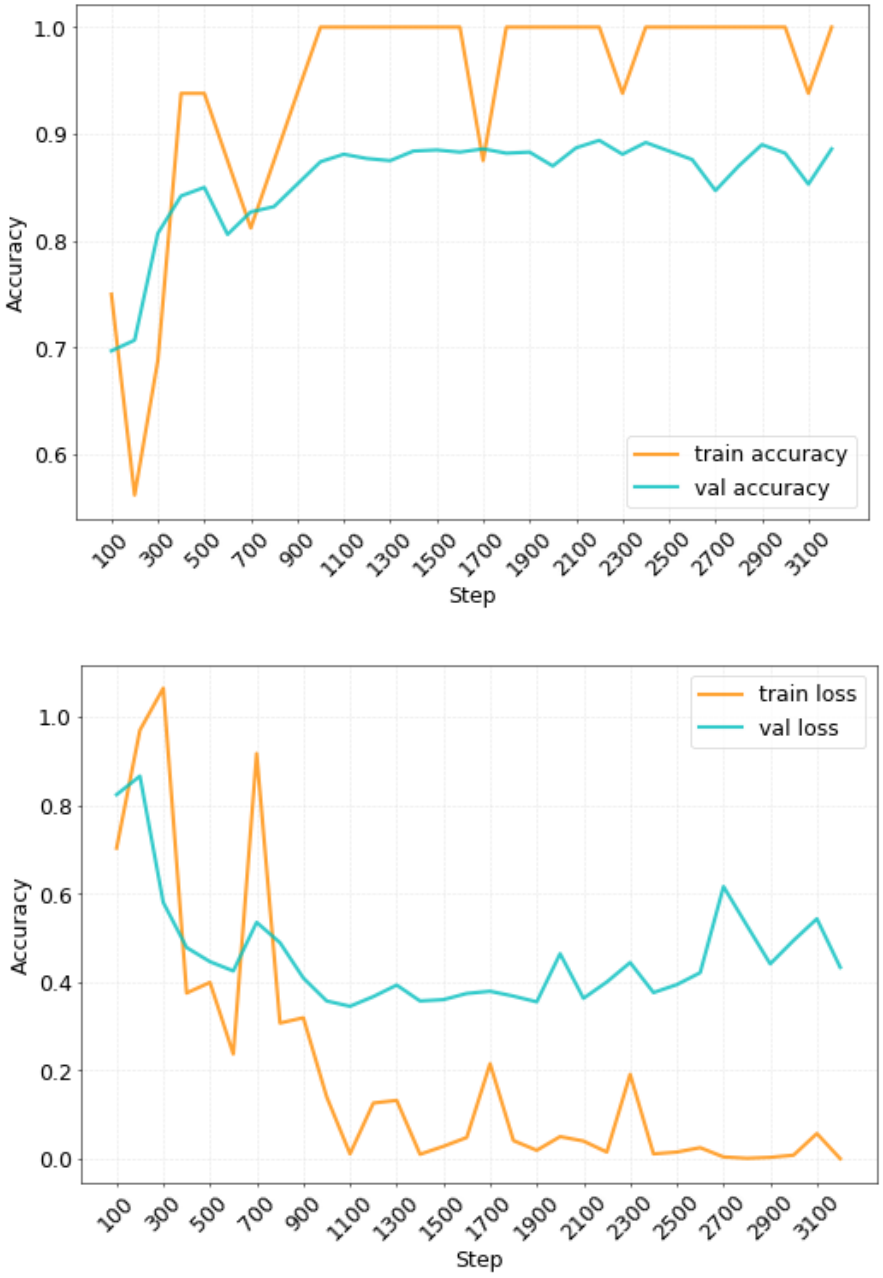

图 2-10 卷积神经网络指标效果图

为了验证本文构建的卷积神经网络分类模型具有优越性，下面将分别对模型预处理、特征提取与模型选择三个阶段采用不同处理方式所得的结果进行比较。

（1）是否剔除冗余的结果对比

在数据预处理阶段，本文提出了首尾冗余识别算法，对数据集“留言详情”部分去除首部与尾部的冗余信息，保留有价值的留言文本。下图 2-11 是未去除冗余与去除冗余的文本使用卷积分类的效果，未去冗余信息的准确率为 89%，

F1 值为 88.5%。而去除冗余信息的准确率为 91.4%，F1 值为 90.4%，分别提高了 2.4%和 1.9%。从中可以看出，本文提出的首尾冗余识别算法能提高分类效果。

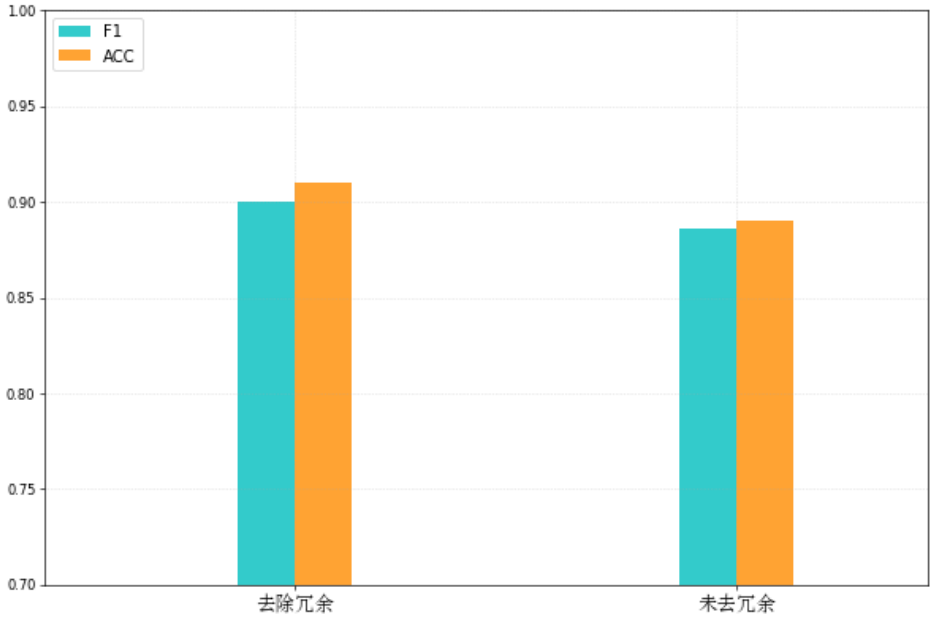

图 2-11 去除冗余信息效果对比

（2）不同词向量下的结果对比

在特征提取阶段，本文使用 Word2Vec 与 FastText 结合的方式训练词向量，将结果按顺序拼接作为最终使用，下图 2-12 是单独使用 Word2Vec、单独使用 FastText 与两者结合后放入卷积神经网络的效果。结果显示，单独使用 Word2Vec 的模型的准确率为 90.2%，F1 值为 90.1%，单独使用 FastText 的模型的准确率为 90.8%，F1 值为 90.3%，两者结合使用的模型的准确率为 91.4%，F1 值为 90.4%。从中可以看出，两者结合的模型分类效果更优。

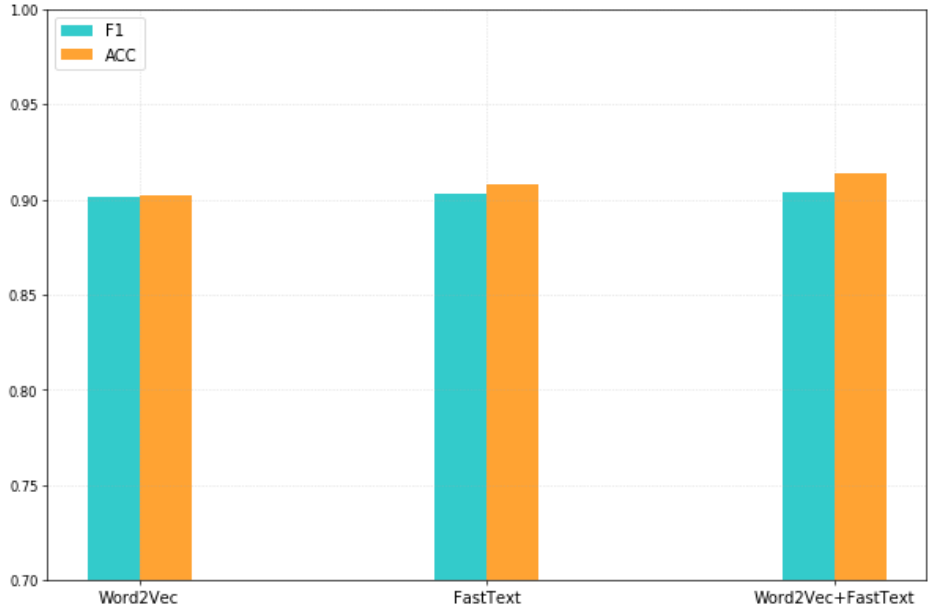

图 2-12 词嵌入模型效果对比

（3）不同模型的结果对比

在模型选择阶段，本文曾使用随机森林、XGBoost 与这两种模型的加权平均模型（1：1 融合）来进行训练，随机森林与 XGBoost 都是较为成熟的分类算法，在分类问题中表现出优良的特性。但是相比于本文所建立的卷积神经网络，随机森林、XGBoost 以及两者的加权平均模型的效果都不佳。下图 2-13 显示，随机森林模型中，准确率为 84%，F1 值为 83%；XGBoost 模型中，准确率为 85%，F1 值为 83.5%；两者的加权平均模型中，准确率为 86%，F 值为 84%。可见，对单个模型采用加权平均法融合得到的结果优于单模型的结果，再对比本文所使用的卷积神经网络模型，本文所使用的卷积神经网络模型得到的准确率为 91.4%，，F1 值为 90.4%，其效果远远优于随机森林模型、XGBoost 模型以及两者的加权平均模型。

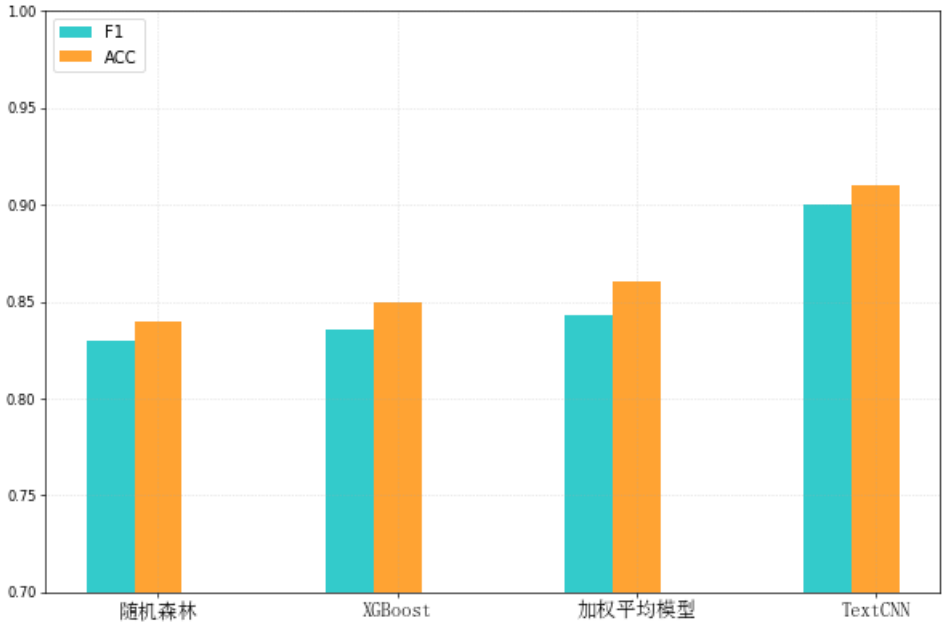

图 2-13 分类模型效果对比

第三章 热点问题挖掘

3.1 数据准备

3.1.1 数据描述

针对任务二的数据集共有 4327 条留言，分为 7 个部分，包括“留言编号”、“留言用户”、“留言主题”、“留言时间”、“留言详情”，“点赞数”与“反对数”。“点赞数”与“反对数”代表了网友对这条留言的认可程度，其余部分同任务一数据集。

3.1.2 数据预处理

对任务二中的留言挖掘热点问题前，需要先对数据集中的“留言主题”与“留言详情”部分进行去除特殊字符、去除重复文本与去除首尾冗余的操作，再对文本分词、去除停用词，这些操作与任务一中的数据预处理阶段一样，不再赘述，但与任务一不同的是，考虑到同一用户在不同时间的相似留言会影响后续热度值的计算，所以本文不对任务二对应的数据集进行相似留言的操作，仅去除完全相同的数据，即“留言用户”、“留言时间”与“留言详情”均相同的数据。

3.2 提取热点问题

本文在提取热点问题之前，首先利用 TextRank 算法对“留言主题”提取关键词，计算关键词的词频，并设定阈值过滤噪声留言。然后对数据集进行“粗分类”，即把每条数据的“留言主题”和“留言详情”部分合并，用 Doc2Vec 训练为句向量，依据语义相似度用 K-means 算法将数据集聚为 7 个大类，每一类都带有特定的主题。最后在每一类内使用命名实体识别以及模糊匹配提取“特定地点”相关的地名，结合高频词提取“特定人群”相关的事件词汇，并在类内精确匹配得到留言数大于 10 的热点问题，最终得到 16 个热点问题。

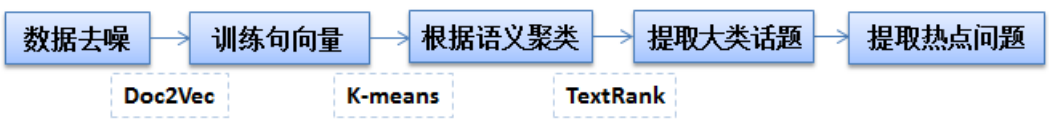

图 3-1 提取热点问题流程图

3.2.1 文本去噪

本文将“噪声留言”定义为：“留言主题”的关键词在全部留言中出现频数小、关注人数少的留言。因为噪声留言不满足成为热点问题的条件，所以在原始数据集中剔除噪声留言[5]不仅能减少相似矩阵的维度，还能提高后续热点抽取的效果。具体步骤为：对每条数据的留言主题提取 3 个关键词，汇总所有关键词并统计词频，提取出词频大于 4 的关键词作为筛选词集；然后利用程序遍历，找到不包含筛选词集的留言文本，将其中“点赞数”与“反对数”均为 0 的留言作为噪声留言剔除。

1．提取关键词

TextRank 算法是一种用于文本的基于图的排序算法，其基本思想来源于 PageRank 算法, 通过把文本分割成若干组成单元并建立图模型, 利用投票机制对文本中的重要成分进行排序。提取关键词是指从一段给定的文本中自动抽取若干有意义的词语或词组。它利用局部词汇之间关系（共现窗口）对后续关键词进行排序，直接从文本本身抽取。

留言详情体现用户的情感，包含大量信息价值不高的词，语句冗长，语言多样，难以把握中心，提取出来的关键词存在并 “不关键”的情况。相对而言，留言主题结构简单，高度概括了留言详情的内容，可以视为一条留言的中心句，因此本文利用 TextRank 算法对数据集每条留言的留言主题提取关键词。表 3-1 为提取出的留言主题关键词示例。

表 3-1 留言主题关键词

留言编号	留言主题	关键词
188006	A3 区一米阳光婚纱艺术摄影是否合法纳税了？	艺术摄影、婚纱、纳税
188007	咨询 A6 区道路命名规划初步成果公示和城乡门牌问题	门牌、公示、城乡
188031	反映 A7 县春华镇金鼎村水泥路、自来水到户的问题	水泥路、金鼎村、到户
188039	A2 区黄兴路步行街大古道巷住户卫生间粪便外排	黄兴路、外排、步行街

2．计算关键词词频

提取出所有留言主题的关键词后，汇总并统计各个关键词的词频，设定阈值为 4，将词频大于阈值的关键词设为筛选词集，图 3-2 为筛选词集词云图。从中可以看出，“西地省”、“扰民”、“噪音”、“星沙”等词具有较高的词频。

图 3-2 筛选词集

3．去除噪声留言

利用 Python 程序对留言主题的关键词进行遍历，将不包含筛选词集的留言标记为 0，包含筛选词集的留言标记为 1。由本文对噪声留言的定义可知，噪声留言还应有关注度低的特点，因此标记为 0 的留言中，“点赞数”与“反对数”均为 0 的留言才能视为噪声留言去除，剩余留言组成的数据集将用于后续提取热点问题。经过以上操作，本文共去除噪声留言 501 条，剩余留言 3825 条。表 3-2 为噪声留言示例。

表 3-2 噪声留言

留言编号	留言主题	点赞数	反对数
188006	A3 区一米阳光婚纱艺术摄影是否合法纳税了？	0	0
188251	A7 县特立路与东四路口晚高峰太堵，建议调整信号灯配时	0	0
188856	A3 区谷园路 39 号维也纳智好酒店有卖淫团伙	0	0

3.2.2 话题聚类

去除噪声留言后的数据集包含了各种话题的留言信息，为了更加精准的找到候选热点问题，本文首先利用 Doc2Vec 工具对合并后的留言训练句向量，再利用 K-means 算法依据语义相似性将数据集聚类成 7 大类，每个类含有特定的主题。然后，对原始数据集提取热点问题的任务将转变为在带有特定主题的 7 类内分别提取热点问题的任务。显而易见，后者更加容易，提取的热点问题更加精确。

1．训练句向量

Doc2Vec 是一种非监督算法，用于获得句子/段落/文档的向量表达[6]。训练出来的向量可以通过计算距离来找句子/段落/文档之间的相似性，常用于文本聚

类问题。Doc2Vec 有两种训练方法，本文使用的是 PV-DBOW。

Distributed Bag of Words version of Paragraph Vector（PV-DBOW）训练方法是忽略输入的上下文，让模型去预测段落中的随机一个单词。在每次迭代的时候，从文本中采样得到一个窗口，再从这个窗口中随机采样一个单词作为预测任务，让模型去预测，输入就是段落向量，如下图 3-3 所示。本文对合并留言部分训练句向量。

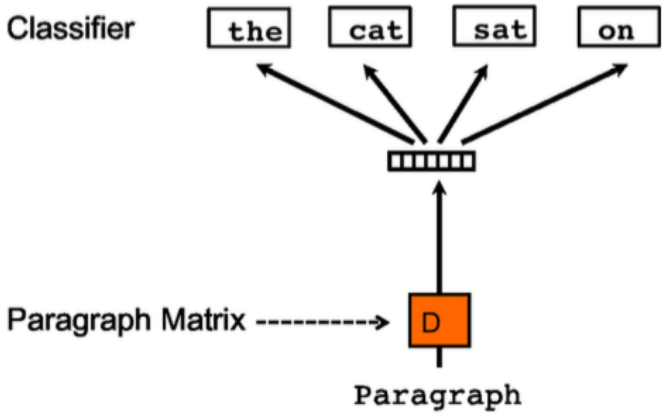

图 3-3 Doc2Vec 原理示意图

2．K-means 聚类

（1）K-means 原理

K-means 算法（K-均值聚类算法）是一种已知聚类类别数的划分算法[7]。它是一种典型的距离聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到的紧凑且独立的簇作为最终目标。

①基本思想

它是基于给定的聚类目标函数，采用迭代更新的方法，每一次迭代过程都是向目标函数减小的方向进行，最终聚类结果使得目标函数取得极小值，达到较好的分类效果。

②算法框架

Step1：给定大小为 n 的数据集，令 $O=l$，选取 K 个初始聚类中心 $Z_j(O)$，$j=1,2,...K$，O 代表不同迭代轮数的聚类中心；

Step2：计算每个样本数据对象与聚合中心的距离 $D(x_i,Z_j(O))$，$i=1,2,...,K$，并分类；

Step3：令 $O=O+1$，计算新的聚类中心和误差平方和准则 f（目标函数）值：$J_e(2)=\sum_{j=1}^{k}\sum_{k=1}^{n}\left\|x_k^{(j)}-Z_j(2)\right\|^2$；

Step4：判断：若 $If\left(O+1\right)-f\left(O\right)I<theta$（ f 收敛）或者对象无类别变化，则算法结束，否则， $O=O+1$，返回 Step2 步。

（2）聚类

本文利用 K-means 算法对数据集聚类[8、9]，将内容相似的留言聚在一起。经过反复试验确定最佳 K 值为 7，当 K 值为 7 时，每个大类的话题最为明显。最终得到第 1 大类 191 条留言，第 2 大类 2207 条留言，第 3 大类 336 条留言，第 4 大类 316 条留言，第 5 大类 294 条留言，第 6 大类 284 条留言，第 7 大类 697 条留言。

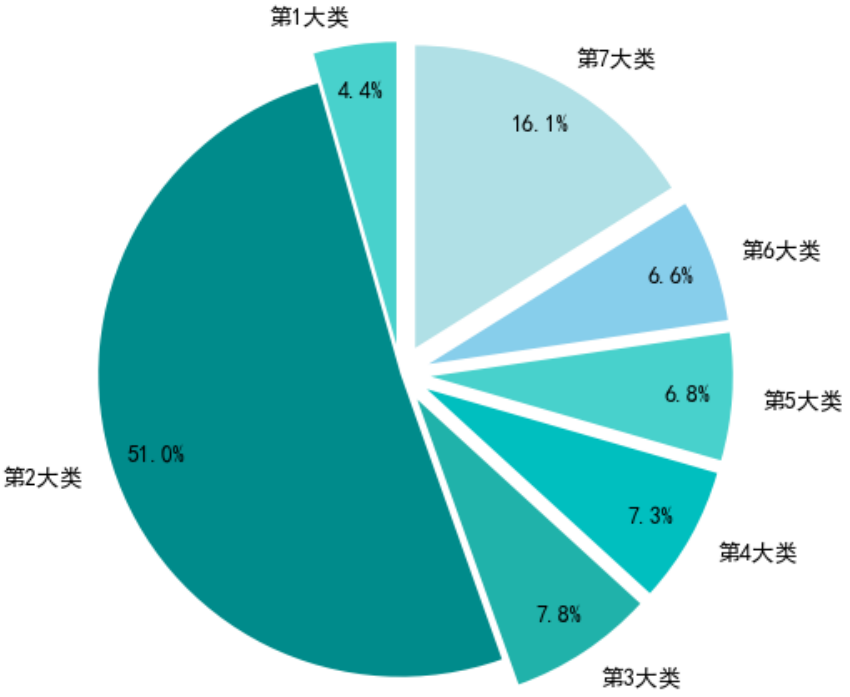

图 3-4 7 大类留言数占比图

3．提取话题

用 TextRank 算法对 7 个大类内部提取话题，表 3-3 显示了 7 个大类的主要话题，第 1 大类的留言主要关于房屋拆迁以及征地补偿等问题；第 2 大类的留言主要关于噪音扰民以及住房相关问题；第 3 大类的留言主要关于小区内物业与业主的纠纷；第 4 大类的留言主要关于小区房屋质量问题；第 5 大类的留言主要关于诈骗活动与拖欠工资；第 6 大类的留言主要关于学生教育；第 7 大类的留言主要是对地铁、公交车等公共交通的建议。

表 3-3 大类话题

大类	话题
1	拆迁、房屋、举报、补偿、村民、安置
2	扰民、噪音、施工、投诉、公积金
3	小区、物业、违规、投诉、问题
4	开发商、小区、质量、房屋、投诉
5	公司、涉嫌、诈骗、拖欠、举报
6	学生、学校、教育、培训机构、学区
7	建议、地铁、公交车、乱停、居民出行

为了更加直观的体现每个大类的话题，本文对每个大类“合并留言”中的所有词汇计算词频，并绘制词云图。对比表 3-3 的大类话题与图 3-5 各大类的词云图，可以看出，各个大类按话题归类的效果较明显，每一大类能比较好的代表对应话题。

第 1 大类

第 2 大类

第 3 大类

第 4 大类

第 5 大类

第 6 大类

第 7 大类

图 3-5 7 大类词云图

3.2.3 提取热点

经过上述步骤，原来混乱、复杂的留言数据集已被分成了带有话题的大类，对每个大类提取热点问题是本节需要完成的任务。热点问题是反映某一时段内“特定地点”或“特定人群”的问题。反映“特定地点”的热点问题会反复出现同一地名，因此本文同时使用命名实体识别与模糊匹配提取相关地名，选择可能的地名进行类内精确匹配找出热点问题留言条目；而反映“特定人群”的热点问题会反复出现同一事件词汇，如：“58 车贷”，因此本文对“留言主题”提取高频词，选择可能的事件词汇进行类内精确匹配。

由于热点问题数量众多，且本文的目标是提取出热度值排名前 5 的热点问题，因此在提取热点问题时，本文将不考虑留言数较少的问题。

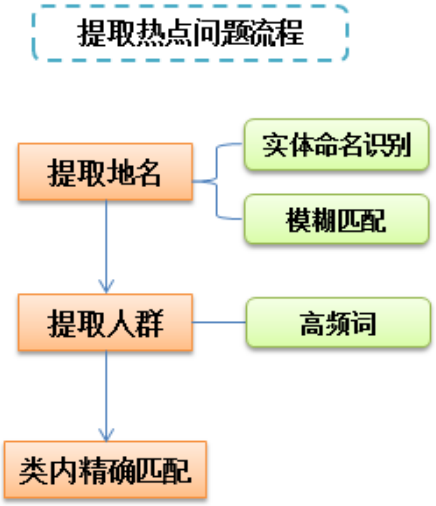

图 3-6 提取热点问题流程

1．识别特定地点和人群

（1）命名实体识别

利用 HanLp 中命名实体识别技术对每个大类的“留言详情”识别地名，并统计其出现的次数，如“新城”这个词出现 37 次。选择词频大于 10 的有效地名作为热点问题筛选词汇，但是命名实体识别技术无法识别所有特定地点的名称，像“小区”名称所提取的并不多。

（2）模糊匹配

模糊匹配的思想是设置特定格式后，利用正则表达式在“留言详情”中提取符合这种格式的所有词汇，如设置格式为“XX 小区”，将提取出“新城小区”、“之城小区”等结果。将得到的小区名称在类内统计词频，同样选择词频大于 10 的有效地名作为热点问题筛选词汇；

（3）高频词

对每个大类的高频词进行整理，删除地名相关的高频词，提取可能的事件词汇作为热点问题筛选词汇，在各大类内进行匹配提取出“特定人群”相关的热点问题留言，对出现次数大于 10 的高频词选择有效词频作为热点问题筛选词汇。

以第 2 大类为例用于说明结合使用命名实体识别、模糊匹配与高频词提取热点问题的效果。首先使用命名实体识别后提取地名，得到的结果为：

（新城，37）（暮云街道，16）

使用模糊匹配得到的结果为：

（滨河苑，52）（新城小区，48）（暮云街道，29）（星沙街道，18）

（之城小区，18）（泉塘街道，17）（国际城，10）

对高频词进行处理后，得到以下筛选词汇：

(公积金，17)（人才，21）（户口，10）

2．精确匹配热点问题

根据筛选的词汇对数据集进行精确匹配，即利用程序在每类内根据筛选词汇遍历匹配，将属于同一热点问题的留言归类。最后对每一大类按照上述步骤提取出热点问题后，一共得到 16 个热点问题。如第 2 大类可得到以下 7 个热点问题。

表 3-4 第 2 大类热点问题

问题 ID	代表留言主题	地点/人群
0	关于伊景园滨河苑捆绑销售车位的维权投诉	伊景园滨河苑
1	投诉 A2 区丽发新城附近建搅拌站噪音扰民	丽发新城
2	A7 县星沙四区凉塘路的旧城改造要拖到何时？	凉塘路
3	A 市万科魅力之城商铺无排烟管道，小区内到处油烟味	魅力之城
8	询问 A 市住房公积金贷款的相关问题	住房公积金
9	咨询 A 市人才购房补助发放问题	人才
11	A3 区中海国际社区空地夜间施工噪音太大了	中海国际社区

3.3 热度度量

提取出热点问题后，本文已将某一时段内反映特定地点或特定人群问题的留言进行归类，为了找出最具热度的前 5 个热点问题，本文首先使用本文提出的异常时间点留言识别算法剔除每个热点问题中时间异常的留言，随后定义了问题相关留言数量比、问题相关用户数量比、问题相关留言关注度、问题相关留言集中度四个热度评价指标，利用乘法合成法来合成熵值法和变异系数法两种方法确定权重，再通过综合评价法 TOPSIS 计算各个候选热点问题的热度值，取排名前 5 作为本文所求的热度值前 5 的热点问题。

3.3.1 异常时间识别

在每个热点问题中会存在某两条留言之间留言时间间隔异常的问题。以第 2 大类中问题 ID 为 1 的热点问题为例，下图 3-7 为该热点问题在不同时间的留言数量。

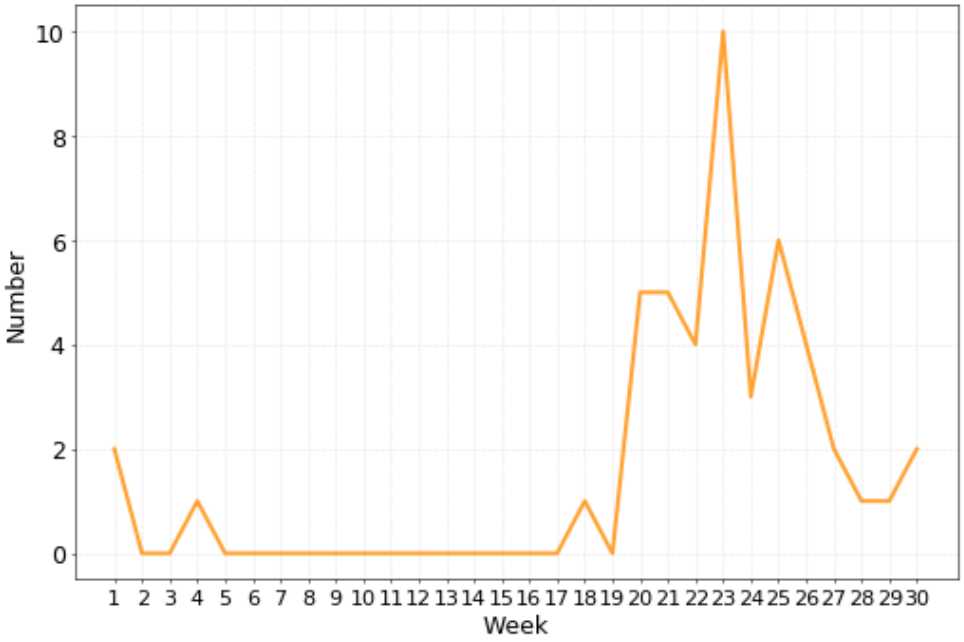

图 3-7 问题 ID 为 1 的热点问题各时间点留言数

从上图可以看出，在第 1 周有 2 条留言，第 4 周仅有 1 条留言，从第 4 周到第 18 周隔了近 3 个月，问题相关留言数为 0。在后续计算热度值时无疑会考虑热点问题的持续时长，如果把第 1 周与第 4 周的一条留言也算入其中，会影响热点问题的热度值，所以将第 1 周与第 4 周的留言视为异常去除。

本文针对上述这种情况设计了一种异常时间点识别算法，在提取热点问题之后对每一类内识别出异常时间点的留言并将其从中剔除。首先，将每一类留言类别中的留言数据按留言时间升序重新排列，计算相邻两条留言之间的留言时间差，其中，留言时间差等于后一条留言的留言时间减去前一条留言的留言时间；遍历所有留言时间差，若某一留言时间差大于 90 天，则计算得出此留言时间差的后一条留言之前的所有留言数量，若留言数量小于此类留言类别总数量的三分之一，则继续寻找下一个大于 90 天的留言时间差，若留言数量大于此类留言类别总数量的三分之一，则停止遍历，删除相应异常时间点留言；上述内容为按留言时间升序删除异常时间点留言，此外，还需按留言时间降序删除异常时间点留言，步骤与升序相似，但略有不同，具体算法如下所述：

Step1：将第 m 类留言类别中的留言按留言时间升序重新排列后记为 mx1，求出相邻两条留言的留言时间差序列 delta1，delta1 序列中的第 i 个值等于 mx1 中第 i+1 条留言的留言时间减去第 i 条留言的留言时间；

Step2：找出 delta1 序列中大于 90 的数值所处位置序列 loc1，loc1 中第 j 个数值为 delta1 中第 j 个大于 90 的数值所处位置；

Step3：设 mx1 中的留言数量为 n，则 r1 为 loc1 序列中小于 n/3 数值中的最

大值；

Step4：将 mx1 中的第 1 条到第 r1 条留言数据删除，记为 mxx；

Step5：将 mxx 中的留言按留言时间降序排列后记为 mx2，求出相邻两条留言的留言时间差序列 delta2，delta2 序列中的第 i 个值等于 mx2 中第 i+1 条留言的留言时间减去第 i 条留言的留言时间的绝对值；

Step6：找出 delta2 序列中大于 90 的数值所处位置序列 loc2，loc2 中第 j 个数值为 delta2 中第 j 个大于 90 的数值所处位置；

Step7：找出 r2，r2 为 loc2 序列中小于 n/3 数值中的最大值；

Step8：将 mx2 中的第 1 条到第 r2 条留言数据删除，记为 mz，则 mz 即为第 m 类留言类别中剔除异常时间点留言后的最终数据。

3.3.2 定义热度指标

热点问题是指某一时间段内反映“特点地点”或“特点人群”的一系列问题，由此可见，在定义热度评价指标时，需要综合考虑时间段、问题相关留言条目数量、问题相关留言的用户情况以及某一问题的受关注程度。从上述几个方面出发，可以更加合理地定义热度指标，确定热点问题。因此通过查阅相关文献[10、11]后，选取以下指标进行热度评价：

1．问题相关留言数量比（X1）

一个热点问题必须要在留言数量上有所保证，没有达到一定的留言条目数量则说明这类问题反映的人较少，那么在这段时期内该问题不可能成为热点问题。问题相关留言数量比是指在该问题出现的时间段内，问题相关留言条目数与所有留言条目数之比；

2．问题相关用户数量比（X2）

由于一个热点问题必定会被大量用户反映，因此针对某一问题发布留言的用户数量也可以反映热度。问题相关用户数量比是指在该问题出现的时间段内，发布问题相关留言的用户数量与所有在该时间段发布留言的用户数量之比；

3．问题相关留言关注度（X3）

针对某些问题，某些用户不会自己发表留言，但会对该问题的相关留言进行表态，比如：“点赞”或“反对”。“点赞”则表示其认为这一留言所反映的问题同样对他造成了困扰，“反对”则表示其认为这一留言所反映问题并不属实。但不论是“点赞”还是“反对”，都表明这一问题得到了群众的关注，且存在争议。因此，某一问题相关留言的“点赞数”和“反对数”也可以反映热度。问题相关留言关注度是指在该问题出现的时间段内，问题相关留言的“点赞数”与“反

对数”总和与该时间段内所有留言的“点赞数”与“反对数”总和之比；

4．问题相关留言集中度（X4）

一个热点问题在短时间内应当具有较高的留言集中度。若一个问题的相关留言数量很多，但却分布在较长的时间段上，导致一定时间段内的相关数目很少，则不能称该问题为热点问题。问题相关留言集中度是该问题的留言条目数与该问题出现时间段之比。

上述四项指标的具体公式计算见表 3-5，其中：

n、m、p——分别代表问题 C 所包含的留言数量、留言用户数量（即问题 C 中所包含用户数量）、问题 C 中所有留言的“点赞数”和“反对数”总和；

T_e、T_b——分别代表问题 C 中最晚的留言时间和最早的留言时间；

N、M、P——分别代表在时间T_e和T_b内所包含的全部留言数量、留言用户数量以及全部留言的“点赞数”和“反对数”总数。

表 3-5 热度指标体系

指标	计算方法
问题相关留言数量比	$\frac{n}{N}$
问题相关用户数量比	$\frac{m}{M}$
问题相关留言关注度	$\frac{p}{P}$
问题相关留言集中度	$\frac{n}{T_e - T_b}$

3.3.3 热度计算

基于上节提出的热度评价指标，本文采取以下步骤计算 16 个热点问题的热度值：首先，分别采用熵值法和变异系数法得出四个热度指标的两种不同权重；随后，采用乘法合成法将两种权重组合得出组合权重；最后，基于组合权重，使用 TOPSIS 综合评价法得出各热点问题的热度值进行后续分析。

1．方法原理

（1）熵值法

在信息论中，熵是对不确定性的一种度量。信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也越大[12]。根据熵的特性，可以

通过计算熵值来判断一个事件的随机性及无序程度，也可以用熵值来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的影响越大。因此，可根据各项指标的变异程度，利用信息熵这个工具，计算出各个指标的权重，为多指标综合评价提供依据。具体步骤如下：

Step1：选取 n 个样本， m 个指标， x_{ij} 则为第 i 个样本的第 j 个指标的数值 $(i=1,2,...,n;\ j=1,2,...,m)$；

Step2：指标的标准化处理——异质指标同质化。由于各项指标的计量单位并不统一，因此在用它们计算综合指标前,先要对它们进行标准化处理，即把指标的绝对值转化为相对值，从而解决各项不同质指标值的同质化问题。并且，由于正向指标和负向指标数值代表的含义不同（正向指标数值越高越好,负向指标数值越低越好），因此，对于正负向指标需要采用不同的算法进行数据标准化处理；

正向指标处理方法： $x_{ij}^{'}=\dfrac{x_j-x_{\min}}{x_{\max}-x_{\min}}$

负向指标处理方法： $x_{ij}^{'}=\dfrac{x_{\max}-x_j}{x_{\max}-x_{\min}}$

Step3：计算第 j 项指标下第 i 个样本占该指标的比重；

$$y_{ij}=\frac{x_{ij}^{'}}{\sum_{i=1}^{n}x_{ij}^{'}}\left(0\le y_{ij}\le 1\right)$$

Step4：计算第 j 项指标的信息熵值，其中 $e_j=-K\sum_{i=1}^{n}y_{ij}\ln y_{ij}$ ， $K=\dfrac{1}{\ln n}$ 和信息效用值 $d_j=1-e_j$；

Step5：计算各评价指标权重 $w_j=\dfrac{d_j}{\sum_{i=1}^{m}d_j}$。

（2）变异系数法

变异系数法是直接利用各项指标所包含的信息，通过计算得到指标的权重。是一种客观赋权的方法[13]。此方法的基本做法是：在评价指标体系中，指标取值差异越大的指标，也就是越难以实现的指标，这样的指标更难反映被评价单位的差距。具体步骤如下：

Step1：计算每个指标的平均值 $\bar{x}$，标准差 σ；

Step2：计算每个指标的变异系数 $v_i = \frac{\sigma_i}{\bar{x}_i}\left(i = 1,2,...,n\right)$；

Step3：计算每个指标的权重 $w_i = \frac{v_i}{\sum_{i=1}^{n} v}$。

（3）乘法合成法

将熵值法和变异系数法得到的权重利用乘法合成法进行组合，即将两种方法得出的某一指标的权重相乘，然后再进行加权调整得到组合权重。

（4）TOPSIS 综合评价法

TOPSIS 法是一种常用的组内综合评价方法，能充分利用原始数据的信息，其结果能精确地反映各评价方案之间的差距[14]。基本过程为基于归一化后的原始数据矩阵，采用余弦法找出有限方案中的最优方案和最劣方案，然后分别计算各评价对象与最优方案和最劣方案间的距离，获得各评价对象与最优方案的相对接近程度，以此作为评价优劣的依据。该方法对数据分布及样本含量没有严格限制，数据计算简单易行。具体步骤如下所示：

Step1：指标属性同向化，一般选择指标正向化：TOPSIS 法使用距离尺度来度量样本差距，使用距离尺度就需要对指标属性进行同向化处理（若一个维度的数据越大越好，另一个维度的数据越小越好，会造成尺度混乱）。通常采用成本型指标向效益型指标转化（即数值越大评价越高，事实上几乎所有的评价方法都需要进行转化），此外，若需要使用雷达图进行展示，则需将所有数据都变成正数；

Step2：构造归一化初始化矩阵；

Step3：确定最优方案和最劣方案；

最优方案 Z^+ 由 Z 中每列元素的最大值构成：

$$Z^+ = \left(\max\{z_{11}, z_{21}, ..., z_{n1}\}, \max\{z_{12}, z_{22}, ..., z_{n2}\}, \cdots, \max\{z_{1m}, z_{2m}, ..., z_{nm}\}\right) = \left(Z_1^+, Z_2^+, \cdots, Z_m^+\right)$$

最劣方案 Z^- 由 Z 中每列元素的最大值构成：

$$Z^- = \left(\min\{z_{11}, z_{21}, ..., z_{n1}\}, \min\{z_{12}, z_{22}, ..., z_{n2}\}, \cdots, \min\{z_{1m}, z_{2m}, ..., z_{nm}\}\right) = \left(Z_1^-, Z_2^-, \cdots, Z_m^-\right)$$

Step4：计算各评价对象与最优方案、最劣方案的接近程度；

$D_i^+ = \sqrt{\sum_{j=1}^{m} w_j \left(Z_j^+ - z_{ij}\right)^2}$；$D_i^- = \sqrt{\sum_{j=1}^{m} w_j \left(Z_j^- - z_{ij}\right)^2}$，其中 w_j 为第 j 个属性的权

重（重要程度）。

Step5：计算各评价对象与最优方案的贴近程度C_i；其中，$C_i=\frac{D_i^-}{D_i^+ + D_i^-}$，

$0 \le C_i \le 1$，$C_i \to 1$表明评价对象越优；

Step6：根据C_i大小进行排序，给出评价结果。

2．计算热度值

（1）组合权重

根据熵值法和变异系数法得出的各指标权重以及利用乘法合成法得到的组合权重如表 3-6 所示。

表 3-6 指标权重

方法	X1	X2	X3	X4
熵值法权重	0.2492	0.2494	0.2476	0.2538
变异系数法权重	0.2079	0.2078	0.3250	0.2592
组合权重	0.2074	0.2074	0.3219	0.2633

（2）计算热度值

基于乘法组合权重，使用 TOPSIS 综合评价法计算热点问题的热度值并排序，表 3-7 显示了热度值排名前 5 的热点问题指标值。

表 3-7 TOP5 热点问题热度值

问题 ID	热度值	X1	X2	X3	X4
1	0.6226	0.6286	0.6176	0.3333	0.5176
0	0.6099	0.4561	0.4537	0.2313	1.0000
15	0.4307	0.1029	0.1019	0.8557	0.0773
3	0.2404	0.1552	0.1522	0.2466	0.2727
14	0.1964	0.2250	0.2000	0.0270	0.1837

从表 3-8 可知热度值排名第 1 的类别号为 1，其问题相关留言数量比指标值为 0.6226，问题相关用户数量比指标值为 0.6286，问题相关留言关注度指标值为 0.3333，问题相关留言集中度指标值为 0.5176。由此可见，类别号为 1 的热点问题相关留言数量占该问题出现时间段内所有留言数量的 62%以上，发表该热点问题的留言用户数占该问题出现时间段内所有发表留言的用户数量的 62%以上，受关注度为 33%以上，且该热点问题留言时间较为集中。可见，本文设计的热点指标评价体系效果较好。

表 3-8 是热度值排名前 5 的热点问题的代表“留言主题”、相关地点或人群与留言数量。

表 3-8 TOP5 热点问题代表留言主题

问题 ID	代表留言主题	地点/人群	留言数量
1	投诉 A2 区丽发新城附近建搅拌站噪音扰民	丽发新城	45
0	关于伊景园滨河苑捆绑销售车位的维权投诉	伊景园滨河苑	52
15	请书记关注 A 市 A4 区 58 车贷案	58 车贷	14
3	A 市万科魅力之城商铺无排烟管道，小区内到处油烟味	魅力之城	18
14	A3 区青青家园小区公共通道被架了烟道	青青家园	10

3.4 事件提取

在找到排名前 5 的热点问题后，使用 TextRank 算法提取每个热点问题的关键词和关键句[15]，经过整理汇总得到热点问题表，见下表 3-9。热度值排名前 5 的热点问题明细表见附件“热点问题留言明细表”。

表 3-9 TOP5 热点问题汇总表

热度排名	问题 ID	热度值	时间范围	地点/人群	问题描述
1	1	0.6226	2019/070/2 至 2020/01/26	A 市丽发新城小区	搅拌站噪音污染
2	0	0.6099	2019/7/11 至 2019/09/01	A 市伊景园滨河苑	车位捆绑销售
3	15	0.4307	2019/01/08 至 2019/07/08	A 市西地省 58 车贷案件	58 车贷案件案情拖延
4	3	0.2404	2019/07/21 至 2019/09/25	A 市 A5 区魅力之城	小区餐馆油烟噪声扰民
5	14	0.1964	2019/4/22 至 2020/01/08	A 市 A3 区青青家园小区	消防通道违法架设烟管

第四章 答复意见评价体系

4.1 数据准备

4.1.1 数据描述

针对任务三的数据集共有 2816 条留言，分为 7 个部分，包括“留言编号”、“留言用户”、“留言主题”、“留言时间”、“留言详情”、“答复意见”与“答复时间”。“答复意见”是有关部门对相关留言做出的回复，“答复时间”是相关部门做出回复的时间，其余部分同任务一数据集。

4.1.2 数据预处理

对任务三中的“答复意见”质量进行评价时，为提高评价的效果，需要先对“留言主题”、“留言详情”与“答复意见”进行去除特殊字符的操作，然后去除“答复意见”中字数不足 10 个的留言，最后对文本分词、去除停用词。这些操作与任务一中的数据预处理阶段一样，不再赘述。

4.2 评价指标体系

对相关部门的答复内容进行评价时，为了全方位、多角度的考查答复质量，本文通过查阅文献[16-19]，从相关性、完整性、可解释性和及时性四个方面选取了 8 个指标，构成“答复意见”质量评价指标体系。

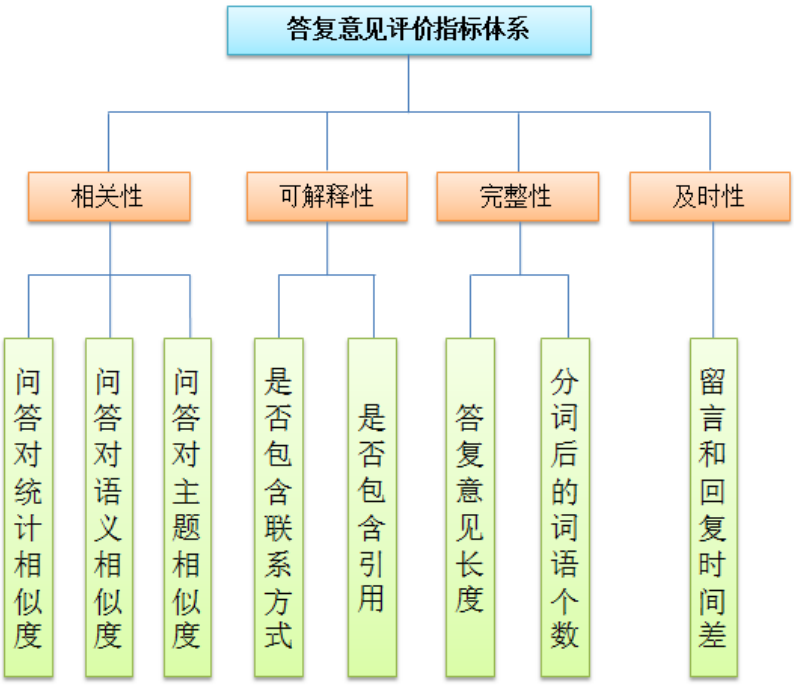

图 4-1 答复意见评价指标体系

4.2.1 相关性

从相关性角度来看，相关部门的答复意见必须与留言问题具有一定的相关性。如果答复意见与留言问题不相关，即所谓的文不对题、答非所问，则认为此答复意见的质量较低，因此本文选取相关性作为一个一级指标，并选取以下指标作为相关性的二级指标。

1．问答对统计相似度（X1）

该指标指的是在分词及去停用词之后，一条留言与它的相应的答复意见之间重叠词的个数。一个高质量的答复意见应当包含一个相当大比例的同样出现在留言问题中的词语。

2．问答对语义相似度（X2）

该指标指的是留言与它的答复意见之间语义的相似性。回复内容应该与留言内容具有一定的语义相似，这样才能说明该答复意见是针对该留言问题提出的。本文使用 Word2Vec 工具对一个句子中的所有词进行加权平均求得问答对的句向量，然后计算问答对之间的余弦相似度。

3．问答对主题相似度（X3）

该指标指的是留言主题与答复意见所表达主题的相似性，值越大，越能说明问答对具有一样的主题，体现问答对的相关性。本文使用 LDA 模型分别提取问答对的相应主题，再计算主题相似度。LDA 是一种三层贝叶斯主题模型，通过

无监督的学习方法发现文本中隐含的主题信息，目的是要以无指导学习的方法从文本中发现隐含的语义维度。

4.2.2 可解释性

从可解释性角度来看，相关部门的答复意见应具有可解释性，即答复意见中所给的建议、解决问题方案以及解决效果等应有理有据，具有较强的可解释性，不能凭空捏造，因此，本文选择可解释性作为一个一级指标，并选取以下指标作为可解释性的二级指标。

1．答复意见中是否包含联系方式（X4）

该指标指的是答复意见中是否包含电子邮箱、网址链接与电话号码。一个高质量的答复意见不仅要能提出解决问题的办法，还应该提供给留言用户回访的机会，比如电子邮箱、网址与电话号码。

2．答复意见中是否包含引用（X5）

该指标指的是答复意见中是否包含法律法规的引用。一个高质量的答复，应该包含相关法律法规及规定等权威信息来源的引用，据此来给出相应回复建议及解决方案，来确保答复的权威性和可信性。

4.2.3 完整性

从完整性角度来看，一个高质量的答复意见应该是完整的，对问题的方方面面都进行了回复，包括了解决方案的方方面面，因此，本文选择完整性作为一个一级指标，并选取以下指标作为完整性的二级指标。

1．答复意见长度（X6）

该指标指的是相关部门答复意见的句子长度。一个高质量的答复必然涵盖了问题的方方面面，则应该拥有一个较长的长度。

2．答复意见分词后的词语个数（X7）

该指标指的是答复分词以及去停用词之后的词语个数，上述分析中已经表明一个高质量的答复应该拥有一个较长的长度，但一个较长答复中可能包括大量无用的停用词，因此本文认为一个高质量的答复在分词及去停用词之后必然包括较多的词语。

4.2.4 及时性

从及时性角度来看，及时对用户的留言问题进行回复，这体现了相关部门解决问题的效率。因此，本文选取及时性作为一个一级指标，并选取以下指标作为

及时性的二级指标。

1．留言和回复时间差（X8）

该指标指的是相关部门给出答复意见的时间与问题留言的时间差，时间差越短，说明相关部门回复的速度越快，时间差越长，说明回复速度越慢。

4.3 答复意见质量评价

基于上述所建立的评价指标体系，本文首先建立以下模型对相关部门答复意见的质量进行评价：以上述 8 个二级评价指标作为特征，利用 K-means 算法将原始“答复意见”聚为三类，并计算每一类中各指标的均值以判断每一类别所对应的标签（高、中、低）。

4.3.1 聚类结果

以上述建立的 8 个评价指标为特征，利用 K-means 算法将“答复意见”聚为三类，即高质量答复意见、中等质量答复意见、低质量答复意见。图 4-2 是对“答复意见”聚类的效果图。

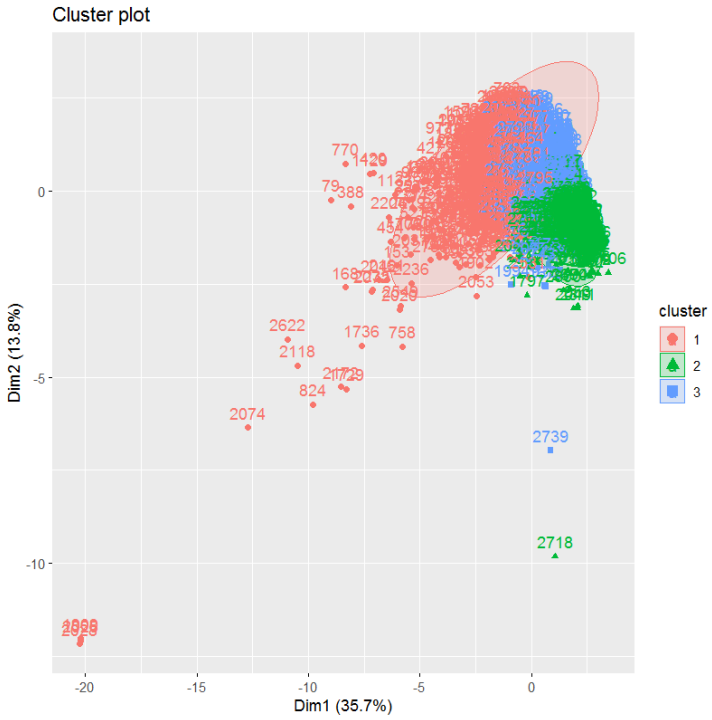

图 4-2 聚类效果图

由图 4-2 中可以看出，聚类效果良好，但由于聚类结果中的类别数字并无实际意义，不能代表答复意见质量的高低，因此本文计算了各类别下各个指标的均值，用以判断各类别所对应标签，各指标均值如表 4-1 所示。根据各类别下各指

标的均值，综合判断得出各类别对应标签为：1 类别为高质量，2 类别为低质量，3 类别为中等质量。

表 4-1 各类别下各指标均值

类别	X1	X2	X3	X4	X5	X6	X7	X8
1 类别	13.76	0.45	0.88	0.31	0.92	613.61	161.02	20.59
2 类别	3.32	0.35	0.32	0.13	0.13	147.42	36.22	20.13
3 类别	6.84	0.40	0.92	0.21	0.01	259.67	66.46	19.25

4.3.2 结果分析

本文对数据集中所有“答复意见”的评价（高质量、中等质量、低质量）见附件“答复意见质量评价表”，从中可以看出，本文对答复意见的质量评价非常合理，能正确评价答复意见的质量。

表 4-2 是部分“答复意见”的评价指标值和类别。从中可以看出，高质量的“答复意见”问答对统计相似度（X1），问答对语义相似度（X2）和问答对主题相似度（X3）都比较高，说明留言内容和回复内容具有高度相关性。高质量的“答复意见”一般会包含联系方式（X4）和引用（X5），说明“答复意见”具有可解释性。高质量的“答复意见”长度（X6）较长，分词后的词语个数（X7）也较多，说明答复内容较完整。高质量的“答复意见”与留言的时间差（X8）都比较短，体现了高质量答复的及时性。而中等质量与低质量的“答复意见”在这 8 个指标上的表现相对较差。

表 4-2 答复意见评价指标值

ID	X1	X2	X3	X4	X5	X6	X7	X8	质量
2549	18	0.44	1.00	0	1	426	108	15	高
2555	9	0.42	0.94	0	1	348	98	14	高
2557	7	0.50	1.00	1	1	287	73	14	高
2554	2	0.41	0.68	0	0	293	74	14	中
2574	8	0.37	1.00	0	0	152	37	15	中
3683	10	0.44	0.52	0	0	464	119	16	中
2759	2	0.30	0.28	1	0	191	31	31	低
3684	12	0.44	0.47	0	0	224	52	16	低
3685	10	0.54	0.20	0	0	451	119	70	低
…	…	…	…	…	…	…	…	…	…

4.4 模型构建

对“答复意见”按照指标值聚类后，所有“答复意见”已被打上高质量、中等质量、低质量三类标签。由于 K-means 算法中初始聚类中心的选择具有随机性，导致聚类结果并不唯一，使得对“答复意见”的评价具有随机性和不唯一性。因此，鉴于本次聚类结果很好，本文利用已有标签对“答复意见”的质量评价构建分类器，后期直接利用分类器对“答复意见”进行质量评价。考虑到数据类别并不均衡，因此本文建立二阶段分类器进行分类，先将原始数据分为高和非高分类的数据集以及中和非中分类的数据集，第一阶段分类器由高和非高分类的数据集训练而得，用以得出数据的高和非高分类结果，第二阶段分类器由中和非中分类的数据集训练而得，用以得出数据的中和非中分类结果，用此二阶段分类器可对“答复意见”的质量进行快速、有效评价。

4.4.1 二阶段分类器

考虑到数据类别并不均衡，因此本文建立二阶段分类器进行分类：首先，将原始数据分为高和非高分类的数据集以及中和非中分类的数据集，使得数据类别均衡，其次，用这两个数据集分别训练第一阶段分类器和第二阶段分类器，训练过程中，使用十折交叉验证分别得到各阶段 Boosting、Bagging 以及 Stacking 三种集成学习算法下各算法的分类效果，最终选取各阶段分类效果最好的算法作为各阶段分类器。二阶段分类器分类过程如图 4-3 所示：

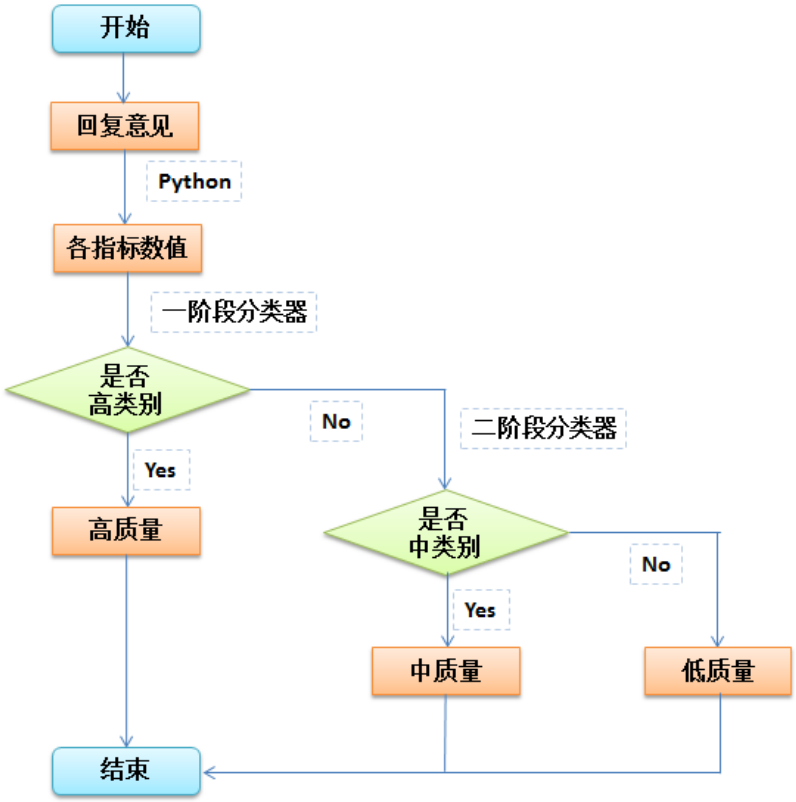

图 4-3 二阶段分类器分类过程

1．集成学习算法原理

（1）Boosting 算法原理

初始对每个样本分配相同的权重，每次经过分类，把对的结果的权重降低，错的结果权重增高，如此往复，直到阈值或者循环次数[20]。梯度提升算法首先给定一个目标损失函数，它的定义域是所有可行的弱函数集合；提升算法通过迭代的选择一个负梯度方向上的基函数来逐渐逼近局部极小值。这种在函数域的梯度提升观点对机器学习的很多领域有深刻影响。

提升的理论意义：如果一个问题存在弱分类器,则可以通过提升的办法得到强分类器。

给定输入向量 X 和输出变量 Y 组成的若干训练样本$(x_1,y_1),\ldots,(x_n,y_n)$，目标是找到近似函数$\hat{F}(\vec{x})$，使得损失函数$L\left(y,\hat{F}(\vec{x})\right)$的损失值最小。损失函数的典型定义为：$L\left(y,F(\vec{x})\right)=\frac{1}{2}\left(y-F(\vec{x})\right)^2$。

假定最优函数为$F^*(\vec{x})$，即$F^*(\vec{x})=\arg\min \mathrm{E}_{(x,y)}\left[L\left(y,F(\vec{x})\right)\right]$，假定最优函数是一组基函数$f_i(\vec{x})$的加权和$F(\vec{x})=\sum_{i=1}^{M}\gamma_i f_i(\vec{x})+C$，梯度提升方法寻找最优解

$F(\vec{x})$，使得损失函数在训练集上的期望最小，步骤如下：

Step1：初始给定模型为常数 $F_0(\vec{x})$，对于 $m=1$ 到 M；

Step2：计算伪残差 $\gamma_{im}=\left[\frac{\partial L(y_i,F(\vec{x}_i))}{\partial F(\vec{x}_i)}\right]_{F(\vec{x})=F_{m-1}(\vec{x})}, i=1,2,\ldots n$；

Step3：使用数据 $(\vec{x}_i,\gamma_{im})_{i=1}^{n}$ 计算拟合残差的基函数 $f_m(x)$；

Step4：计算步长 $\gamma_m=\arg\min\sum_{i=1}^{n}L(y_i,F_{m-1}(\vec{x}_i)-\gamma * f_m(\vec{x}_i))$；

Step5：更新模型 $F_m(\vec{x})=F_{m-1}(\vec{x})-\gamma_m f_m(\vec{x}_i)$。

（2）Bagging 算法原理

Step 1： 通过自助法（有放回抽样）生成 K 个数据集，即在所有的样本中通过有放回的随机抽样，生成 K 个数据集。

Step 2： 对这 K 组数据集分别进行训练，从而得到 K 个分类器

Step 3： 将这 K 个分类器组合到一起，各个分类器的权重相同，从而得到最终的分类器。

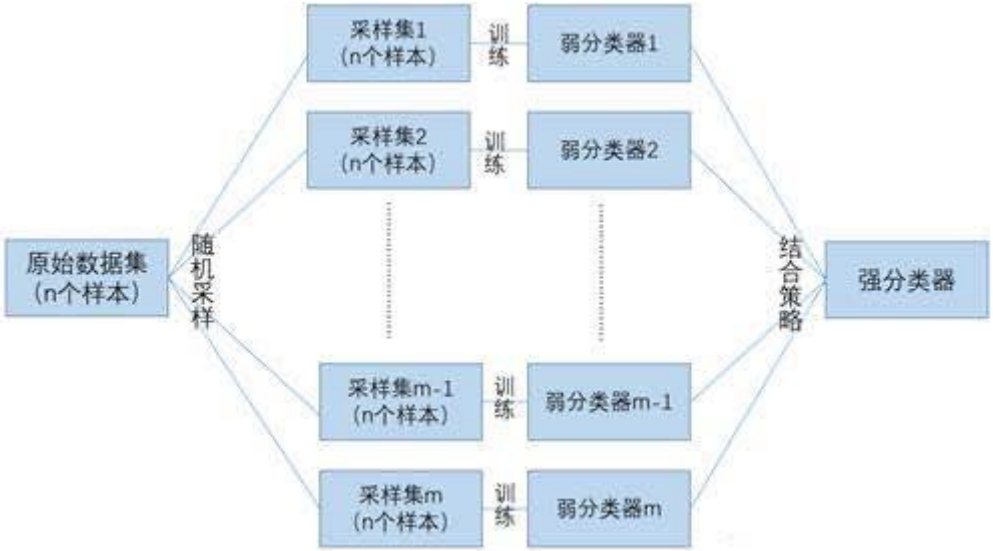

图 4-4 Bagging 算法流程图

（3）Stacking 算法原理

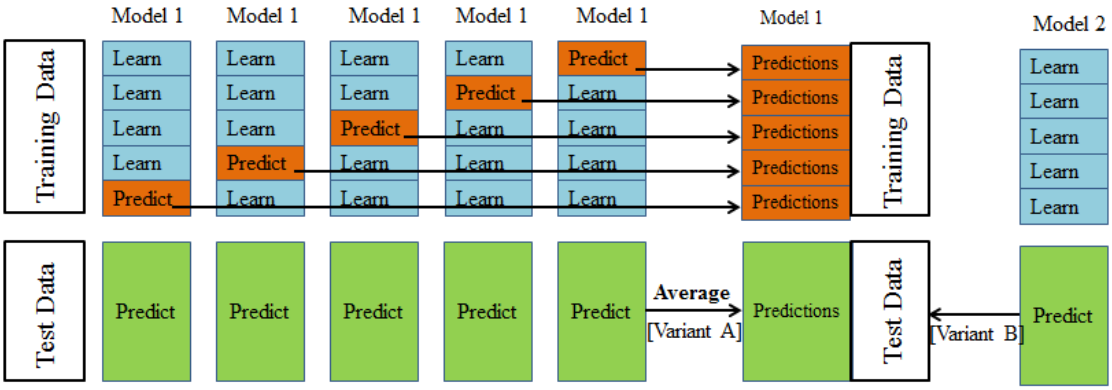

图 4-5 Stack 算法流程图

以图 4-5 为例，现在有训练集 train_x，train_y,测试集 test

Step1：首先选择一种模型例如随机森林 RF（未经训练）；

Step2：假设训练集被均分成 5 份，把其中四份作为小的训练集 s_train_x,s_train_y 另外一份作为小的测试集 s_test，测试集 test 不变；

Step3 以 s_train_x,s_train_y 训练 RF 模型，训练出的模型预测 s_test 得出对应的 s_pred，再预测 test 得出 y_pred；

Step4：在训练集再选择另外一份作为小的测试集 s_test_x,其他四份作为训练集训练模型 RF；

Step5：重复 Step2，Step3，Step4 步骤五次，会得到五个 s_pred 和五个 y_pred，五个 s_pred 作为一个 train_X，原始的 train_y 作为 train_Y 训练模型得到模型 G，五个 y_pred 取个平均值作为新的 test_X,把 test_X 带入到模型 G 中得出预测结果；

Step6：以上就是 Stacking 的第一层，在第二层中，以第一层的输出 train 再结合其他的特征集再做一层 Stacking。不同的层数之间有各种交互，还有将经过不同的 Preprocessing 和不同的 Feature Engineering 的数据用 Ensemble 组合起来的做法。

上述是以一种模型随机森林进行模型训练，也可用于其他不同模型。本文所采用的就是不同种模型的 Stacking 算法。

2．数据再归类

根据每类别所对应标签，将原始数据集分为高和非高数据集以及中和非中数据集，各数据集示例如表 4-3 所示：

表 4-3 数据再归类结果

ID	原类别	高和非高	中和非中
2549	1	高	非中
2554	3	非高	中
2555	1	高	非中
2557	1	高	非中
2574	3	非高	中
2759	2	非高	非中
2849	1	高	非中
3681	1	高	非中
3683	3	非高	中
...	...	...	...

3．第一阶段分类器——高和非高

训练第一阶段分类器时，本文基于高和非高数据集，使用十折交叉验证分别得出 Boosting、Bagging 以及 Stacking 这三种集成学习算法下各个模型的分类效果并进行比较，选出分类效果最好的模型作为第一阶段分类器，用来进行数据高和非高类别的判断。其中，Boosting 集成算法下选取 C5.0 和 Stochastic Gradient Boosting 这两种模型分别进行十折交叉验证；Bagging 集成学习算法下选取 Bagged CART 和 Random Forest 这两种模型分别进行十折交叉验证；Stacking 集成学习算法下选取以简单线性模型为元分类器组合子模型的算法（Stack.glm）和以随机森林为元分类器组合子模型的算法（Stack.rf）分别进行十折交叉验证。各分类模型十折交叉验证结果如表 4-4 所示：

表 4-4 各分类模型十折交叉验证结果

	model	Accuracy	Kappa
Boosting	C5.0	0.9891	0.9737
	GBM	0.9804	0.9769
Bagging	Bagged CART	0.9876	0.9702
	Random Forest	0.9892	0.974
Stacking	Stack.glm	0.9900	0.9759
	Stack.rf	0.9920	0.9806

从表 4-4 中可以看出，以随机森林为元分类器组合子模型的 Stack 算法（Stack.rf）分类效果最好，因此选取此算法作为第一阶段分类器，Stack.rf 内部

各子模型分别为 CART、KNN 及 SVM ，各子模型分类效果如表 4-5。

表 4-5 Stack.rf 子模型分类效果

	Accuracy			Kappa		
	min	max	mean	min	max	mean
CART	0.92	0.9681	0.9475	0.8095	0.9245	0.8737
KNN	0.745	0.828	0.7879	0.3401	0.561	0.437
SVM	0.976	1	0.9892	0.9424	1	0.974

4．第二阶段分类器——中和非中

训练第二阶段分类器时，本文基于中和非中数据集，同样使用十折交叉验证分别得出 Boosting、Bagging 以及 Stacking 这三种集成学习算法下各个模型的分类效果并进行比较，选出分类效果最好的模型作为第二阶段分类器，用来进行数据中和非中类别的判断。训练第二阶段分类器时所采用的各种模型与第一阶段相同，在此不再赘述。各分类模型十折交叉验证结果如表 4-6 所示：

表 4-6 各分类模型十折交叉验证结果

	Model	Accuracy	Kappa
Boosting	C5.0	0.9809	0.9613
	GBM	0.9775	0.9542
Bagging	Bagged CART	0.9788	0.9569
	Random Forest	0.9792	0.9578
Stacking	Stack.glm	0.9842	0.9680
	Stack.rf	0.9862	0.9720

从表 4-6 中可以看出，以随机森林为元分类器组合子模型的 Stack 算法（Stack.rf）分类效果最好，因此选取此算法作为第二阶段分类器，Stack.rf 内部各子模型与第一阶段相同，各子模型分类效果如表 4-7 所示：

表 4-7 Stack.rf 子模型分类效果

	Accuracy			Kappa		
	min	max	mean	min	max	mean
CART	0.92	0.972	0.9527	0.8377	0.9432	0.9039
KNN	0.536	0.652	0.5917	0.0619	0.2931	0.1746
SVM	0.968	0.996	0.984	0.9353	0.9919	0.9675

4.4.2 模型评价

使用本文所建立的二阶段分类器对测试集数据进行二阶段分类，具体分类过程为：首先使用第一阶段分类器对测试集数据进行分类，得出测试集数据“高”和“非高”类别的分类结果；其次，对“非高”类别的数据使用第二阶段分类器进行分类，得出数据“中”和“非中”类别的分类结果；最终，得出测试集数据的分类结果。分类结果示例如表 4-8 所示：

表 4-8 二阶段分类器分类结果

ID	第一阶段分类器结果	第二阶段分类器结果	最终分类结果
7297	高		高
3978	高		高
4173	高		高
4287	高		高
4418	高		高
2759	非高	非中	低
3713	非高	中	中
3777	非高	非中	低
3954	非高	非中	低
5375	非高	中	中
...	...	...	...

表 4-9 为测试集分类结果的混淆矩阵，从混淆矩阵中计算得出准确率为 98%，F1 值为 98.09%，Kappa 值为 96.93%，表明模型分类效果很好，可对回复意见的质量进行分类和有效评价。

表 4-9 混淆矩阵

		Predict class		
		高	中	低
Actual class	高	99	3	0
	中	1	123	0
	低	0	2	72

第五章 总结

随着互联网的快速发展，网络问政平台中的文本数据不断攀升，为了满足对大体量文本数据的挖掘与分析要求，本文主要基于卷积神经网络模型和集成学习方法解决了留言文本分类、热点问题挖掘与答复意见评价三个任务。

为了对任务一的留言文本进行分类，本文对原始数据集进行预处理后，构建了一个基于卷积神经网络的分类模型，并与其他模型进行对比，结果显示本文构建的分类器效果最优。

为了挖掘任务二要求的热点问题，本文首先利用留言文本中的关键词剔除噪声，再通过文本语义相似性聚类，将同一话题的留言归为一类。随后，在各大类内通过命名实体识别、模糊匹配及高频词的方式提取地名和人群，然后在类内精确匹配提取热点问题。最后，本文构建了合理的热点评价指标，用 TOPSIS 方法计算热度值，找出了热度值前 5 的热点问题，提取相关事件。

为了对任务三的答复意见进行评价，本文从相关性、可解释性、完整性和及时性四个方面构建了答复意见质量评价指标体系，以各指标值为特征，将数据集聚为高质量答复意见、中等质量答复意见、低质量答复意见三类，并以聚类结果为基础构建二阶段分类器，以便后期直接使用二阶段分类器对答复意见进行评价。

在对赛题进行充分研究后，本文对三个任务提出了合适的解决方案，基本实现了赛题设立的目标。

参考文献

[1] 樱桃小胖子同学. Word2vec(CBOW 和 Skip-Gram)原理理解及模型训练过程推理.https://blog.csdn.net/weixin_40771521/article/details/103893982.2020.4.30

[2] Magician~. FastText 原理总结. https://blog.csdn.net/qq_16633405/article/details/80578431.2020.4.30

[3]王芝辉,王晓东.基于神经网络的文本分类方法研究[J].计算机工程,2020,46(03):11-17.

[4]陈志,郭武.不平衡训练数据下的基于深度学习的文本分类[J].小型微型计算机系统,2020,41(01):1-5.

[5]李东辉. 基于微博结构的热点话题预测方法研究[D].中国地质大学(北京),2015.

[6] 不会停的蜗牛. 用 Doc2Vec 得到文档/段落/句子的向量表达.https://www.jianshu.com/p/854a59b93e09.2020.05.02

[7] lj_tang_tf. k-means 聚类算法过程与原理.https://blog.csdn.net/qq_39742013/article/details/81675050.2020.05.03

[8]张亚男,冯建文.基于混合聚类的微博热点话题发现方法[J].杭州电子科技大学学报(自然科学版),2018,38(01):59-64.

[9]郑飘飘,万健,司华友.基于评论的热点新闻事件识别方法研究[J].浙江科技学院学报,2019,31(05):392-399.

[10]李良. 突发事件微博舆情的话题发现和热度预测研究[D].西安理工大学,2018.

[11]申晓燕. 网络舆情热点分析技术的研究与实现[D].辽宁大学,2018.

[12] 达微. 熵值法确定权重算法及实现.https://www.jianshu.com/p/df1fa57e5532.2020.05.04

[13] 开心果汁. 【综合评价方法 变异系数权重法】指标权重确定方法之变异系数权重法.https://blog.csdn.net/u013421629/article/details/81171361.2020.05.04

[14]三月和九月.评价类模型——TOPSIS 法（优劣解距离法）学习笔记（一） .https://blog.csdn.net/qq_36384657/article/details/98188769.2020.05.04

[15]周炜翔,张仰森,张良.面向微博热点事件的话题检测及表述方法研究[J].计算机应用研究,2019,36(12):3565-3569.

[16]胡泽. 在线问诊服务回答质量评价方法研究[D].哈尔滨工业大学,2019.

[17]杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学,2018,45(09):237-242.

[18]易明,张婷婷.大众性问答社区答案质量排序方法研究[J].数据分析与知识

发现,2019,3(06):12-20.

[19]袁红,张莹.问答社区中询问回答的质量评价——基于百度知道与知乎的比较研究[J].数字图书馆论坛,2014(09):43-49.

[20] 村头陶员外. 机器学习 --> 集成学习 -->Bagging,Boosting,Stacking.https://blog.csdn.net/Mr_tyting/article/details/72957853.2020.05.03