基于 BERT 模型的“智慧政务”文本挖掘

摘要

本文就题目“‘智慧政务’中的文本挖掘应用”进行研究。为方便相关政府部门处理民生意见，提高政府的管理水平以及市政效率，本文基于文本挖掘领域最为前沿处理模型 BERT 模型，对于民生意见文本进行分类、热点问题聚类以及政府部门回复质量进行评估。

针对问题一：题目要求的文本分类任务，基本的操作包括文本的预处理、文本向量化以及构建分类器并且分类，这三个部分构成整个分类模型。对题目中提供的样本数据，首先我们进行数据预处理操作。其中包括文本清洗、去除停用词、类别匹配等操作。其次，经过预处理的文本数据需要进行向量化操作，才能进一步输入到分类模型中进行分类。

在本问题中，我们对文本的向量化以及分类器构建的方法，采取用的是 BERT 模型的算法思想，避免文本向量词义多义化的同时，提高了文本分类的精度。在分类的结果显示，在小样本中文本的召回率可以达到 86.5%，而在整体数据即大样本中可以达到 91.7%的召回率。相比较传统机器学习分类算法（以 VSM-SVM 模型为例）而言，有着更加优秀的表现。

针对问题二：题目要求进行热点问题的聚类以及热度评价指标构建。因此本文中采取的策略是先对文本地区分类，即对相同区域进行文本划分，接着进行同样的文本预处理，得到预处理的文本。最后采用 DBSCAN 算法进行文本聚类，本文实验结果表明聚类效果良好。

在热度评价指标构建中，本文结合了聚类文本的个数、聚类文本的总点赞数以及总反对数，进行合理的权重指数设计，构建得到了最终的热度评价指标公式。

针对问题三：题目要求进行政府部门相关回复的文本质量评估。本文通过对文本进行包括 TF-IDF、余弦相似度计算等多种处理，共计提取用于评估文本质量的十个特征，即包括回复文本五种词性个数、句子长度、词语数目、留言文本和回复文本的相似度，以及同一标签领域下回复文本的相关程度。通过主成分分析的方法得到三个主要成分，再根据不同特征对于主成分的载荷因子分析，可以很好的解释回复文本的完整性、相关性以及可解释性，从而得到评估政府部门回复文本质量的方案。

关键词：BERT 模型、DBSCAN 聚类、主成分分析、TF-IDF、VSM-SVM 模型、余弦相似度

一、问题重述

1.1 问题背景

近些年来，随着微信、微博、市长信箱、阳关热线等网络问政平台逐步成为政府汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

1.2 题目重述

问题一： 在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续群众留言分派至相应的职能处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

问题二：某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映 入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关 部门进行有针对性地处理，提升服务效率。请根据附件 3 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，定义合理的热度评价指标，并给出评价结果。

问题三：针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对 答复意见的质量给出一套评价方案，并尝试实现。

二、模型假设

假设一：意见文本均匀具体的地点标明，如果为出现市、县、区中其中一个地点单位，则归为最大地点单位，否则按照最小地点单位处理。

假设二：意见文本中，不考虑恶意刷留言条目的情况。即考虑每一条文本均是单独的用户发送的，这样有助于热度指标的计算。

三、问题一模型建立以及求解

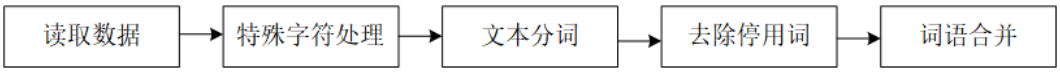
4.1 模型准备

4.1.1 文本预处理

1） **数据清洗**

在进行文本数据读取之后，发现每一个文本数据都存在着或多或少，与句子意思无关的字符，这些字符包括空白符、特殊符号、转义字符等符号，这对接下来的文本分析会产生较大的干扰。除此之外，文本中存在着大量毫无意义的停顿词，比如说“的、不仅仅、虽然”等等，通过停用词表过滤掉句子中的停用词，可以在一定程度上降低文本特征维度，提高文本的分类效果。

为了尽可能的让文本集中关注于语义本身以及本身的词汇特征，接下来就对文本进行相应的数据清洗操作，处理的流程为：



本题采用 Python 语言作为主要语言开发环境，采用 Pandas 包进行数据读取操作，引入哈工大停用词表，同时借助 jieba 包进行分词操作，数据清洗结果部分如下所示：

原始文本（中间省略部分内容）
1、 '\n\t\t\t\t\t\n\t\t\t\t\t\u3000\u3000 书记您好，我是来自西地省经济学院体育学院的一名即将大四的学生，系里要求我们在实习前分别去指定的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口，而且我学的高尔夫管理，却分配到其他系的实训点，老师说绝对不准提前回学校，……求问我们该怎么办？在这实训纯属浪费时间，感觉自己就是被系里连哄带骗过来了。之前说好的做青少年高尔夫体适能教练，过来完全不是一码事，天天陪 3，4 岁的小孩玩，做前台接待，当儿童顾问。很无奈\n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t'
2、 ……（省略后面文本数据）
处理后合并的文本
1、 书记您好西地省经济学院体育学院一名大四学生系里实习前指定公司实训公司内容老师介绍专业对口我学高尔夫管理分配系实训点老师说不准提前回学校做实训分毕业商量学校公司签了合同公司签合同大四第一学期结束家花几万块钱供读完高尔夫管理跑北京做儿童运动馆销售人员求问实训纯属浪费时间感觉系里连哄带骗说好做青少年高尔夫体适教练一码事天天陪岁小孩玩做前台接待儿童顾问无奈
2、 ……（省略后面处理后文本数据）

表 1：数据清洗文本列表

2） **类别匹配**

由于本文采取的是监督性学习算法，在文本输入模型之前，需要对文本进行标签匹配，匹配后每一段语料呈现形式均为“标签+制表符+预处理后文本”的形式。

4.1.2 BERT 模型引入

对文本数据的分析，本文初始实验数据条数共计 497 条，这个样本数据规模无疑是比较小的，并不适合一些需要在大规模数据集下才能表现优异神经网络模型，其中包括 TextCNN、RNN、LM 模型等，故本文将这些模型从实践方案中剔除。

其次，在文本训练的过程中，存在着一词多义的情况，即该词在不同上下文环境下的语义不同，这就要求考虑文本语义的过程中引入上下文环境，以此来进行更好的文本分类。

综合上述要求，为了在小样本数据集上取得较好的分类准确率，本文采用了 BERT 模型作为分类模型，BERT 模型基于 Transformer 的双向编码器表征，该模型有以下优点符合本次题目特点：

a) BERT 模型采用双向的 Transformer 编码层，该层为特征提取层，而 Transformer 的结构是由 Encoder 组成的，而通过 Encoder 可以充分的挖掘到每个词左右和上下文的信息，解决了语义环境上的问题。

b) 该模型可以通过采用通用语言模型，通过一定的微调（fine-tuning）操作，使得由外部庞大语料库生成预训练模型能够应用到任意一个场景之中，相比较于其他模型而言，采用迁移式学习的思想，在小样本分类的准确率表现更加优异。

BERT 模型整体的核心结构层次可以如下图所示：

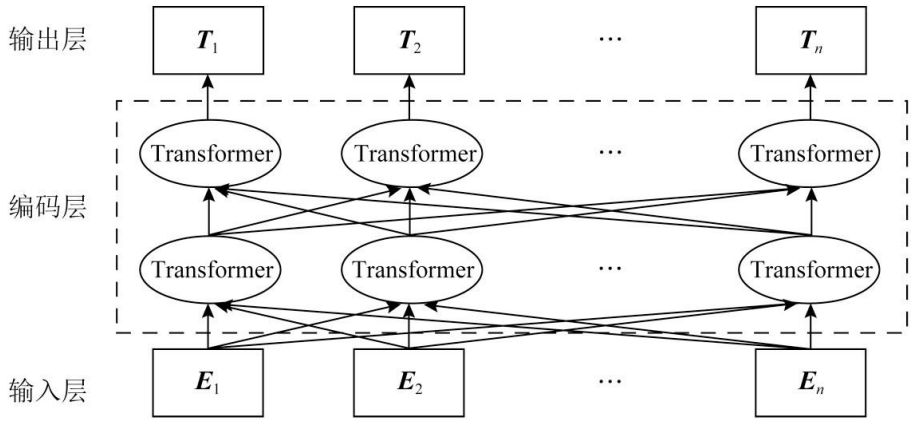


图 1：BERT 模型层次解构

4.2 模型构建

4.2.1 预训练模型

1）输入嵌入层

输入的内容表示可以在一个词的序列中表示单个文本句或者一对文本，即问题和答案，而在这里我们的输入的内容为标签匹配后的文本，即标签作为答案，而文本作为问题的形式进行输入。其输入表示可以通过以下三部分 Embedding 求和组成。如下图所示：

图 2：输入嵌入层解构

其中各层的含义如下：

a） Token Embeddings 表示层词向量，第一个单词为 CLS 标志，即为训练样本的分类标签，可以用于后续分类任务。

b） Segment Embeddings 用来区分两种句子，以保证接下来的分类任务

c） Position Embeddings 通过模型学习得到该层次嵌入值

2）Transformer 层

Transformer 层是一个基于 self-attention 的 Seq2seq 模型，而 Seq2seq 是由 Encode 以及 Decoder 两部分组成的。由于传统的 Encoder-Decoder 是通过 RNN 实现的，无法实现并行处理，故在 Transformer 层中采用 self-attention 作为 Encoder 的核心模块进行处理，解决了并行处理的问题，其中核心层 Attention 的公式如下所示：

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中Q,K,V分别为模型训练计算之后的权重矩阵，而d_k是每一个字 query 以及 key 向量维度，$soft\max$为归一化函数。

Transformer 的处理流程为：

a） 通过输入嵌入层产生的字嵌入表示，再加上每个字的位置信息，得到 Encoder 的输入序列。

b） 在 Encoder 之中，输入序列首先进入 self-attention 层次，计算每一个词与其他词之间的相互关系再不断调整权值得到新的词语表达。接着再通过对输入和输出进行相加，并且将相加之后的结果归一化。

c） 将归一化之后的输出向量传入到全连接的 Feed Forward 神经网络层，接着再次进过相加归一化处理之后，得到最终的词向量列表。

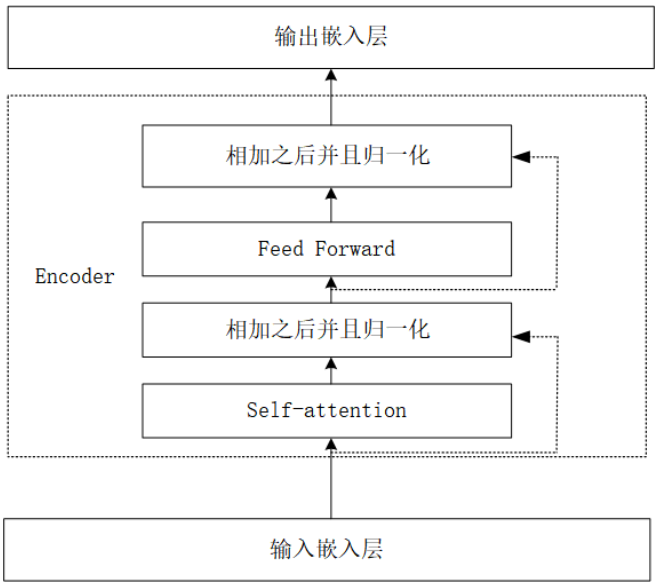


图 3：Transformer 层解构

3）输出向量层

BERT 模型的输出包括文本中每一个字符对应一个向量表示的输出模式、以及句子级别的输出向量模式，在本次分类任务中，我们采用后一种输出向量模式。即 BERT 模型最左端的[CLS]特殊符号的向量，即可以代表整个句子的语义的向量。输出层次如下所示：

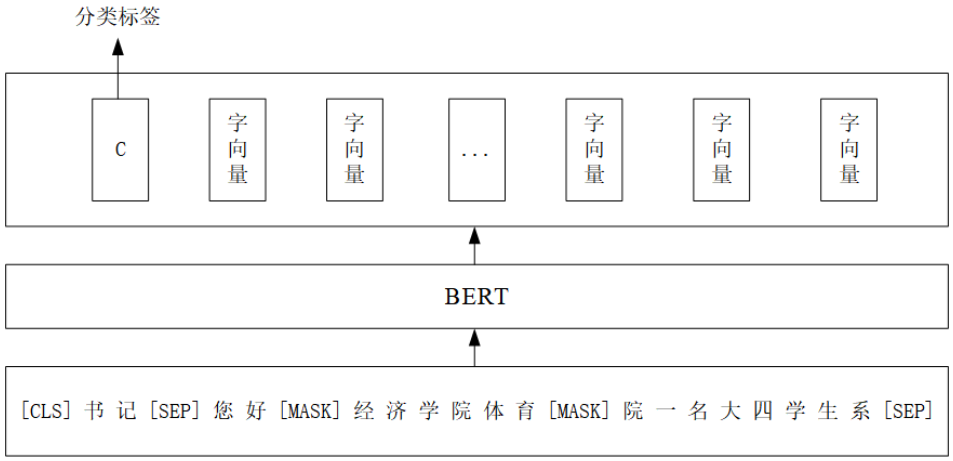


图 4：转换过程解构

4）实验参数

由于构建该层次的时候，需要花费大量的财力和人力，故本文在建立 BERT 层的时候，直接采用的是 Google 发布好的中文模型“BERT-Base，Chinese”. 该模型采用了 12 层 Transformer，即 $L=12$，隐藏层的尺寸为 768，即 $H=768$，而在 Multi-header 的参数调整为 12，共计 110M 的参数总量。通过该载入模型之后，可以直接获得输出的句向量或者字向量，以方便后续的神经网络层次输入。

4.2.2 微调过程

在语言模型 LM 采用了上述的 Transformer Decoder 的方法进行训练，采用文本预测作为语言模型的训练任务，训练完毕之后得到语言模型 LM。以此为基础之上增加少量神经网络层，如加入一层 Linear Porject 来完成一些特定的任务，即本文进行实践的分类任务。

在微调过程中，由于该过程是监督性学习过程，所以我们采用前面进行预处理之后类型匹配文本进行有监督性的训练模型。微调的参数如下所示

参数	值
Batch	32
学习率 Adam	2e-5
epoch	6
句子序列长度	150

表 2：微调参数表格

4.2.3 实验过程

1）评价指标

F1 分数(F1-score)是分类问题的一个衡量标准，它是精准率(precision)和召回率(recall)的调和平均数，能够更加全面的体现模型分类的性能，其最大值为 1，最小值为 0，F1 越接近 1，则说明在该分类下的分类结果越完美，其计算公式如下所示：

$$F_1 = \frac{2(precision \times recall)}{precision + recall}$$

其中精准率以及以及召回率的计算公式如下所示：

$$precision_k = \frac{TP}{TP + FP}$$

$$recall_k = \frac{TP}{TP + FN}$$

TP 代表预测分类正确的个数，FP 代表错误将其他类预测为本类的个数，而 FN 为本类标签预测为其他类标签的个数，同时最终结果得到模型整体分数为：

$$score = \left(\frac{1}{n}\sum f1_k\right)^2$$

2）实验过程

本次实验过程采取了**对照式实验**以及**控制样本大小实验**。其中对照式实验采用传统 VSM-SVM 文本分类模型和 BERT 模型之间的分类效果，来体现模型之间的优异程度。VSM-SVM 模型即是向量空间空间模型下采用 SVM 分类器的模型，以

TF-IDF 作为文本特征向量向量并且输入到 SVM 分类器中，在传统的机器学习分类中有着较好的表现效果。

本文的控制样本实验分为两种情况，第一次是基于初始 497 条数据的模型训练，第二次是基于全部数据 9208 条数据的模型训练，以此检验模型在小样本和大样本上表现效率。同时在模型的训练之前，把预处理之后的全部数据文本分为 8:2 的训练集、测试集样本，分别计算模型出模型的召回率、精确率、F1-score，并且部分结果通过混淆矩阵的形式展现。

3）实验结果

在小样本中，测试集共计包含 39 个文本；在大样本中，测试集共计包含 737 个文本，通过训练之后得到测试结果的混淆矩阵以及模型评价指标

图 5：大小样本下的混淆矩阵

在混淆矩阵中，主对角线颜色较深，其他部分颜色几乎为浅色，故我们可以

很直观的看到整体模型的召回率，准确率效果是十分好的。但由于数据类别的不平衡性，我们需要继续计算模型在不同类别下的 F1-score，以下为不同类别下的 F-score 图表：

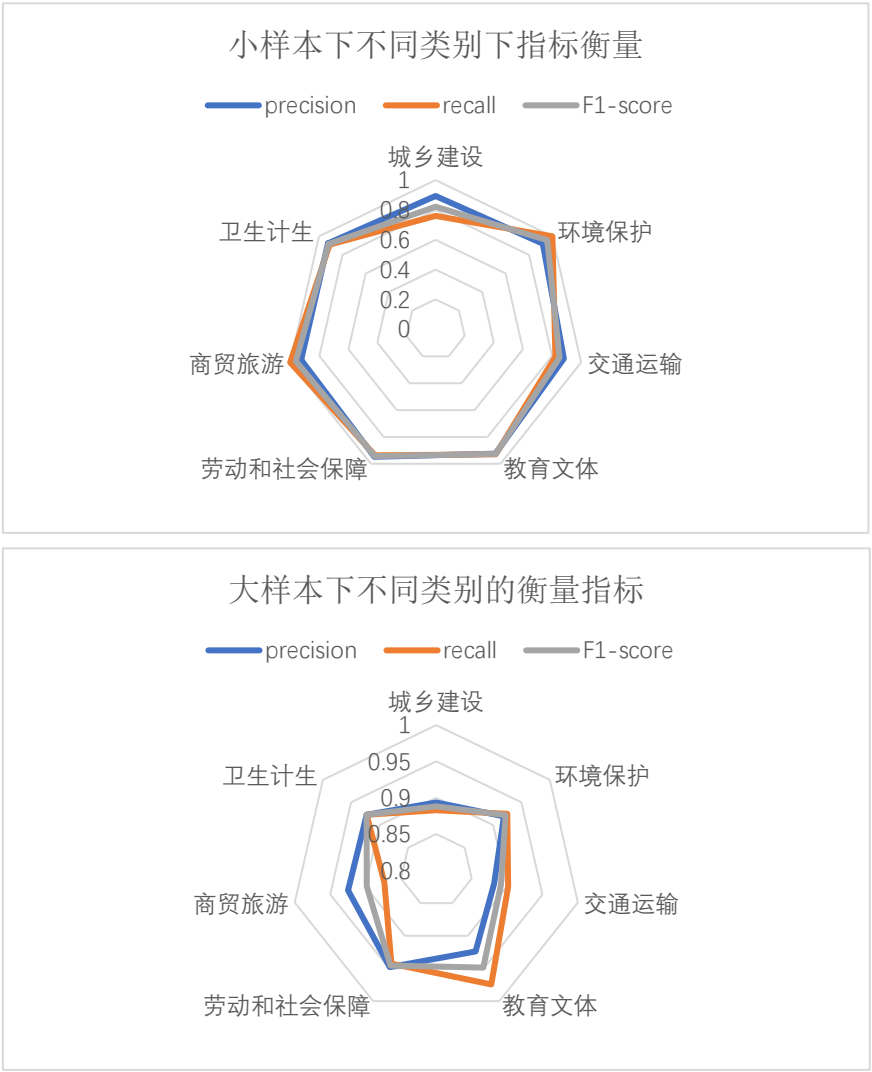


图 6：大小样本下不同类别各指标衡量

在同一样本的不同类别下的不同衡量指标中，其表现的情况各位不同，其数值的相对程度能够体现的是类别不平衡样本下，样本越小分类各项指标越偏低。由于测试集的存在着一定类别上的偏差，故不能够将一次测试集的标准平均作为最终结果。

而最终模型的结果，需要通过 K 折交叉验证之下，打乱整个样本并且进行随机选取，结果才能够避免偶然性误差，并且在测试集上取得比较稳健的表现，在为小样本和大样本平均召回率、平均准确率以及 F1 分数的对比：

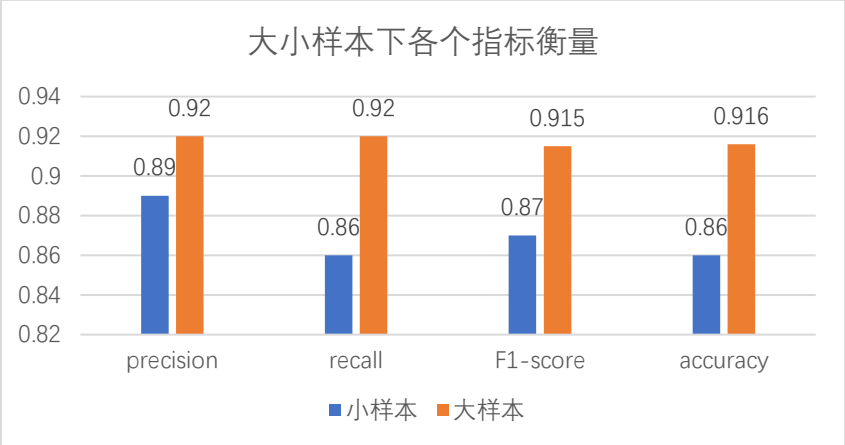


图 7：大小样本下的各指标衡量

在比较完 BERT 模型在不同大小样本的效果之后，接下来通过控制样本不变，选择不同的分类模型进行对照实验，得到的结果如下所示：

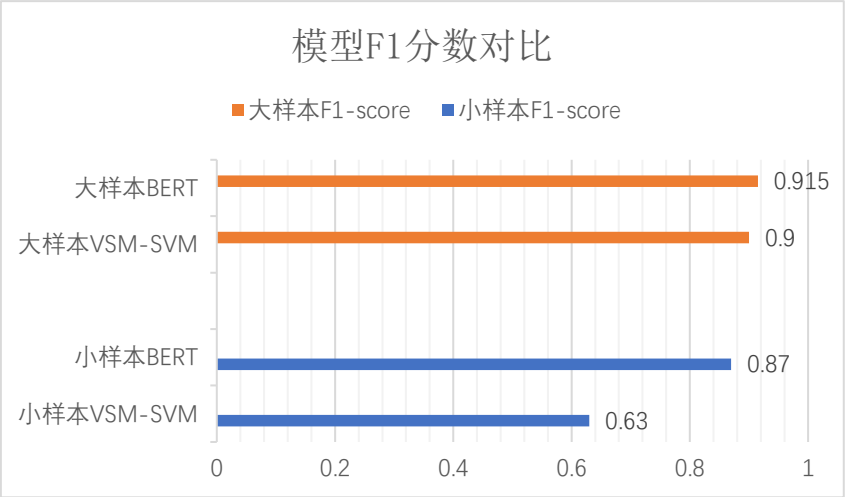


图 8：VSM-SVM 和 BERT 之间表现差异

可以很清楚的看到，在小样本数据集下，BERT 模型的表现程度远远优于传统的机器学习模型，而在较为大样本的数据集下，BERT 模型的表现程度也比一般的传统机器学习模型的表现程度好。预测的标签部分如下所示：

留言文本	预测标签
书记您好，我是来自西地省经济学院体育学院的一名即将大四的学生，系里要求我们在实习前分别去指定的不同公司实……	教育文体
K3 县的乡村卫生室现在大多处于无证行医的状态。造成这种情况原因的是。卫生室的医疗机构许可证……	医疗计生
尊敬的政府：A5 区劳动东路魅力之城小区临街门面长期油烟直排。长期投诉无果。由于各部门无权力强关……	环境保护
(省略后面文本)	(省略后面标签)

四、问题二模型建立以及求解

5.1 模型准备

5.1.1 定性分析

1）数据选取

根据题目分析可得，需要对热点问题进行挖掘。在题目的要求中，需要对特定地点特点人群进行判断，故从常规思维去思考，首先我们需要对文本进行地点的分类，得到每个地区的不同问题的集合。

在题目给附件 3 数据中，通过定性分析，存在着两部分可以利用的文本内容，一部分是留言详情，另一部分是留言主题。留言详情包含的关键词虽然多，进行相似度判断的时候，一是运算量过大，而是无法给出一个合理的阈值进行聚类。故本题选取**留言主题**进行聚类，留言主题的特点十分明显，即包含了一件事构成的几个要素：**特定地点 + 特定人群 + 特定事件 + 时间范围**，言简意赅。

2）处理思路

通过文本数据分析得，特点热点问题的挖掘指标有两个，一个是通过点赞多少以及反对数进行指标衡量，另外一个是通过特定地点相似问题总个数作为衡量指标。为了方便题目说明及处理，做出以下定义

定义	说明	维度
民意指数	点赞数 - 反对数	1

因此，本题可以根据这两个指标，分别从**两个思路**出发。一方面是从点赞高低出发，选择前十个点赞最高的个案，进行相似度计算，从而得到 10 个相似集合；另一方面通过地点分类之后，通过聚类算法，进行相似文本的聚类，聚类簇包含的文本越多，则说明热度越高。综合这两个思路，给出具体的热度评价指标公式。

5.2 模型构建

5.2.1 民意指数聚类

1）民意指数计算

首先本文进行民意指数计算，通过计算点赞数减去反对数，再进行倒序排序。由于我们需要前十个不同主题的高赞文本数据，但是再经过排序之后会存在这主题相同且点赞数较高的情况。故不能够一次性将前十个文本进行聚类计算，而需要先进行某一个文本逐一进行相似度计算之后，从初始文本中剔除该相似文本簇。在进行下一个文本计算。这样就比较好解决了问题。

2）TF-IDF 值计算

TF-IDF（词频-逆向文档频率）是一种用于信息检索与文本挖掘常用的加权技术。用于评估一个字词对于一个文件集或者一个语料库中的一份文件的重要程度。字词的重要性随着文件出现的次数正比增长，但是有会随着语料库出现的频率反比下降。其核心的计算公式为：

$$TF-IDF=TF*IDF$$

其中，TF 为词频，表示关键字在文本中出现的频率。而 IDF 为逆向文本频率，表示为总文件数目除以包含该词语文件数目，再取商的对数得到，进一步计算公式如下：

$$TF_{i,f}=\frac{n_{i,f}}{\sum_k n_{k,j}}$$

$$IDF_i=\log\frac{|D|}{1+|\{j:t_i\in d_j\}|}$$

其中$n_{i,j}$代表的是该词在文件中出现的次数，d_j为所有词汇出现的次数总和。$|D|$为语料库中文件的总数，而$|j:t_i\in d_j|$表示包含词语t_i的文件数目

3）余弦相似度计算

从二维空间向量变换中，可以通过计算向量间夹角的余弦值进行相似度计算。延伸到多维空间向量中，则其计算公式如下所示：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

4）聚类结果

取第一个最高赞的文本，经过把文本转化为 TF-IDF 特征向量之后，在进行余弦相似度计算，得到一个相似文本簇，再将该文本簇从文本中取出，继续算第二个相似文本簇，以此类推，计算到第 10 个文本簇位置。取相似程度为 0.4 以上的阈值、去相似文本簇文本数目前三项，得到以下高赞且相似文本数目较多的前 3 项。

代表文本	文本数簇大小	总民意指数
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	7	2109
承办 A 市 58 车贷案警官应跟进关注留言	5	2344

A市至赣州高铁对绿地海外滩二期小区影响太大了	4	685

5.2.2 相同地区文本聚类

1) 地区分类

在文本分类中，首先需要先将相同地区的文本分在一类，再进行相同地区的文本聚类，效果会较为出色。而本题中采用地区分类的方法为通过正则表达式的方法进行地区聚类，正则表达式如下所示：

$$re.compile(r'([A-Z][市])|([A-Z][0-9][区])|([A-Z][0-9][县])')$$

通过以最小地点单位为基础去进行聚类，不出现具体市区县则归为最大地点单位，最终分为13个不同的地区。

2）DBSCAN聚类算法

该聚类方法基于密度的聚类方法，无需要事先人为确定好聚类的个数。该算法的中心思想就是对于集群中的每一个点，在给定的半径范围内，其相邻的数量必须超过预先设地的某一个阈值。该算法聚类如下图所示:

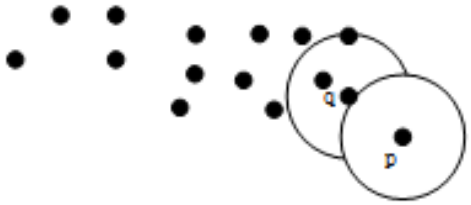


图9 DBSCAN聚类

该算法必须事先设定好两个参数，即Eps、$Minpts$两个参数值，这两个参数的意义以及计算公式如下：

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leqslant Eps\}$$
$$|N_{Eps}(q)| \geqslant MinPts$$

其中，D表示整个数据集集合，$dist(p,q)$表示点p和q的距离。而$MinPts$表示一个中心点Eps所要包含的最小数量。

3）K-dist参数估计

由于在使用DBSCAN聚类的时候，需要不断的去调动参数以得到最优的聚类情况。而一般采用K-dist算法进行最优阈值的范围确定，即对于数据集中的每一个点，分别计算他们与第k个最近邻的距离，然后将这些距离进行逆排序，并且绘制他们距离分布曲线，并且在第一个谷点左右取得最优阈值，以A1地区为例，得到的K-dist曲线如下所示：

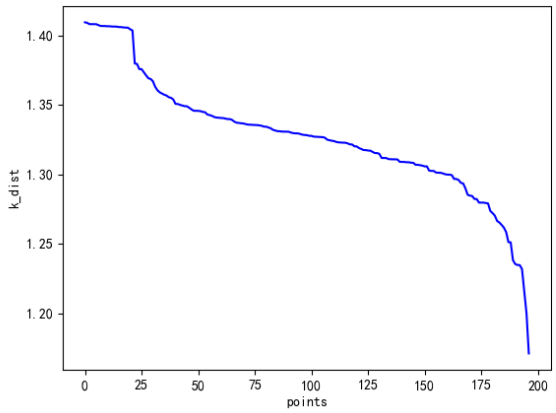


图 10：A1 文本地区 K-dist 曲线

根据 k-dist 图可以发现，在 k_dist 为 1.25-1.30 之间的时候，出现了较陡的曲线，而在这之前，随着 k 值增大，半径内点的变动不为敏感。故此时在缓坡以及陡坡处，在此区间选取最优阈值较为合适。

4) 聚类结果

在地区分类之后，便在同一个地区内进行聚类，最后将每个地区聚类结果总汇，选取聚类条目以及民意指数较高的前五项，作为聚类的结果。部分地区聚类如下所示：

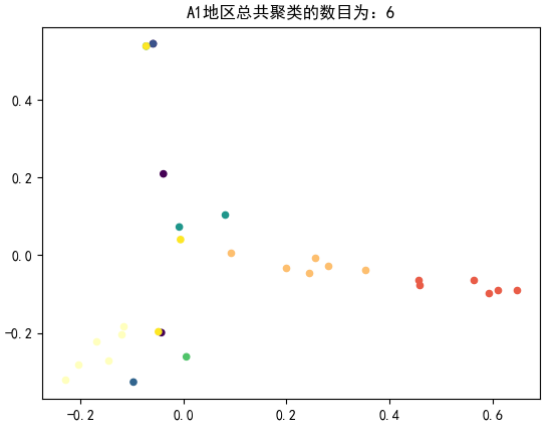


图 11：A1 地区文本聚类情况

最终统计各个地区结果，得到聚类条目最多前五项如下所示：

地区	文本	文本簇大小	民意指数
A2 区	投诉 A2 区丽发新城附近建搅拌站噪音扰民	28	37
A7 区	咨询 A7 县星沙旧城改造项目问题	9	28
A 市伊甸园	关于伊景园滨河苑捆绑销售车位的维权投诉	22	3

A3 区青青家园	A3 区青青家园小区乐果果零食炒货公共通道摆放空调扰民	11	-1
A 市	A 市经开区泉星公园项目规划需优化	6	33

5.3 热度指标以及结果

由于在聚类计算中，存在着民意指数高，但是文本簇的大小较小；民意指数低，但是文本簇较大，为了合理的得到前五个热度问题，通过分析数据的特点，本文就民意指数以及文本簇大小给出具体定义公式

$$\sigma = 0.95 \times \eta + 0.05 \times \mu + \frac{\mu}{|\mu|}\log(|\eta \times \mu|)$$

其中σ代表热度指数，η代表文本簇大小，μ代表民意指数，log 函数为热度平滑指数。最终得到前五问题热榜为：

排行	文本	热度指数
1	承办 A 市 58 车贷案警官应跟进关注留言	134.95
2	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	125.1
3	A 市至赣州高铁对绿地海外滩二期小区影响太大了	49.02
4	投诉 A2 区丽发新城附近建搅拌站噪音扰民	38.45
5	关于伊景园滨河苑捆绑销售车位的维权投诉	27.05

五、问题三模型建立以及求解

6.1 模型准备

6.1.1 影响文本质量特征分析：

评价一个文本的质量可以从多个文本特征去评价。而影响文本文本质量的因素可以从三个方面去入手。

从文本自身的结构上看，包括句子的长度、句子的字数、词语的个数、句子中各种词词性占有词个数。从回复文本之间的相关性上看，留言文本以及回复文本之间的相关性作为指标。从同一标签下领域文本内容，计算文本之间信息量作为指标。

选取这些指标从某一特定方面都可以一一解释对文本质量的重要性，比如说信息量越高的文本，呈现出来的句子有意义的词数量越多等等。因此，从一定程

度上讲，可以分别从可解释性、完整性以及相关性通过这些指标去解释评价。

共计有 10 个特征指标，由于特征指标过多，需要对特征进行降维操作，分析得到对文本质量贡献较大主要成分。指标内容如下所示：

编号	属性	说明	维度
1	词数 (Num_wd)	评论文本进行分词之后的所有词个数	1
2	句数 (Num_sen)	评论文本的句子数量	1
3	平均句长 (Num_wd_sen)	词数/句数	1
4	词性频数 (Pos_tag)	文本中出现各种词性的频数	5
5	留言答复相关度 (msg_reply_sim)	留言内容和回复内容相关程度	1
6	同一标签下不同文本信息量 (same_label_sim)	如环保领域下，不同文本信息量	1

6.1.2 主成分分析基本思想：

主成分分析的思想是利用降维的思想，进行多个变量转化为少数几个综合变量的方法，且得到的主成分能够反映原始变量的绝大部分信息。因此，通过主成分分析，即 PCA 模型可以对特征进行降维，得到不同主成分来评估文本的质量。

6.2 模型求解

6.2.1 PCA 模型建立与求解

由于在数据处理过程中，将数据的条目以及数据的特征进行组合生成数据样本矩阵x，其中r为样本的数目，而n为数据的特征条目

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{r1} & x_{r2} & \cdots & x_{rn} \end{bmatrix}$$

而样本矩阵通过计算均值$\bar{x}$以及标准差S_j可得到标准化后的数据：

$$X_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j}$$

从而得到标准化矩阵X，再通过计算样本与样本之间的协方差r_{ij}，构成的协方差

系数矩阵即为样本的相关系数矩阵，如下所示

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

通过热力图可视化样本系数矩阵，我们可以直观的看到各个样本之间关系以及相关系数具体值：

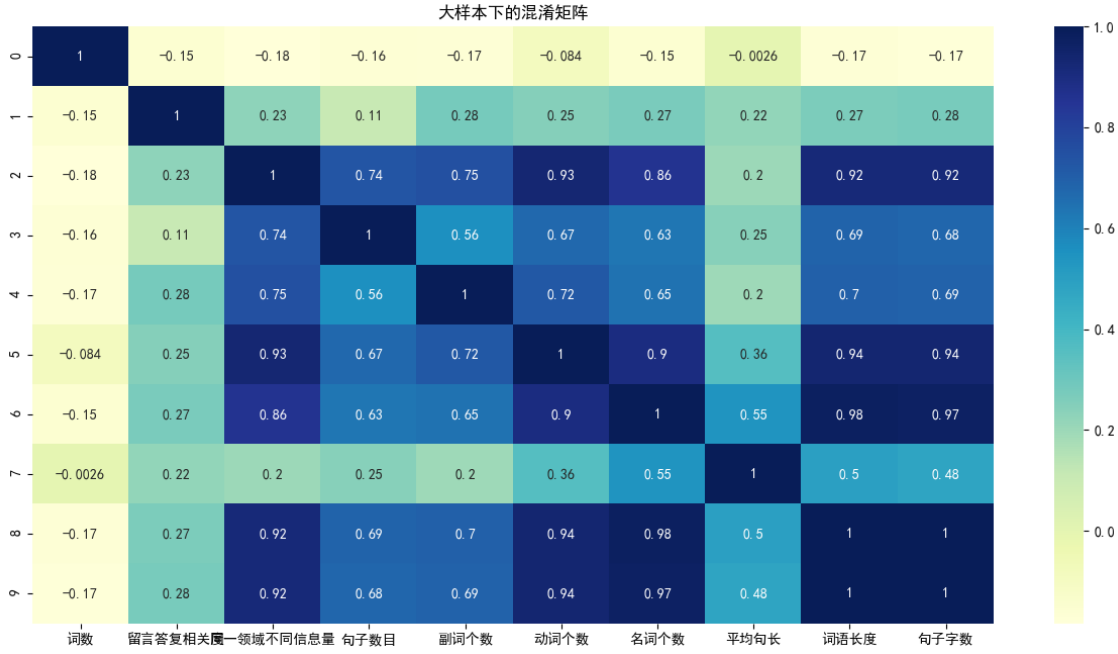


图 12：相关系数矩阵

根据协方差矩阵 R 求出特征值λ、主成分贡献率以及累计方差贡献率。特征值可以通过求解以下特征方程得到：

$$|\lambda E - R| = 0$$

特征值是各个主成分的方差，它的大小反映了各个主成分的影响力。通过特征值可以计算得到累积主成分贡献率：

$$\Phi = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k} (i = 1, 2, \cdots, p)$$

本文借助 SPSS 软件，通过数据计算可以得到各个主成分的特征向量、特征值、贡献率以及累计贡献率通过 SPSS 软件呈现表图如下所示：

总方差解释

成分	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	6.168	61.680	61.680	6.168	61.680	61.680	5.634	56.337	56.337
2	1.060	10.604	72.284	1.060	10.604	72.284	1.518	15.182	71.519
3	1.048	10.479	82.763	1.048	10.479	82.763	1.124	11.244	82.763
4	.783	7.830	90.593						
5	.430	4.297	94.890						
6	.370	3.695	98.585						
7	.077	.774	99.359						
8	.042	.419	99.778						
9	.020	.199	99.977						
10	.002	.023	100.000						

提取方法：主成分分析法。

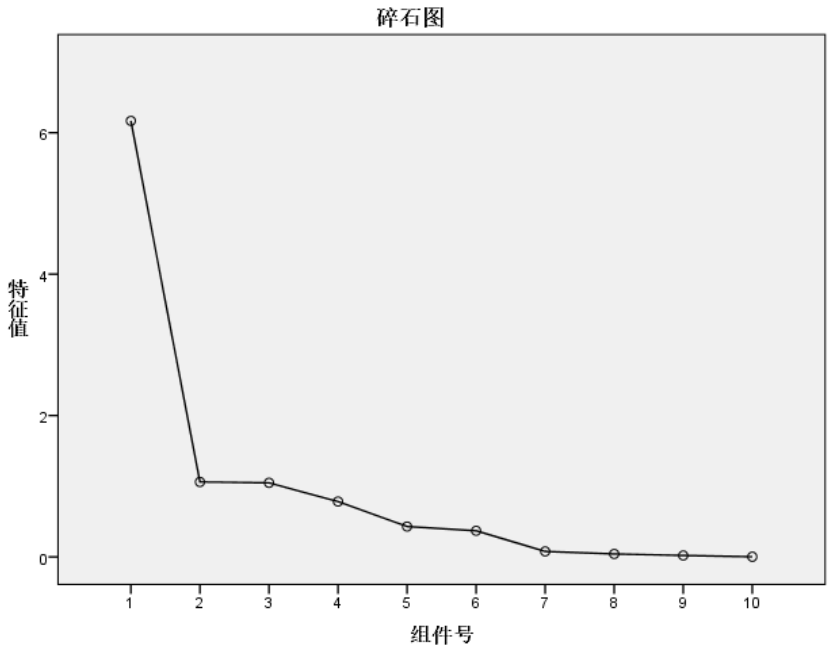


图 13：总方差解释图表和碎石图表现

可以从图表中看到，前三个因子共计占比为 82.7623%的贡献率，而从碎石图中可以清晰的看见，不同组件之间对于的特征值，所以把这十个指标分成三个主成分因子是可以解释一些共性的因素。

通过计算各个特征所对应的主成分得分如下所示：

	主成分 1	主成分 2	主成分 3
同一标签下文本不同信息量	0.027	0.104	0.799
留言答复相关度	-0.173	0.547	-0.456
句子长度	0.214	-0.180	-0.038

形容词个数	0.197	-0.210	-0.003
副词个数	0.156	0.121	-0.134
动词个数	0.182	-0.30	0.078
名词个数	0.125	0.146	0.071
平均句长	-0.107	0.669	0.230
词语个数	0.150	0.078	0.050
句子字数	0.150	0.070	0.036

表 9：各个文本特征所占的主成分

6.2.2 结果分析以及评估方案设计

综合上述得到的结果，可以分为三个主要成分，并且这三个主要成分的解释如下所示：

1) 从第一个主成分出发，句子长度、形容词、动词等句子词性以及词语长度、句子字数等结构性成分具有较大的正载荷，故该主要成分可以解释为回复文本的完整性。
2) 从第二个主成分出发，平均句长，留言和答复句子之间的具有很大的正载荷，而对于平均句长越大，关键词出现的概率就越大，所以可以将该主要成分解释为回复文本的相关性。
3) 从第三个主成分出发，同一标签下不同文本下的信息量不同具有最大的正载荷，而该特征反映的是相同标签领域内，文本是否回答内容与该领域的内容相关，如果越相关，则说明回答的问题越倾向于该领域的内容，就说明其可解释性就越高。

所以得到具体文本质量的评估方案：

1) **解释性：**如果**回复文本信息**与**同一标签下**的其他回复文本正比越高，就说明该文本越能够解释该领域的知识块，文本的可解释性就越高，
2) **相关性：**如果**回复文本信息**与**相应的留言信息**正比越高，并且**平均句长**越高，即句子除去无意义的词语进行分词，有意义词占比越高，说明其与留言文本的相关程度就越高。
3) **完整性：如果回复文本信息**与文本本身结构特征，**如句子的长度、词语个数、名词个数、动词个数等有意义词性个数**的正比越高，则越能说明回复文本的完整性越好。

七、 模型的评价与推广

7.1 模型的优点

该模型较好的解决了题目所提出来的要求，包括在小样本数据以及大样本数据上具有较好的表现，可以推广到实际应用场景，并且通过定量计算以及定性分析得到了不错的结果。其中包括文本分类的准确率，热点文本聚类以及合理的热度系数指标构建、评估回复文本的质量评估方案。

7.2 模型的缺点

在第二问给出的聚类中，通过正则表达式进行定性的判断是不是属于同一个地区，对于扩展到多个市甚至是整个全国的分类而言，那些地址较为混乱的情况表现较差。

在第三问回复质量的评估之中，本文仅仅只是通过主成分分析进行定性定量判断，而不能够算出各个指标对应的权重。我们通过查阅文献以及数据计算说明，如果能够设置回复的点赞数以及反对数，定量给出回复文本质量，则可以通过多元回归分析的方法，进行特征筛选以及权重赋值。

7.3 模型的推广

根据数据计算，本次实验的数据文本大部分为中等偏长的文本序列，由于网络上的短文本数量众多并且具有同样需求，可以基于本文模型进行短文本特征扩展（比如说主题特征、词对特征等），可以更好的推广到短文本领域。

除此之外，该模型通过适当调整之后，同样适用于文本的二分类以及情感极性的判断。

八、参考文献

[1]段丹丹,唐加山,温勇,袁克海.基于 BERT 的中文短文本分类算法的研究[J/OL].计算机工程 1-12[2020-05-07]

[2]刘思琴,冯胥睿瑞.基于 BERT 的文本情感分析[J].信息安全研究,2020,6(03):220-227.

[3]连冬阳. 基于深度学习的新闻评论热度预测研究[D].哈尔滨工业大学,2018.

[4]陈杨楠. 基于改进 SVM 算法的投诉文本分类研究[D].合肥工业大学,2019.

[5]陆玉恒. 基于聚类算法的热点提取技术应用研究[D].东华大学,2016.

[6]张扬武,李国和,王立梅,宗恒,赵晶明.一种基于 PCA 的文本特征混合选择方法[J].计算机应用与软件,2019,36(10):23-29+80.

[7]周董.基于 k-dist 图的变密度 DBSCAN 算法改进研究[D].上海:上海财经大学,2009.