“智慧政务”的文本挖掘

摘要

随着互联网的迅速发展，网络问政平台成为政府工作的重要渠道。其中，对平台留言信息的划分和热点的整理等成为提升政府工作效率的重要手段。

针对问题一，首先对附件 2 的数据进行分析与预处理，对留言详情做 jieba 分词处理，去掉停用词及部分不相关的词语。取出文本数据中留言详情列和一级标签列，划分相应比例的训练集与测试集，将训练集和测试集对应的留言详情转换为 Tf-Idf 权值向量。利用朴素贝叶斯模型、CNN 模型及随机森林模型分别对文本做一级标签分类，对比得出随机森林模型的分类结果更加准确，该模型的 Micro-F1 和 Macro-F1 分别为 0.913842、0.902635。同时根据每个一级分类标签对应的 roc 曲线，验证了随机森林模型分类结果的准确性。

针对问题二，首先对附件 3 的数据进行预处理，用 Python（jieba 分词）对留言主题以及留言详情进行分词，将分词结果进行筛选，提取关键词，并利用 Tf-Idf 将分词文本数据向量化，得到稀疏矩阵。用 K-Means 算法对向量进行聚类分析，得到分类结果。选取留言时间长度，最新留言时间，最早留言时间，单位时间出现的问题频数，留言问题总频次，点赞总数以及反对总数作为留言热度的评价指标，利用 RSR 综合评价法建立热度评价模型:

$$y=0.1537585145951022*Probit-0.22508890877108195$$

对留言热度进行评价，得出热度排名前 5 的热点问题。最后按照表 1 和表 2 的格式导出“热点问题表”和“热点问题留言明细表”。

针对问题三，首先从相关性、完整性、可解释性三个角度出发，对留言详情以及答复意见的数据进行分析，选取相关性、规范性、可读性三个指标并对其进行量化处理。其次，为了降低评价的主观性，将层次分析法和信息权数法结合确定的指标权重与因子分析法确定的权重进行对比，得出前者效果更好。最后构建出质量评价模型：

$$Q=0.3712a_1+0.2543a_2+0.3725a_3$$

得到答复意见的质量得分以及整体的质量分布图。将量化后的指标数据分为训练集和测试集，用SPSS软件对测试集进行线性回归，验证该指标权重的性能。从训练集和测试集的质量分布图看出，二者都是呈偏正态分布，留言的质量得分集中分布在[0.3,0.6]。详细评价结果数据见附件6。

关键词：随机森林模型 ROC 曲线 Tf-idf K-Means 聚类 综合评价法 层次分析法

目录

图录

表录

第一章　问题描述

1.1　问题描述

随着互联网的迅速发展，微信、微博、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智。凝聚民气的重要渠道，各类社情民意的文本数据量不断增加，这给主要靠人工来进行留言划分和人工整理的相关政府部门带来了极大的挑战。

近年来，随着大数据、云计算、人工智能等技术的发展，逐步建立基于自然语言处理技术的智慧政务系统，可通过计算机更快的帮助政府部门实现留言分类与热点整理工作。这也成为有关部门提升政府管理水平和施政效率的有效方法。

针对问题一，根据附件 1 提供的一级分类标签体系对留言进行分类，利用附件 2 提供的数据进行数据分析与处理，建立关于留言的一级标签分类模型。使用 F1-Score 对分类模型进行评价。

针对问题二，根据附件 3 中提供的留言信息提取某一时段内群众集中反映热点问题，有助于相关部门进行有针对性地处理，提升服务效率。利用 K-Means 聚类方法将附件 3 中的留言信息进行聚类，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，利用综合评价法对留言问题进行热度评价，并给出评价结果。导出排名前 5 的热点问题和相应热点问题的留言信息，分别保存为“热点问题表”和“热点问题留言明细表”

针对问题三，根据附件 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。通过对附件 4 中相关数据进行量化处理，选取评价指标，利用广义线性模型建立评价体系。

1.2 论文结构安排

本文共分为八章，各章内容安排如下：

第一章，　对论文需解决的问题进行描述，并简单介绍正片论文的结构安排。

第二章，　对附件 2 数据进行探索与分析。

第三章，　建立文本分类模型。

第四章， 对数据进行预处理，将留言进行分词处理。利用 K-Means 聚类方法将留言归类，并进行可视化处理。

第五章， 选取热度评价指标，对不同单位的数据进行归一化，利用综合评价法对留言热度进行评价。

第六章， 选取质量评价指标，并对附件中的文本数据进行预处理，并将选取的指标进行量化处理，展示部分量化处理结果。

第七章， 利用层次分析法以及信息权数法相结合确定各个指标的权重，并与因子分析确定的权重作比较。利用广义线性模型的方法，通过得到的指标权重值构建质量评价模型。

第八章， 根据得到的模型，对答复意见的质量进行描述，并用测试集验证模型的优劣。

第二章　附件二文本数据的探索分析

2.1　附件 2 数据预处理

在数据挖掘过程中，数据预处理是第一步，同时也是很重要的一步，数据预处理的好坏直接决定着文本特征提取、分类预测等步骤能否顺利进行。在对文本数据的分析过程中，对题目所给的附件数据的认识逐步加深，并最终得出了一套较为完整的数据预处理流程。

2.1.1　一级分类标签的数值转换

由于附件 2 中的一级分类标签的种类较多，我们将其进行数值转换，转换结果如表 1 所示：

表 1 数值转换

一级标签	数值转换
城乡建设	0
环境保护	1
交通运输	2
教育文体	3
劳动和社会保障	4
商贸旅游	5
卫生计生	6

2.1.2　重复数据的处理

通过对文本数据的仔细观察，我们发现某几类留言存在留言用户以及留言内容的重复，如图 1 所示。为了使预测结果更加准确，我们利用程序代码对重复留言进行删除处理，余下数据均为不重复数据。

U0007137	南路A2区华庭楼顶水箱	19/12/6 14:40:	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道，水是我们日常生活必不可少的用品，霉是一种强致癌物，我们住在这里连基本的健康保障都没有，请政府街道各领导重视起来，也请环保部门来检测，还我们一个健康安全的基本生活环境！	0
U0007137	区华庭自来水好大一	19/12/5 11:17:	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道，水是我们日常生活必不可少的用品，霉是一种强致癌物，我们住在这里连基本的健康保障都没有，请政府街道各领导重视起来，也请环保部门来检测，还我们一个健康安全的基本生活环境！	0

图 1 附件 2 中的重复数据

2.1.3 去除文本数据中的*序列

读取文件时，我们发现留言详情中存在地址、电话号码、身份证号码用*代替的现象，如图 2、图 3 所示，为使后续文本分析结果更加准确，我们将其进行删除处理。

得出的中标候选人是西地省梅溪湖建设有限公司第一名，西地省绿林市政景观工程有限公司第二名，西地省望新建设集团股份有限公司第三名。这样的结果更加令人瞠目结舌，第二次中标候选人这3家公司在第一次投标时就已经因为技术标和项目管理机构评审不合格被否决投标，那为什么在复核后又能以前3名的成绩成为候选人？难道这不是一个天大的笑话？后经了解，原来是L8县县住房和0局副局长吴先吉为了帮助意向投标人姚明明（姚明明社会人员，此次投标姚明明拿多家公司串标围标），伙同代理机构西地省通国工程管理有限公司而从中作梗，并扬言摆平此事是小事一桩，一定要帮姚明明得到。近年来，从中央到地方，都在为营造法治营商环境而不断努力推进，L市也出台了《优化营商环境20条措施》，为什么L8县住房和0局副局长吴先吉竟如此猖狂敢顶风作案？希望相关部门能够查明真相，将破坏正常投标程序的人员绳之于法，还一个清清白白的营商环境。反映人：李鸣联系方式：**********

图 2 电话号码*序列

关于请求解决危房重建的报告尊敬的高厅长：　　我叫高明星，男，汉族，1928年1月5日出生，西地省K8县人，身份证号码**********，县邮电局退休职工，家住K8县内环路82号。我于1980年在自家的自留地上建了一座110平方米的平房（当时交纳宅基地款146.16元，并开了收据），因当时没有考虑建楼梯间，生活起居十分不便。1990年我向西郊村委会和舜陵镇政府写了报告，请求批准在属于自己屋檐后的余坪上增建楼梯间。报告递交后，经当时驻西郊村镇干部唐梅雪亲临实地考察和村委L6县意（在镇政府下批文前镇长何生运也带人到现场核实过），报请了舜陵镇政府批准，舜陵镇土地管理服务站到实地进行了丈量，余坪建楼梯面积为14.2平方米，按照当时余坪建房每平方米2.8元的收费标准收取了土地使用费46.86元，并开了票据。　　因房屋建于1980年，年代已久，地形较低，马路新建后，下雨时房屋

图 3 身份证号码*序列

上述异常符号的出现对文本数据的留言详情做结巴分词时会产生部分影响，为了使分类结果更加准确，需要我们对异常符号进行删除处理。

2.2 利用 TF-IDF 进行文本向量化表示

2.2.1 TF-IDF 思想

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术，如果某个词或短语在一篇文章中出现的频率(TF)高，并且在其他文章中很少出现，则认为该词或者该短语具有很好的类别区分能力，适合用来分类。TF-IDF 公式为：$tfidf_j = tf_{i,j} * idf_j$

TF 指的是词频，表示词条在文本中出现的频率。

TF 公式为：$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$，k：所有词语，j：各文件，$n_{i.j}$：词频

IDF 指的是逆文本频率指数，如果包含词条 t 的文档越少，则 n 越小，IDF 越大，说明词条 t 具有很好的类别区分能力。如果某一类文档 C 中包含词条 t 的文档数为 m，而其它类包含 t 的文档总数为 k，显然所有包含 t 的文档数为 n=m+k，当 m 较大时，n 也较大，按照 IDF 公式得到的 IDF 的值会小，就说明该词条 t 类别区分能力不强。

IDF 公式为：$idf_j=\lg\frac{|D|}{1+|\{j:t_i\in d_j\}|}$，D：所有文件，$d_j$：每个文件词语集合，$t_i$：词

2.2.2 将文本数据转换为向量

对文本数据中的留言详情进行分词处理之后，划分训练集与测试集，获取训练集与测试集的词频统计结果，将其转换为 tf-idf 权值向量。

第三章 建立文本分类模型

3.1 朴素贝叶斯模型[1]

3.1.1 朴素贝叶斯思想

朴素贝叶斯是一种简单但十分强大的预测建模算法。它的思想基础是对给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个概率最大，就认为此待分类项属于哪个类别。

3.1.2 朴素贝叶斯的定义

（1）设 $x=\{a_1,a_2,.....a_m\}$ 为一个待分类项，而每个 a 为 x 的一个特征属性。

（2）找到一个已知分类的待分类项集合，即：训练样本集

（3）计算得到在各个类别下各个特征属性的条件概率估计，即：

$$P(a_1|y_1),P(a_2|y_1).....P(a_m|y_1);P(a_1|y_2),P(a_2|y_2).....P(a_m|y_m);.....;P(a_1|y_n),.....P(a_m|y_n)$$

（4）如果各个特征属性是独立的，根据贝叶斯定理有如下推导：$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)}$，因为分母对于所有类别是常数，则只需要将分子最大化即可，又各特征属性是条件独立的，故有：$P(x|y_i)P(y_i)=P(a_1|y_i)P(a_2|y_i).....P(a_m|y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j|y_i)$

3.1.3 朴素贝叶斯模型测试结果

模型(1)：基于朴素贝叶斯算法的一级标签分类模型

总数据集：9055 条留言

训练数据集：7244 条留言

测试数据集：1811 条留言

模型建立步骤：

1. 数据预处理：对附件二中的数据进行重复值、x 序列值等进行删除处理，获取训练集与测试集的 tf-idf 权值向量。

2 计算各个类别下各个特征属性的条件概率估计

3 朴素贝叶斯预测：

a) 以测试集中留言详情内容进行预处理并转换为权值向量作为模型的输入得到测试集每条留言详情相对应的一级分类标签值。

b) 将朴素贝叶斯模型预测的一级标签分类值与测试集中真实的一级标签分类值计算 F1-Score。

通过测试数据得到基于朴素贝叶斯模型的一级标签分类预测值，计算得到“micro”（通过先计算总体的 TP、FN、FP、TN,再计算 F1-Score）和“macro”（分别计算每个类别的 F1 得分，然后根据各类别 F1 的权重相同得出最终的 F1-Score）情况下的 F1-Score 分别为 0.813567，0.819534。如表 2 所示：

表 2 朴素贝叶斯模型的 F1-Score

计算 F1-Score 方法	F1-Score
Micro	0.813567
macro	0.819534

从表 2 我们可以看出，利用朴素贝叶斯模型做文本数据的一级标签分类的效果一般。因此，本文建立的朴素贝叶斯模型不适合做文本分类，模型有待继续优化。

3.2 CNN 模型[2]

3.2.1 利用 CNN 模型提高一级标签分类成功率

上述传统的分类器如：朴素贝叶斯模型，文本表示的方法是将其转换为“词袋模型”，主要是根据文本中出现的词频，这样会导致词与词之间的序列信息丢失。但文本数据在分词之后，句子本身相当于切成一块一块，词和词的组合往往会有局部语意。此时可能存在的一个问题就是粒度和语意的矛盾，如果粒度过大，则太稀疏，此时意义不大，粒度过小模型更加不适用。而使用 CNN 模型的话，通过卷积层之后，把每 k 个词组合之后的语意放在一起，得到比较准确的词向量。

3.2.2 CNN 模型的思想

假设我们有一些句子需要对其进行分类。句子中每个词是由 n 维词向量组成的，也就是说输入矩阵大小为 m*n，其中 m 为句子长度。CNN 需要对输入样本进行卷积操作，对于文本数据，filter 不再横向滑动，仅仅是向下移动。在不同词窗上应用不同 filter，最终会得到卷积后的向量。然后对每一个向量进行最大化池化操作并拼接各个池化值，最终得到这个句子的特征表示，将这个句子向量运用分类器进行分类。

3.2.3 CNN 模型的基本原理

（1）嵌入层

通过一个隐藏层，将 one-hot 编码的词投影到一个低维空间中，本质上是特征提取器，在指定维度中编码语义特征。这样，语义相近的词，它们的欧氏距离或余弦距离也比较近。（本文使用 word2vec 方法训练得到的词向量）。）

（2）卷积层

在文本-CNN 中，卷积核的宽度是与词向量的维度一致。这是因为我们输入的每一行向量代表一个词，在抽取特征的过程中，词作为文本的最小粒度，高度和 CNN 一样，可以自行设置（通常取值为 2，3，4，5）。由于我们的输入是一个句子，句子中相邻的词之间关联性很高，因此，当使用卷积核进行卷积时，不仅考虑了词义而且考虑了词序及其上下文。

（3）池化层

在卷积层过程中我们使用了不同高度的卷积核，使得通过卷积层后得到的向量维度会不一致，因此在池化层中，使用 1-Max-pooling 对每个特征向量池化成一个值，即：抽取每个特征向量的最大值表示该特征，并且认为这个最大值表示的是最重要的特征。当对所有特征向量进行 1-Max-Pooling 操作后，还需要将每个值给拼接起来，得到池化层最终的特征向量。

（4）全连接层

全连接层跟其他模型一样，假设有两层全连接层，第一层可以加上’relu’作为激活函数，第二层则使用 softmax 激活函数得到属于每个类的概率。

3.2.4 CNN 模型进行一级分类标签预测测试

模型（2）：CNN 文本多分类模型

总数据集：9055 条留言

训练数据集：7244 条留言

测试数据集：1811 条留言

模型建立步骤：

1、数据预处理：

a) 使用 Keras 的 Tokenizer 模块实现转换。当我们创建了一个 Tokenizer 对象后，使用该对象的 fit_on_texts()函数，可以将输入的文本中的每个词编号，编号是根据词频的，词频越大，编号越小。使用 word_index 属性可以看到每次词对应的编码。

b) 将数据集中的每条文本转换为数字列表，使用每个词的编号进行编号

c) 由于每句话的长度不唯一，需要将每句话的长度设置一个固定值。将超过固定值的部分截掉，不足的在最前面用 0 填充。

d) 使用 Embedding 层将每个词编码转换为词向量

2、使用 Word2Vec 词向量的 CNN 模型进行一级标签分类预测，计算该模型的 F1-Score

通过测试数据得到基于 CNN 模型的一级标签分类预测值，计算得到“micro”（通过先计算总体的 TP、FN、FP、TN,再计算 F1-Score）和“macro”（分别计算每个类别的 F1 得分，然

后根据各类别 F1 的权重相同得出最终的 F1-Score）情况下的 F1-Score 分别为 0.841543，0.839872 如表 3 所示：

表 3 CNN 模型的 F1-Score

计算 F1-Score 方法	F1-Score
Micro	0.841543
macro	0.839872

从上表我们可以看出，利用 CNN 模型做文本数据的一级标签分类得到的 F1-Score 比朴素贝叶斯模型高。因此，CNN 模型相对于朴素贝叶斯模型有所优化，但 CNN 模型的 F1-Score 仍旧有待继续提高。

3.3 随机森林模型[3]

3.3.1 利用随机森林模型提高一级标签分类成功率

随机森林（Random Forest）作为一种比较新的机器学习方法，近年来在业内的关注度与受欢迎程度得到逐步提升。它是由多个弱学习器（决策树）组成，在运算量没有显著增加的前提下提高了预测精度，同时其运算结果对缺失数据和非平衡的数据也能达到相当稳健的水平。因此本文最终选择运用“随机森林”算法应用于文本一级标签的分类预测并探究其表现。

3.3.2 随机森林模型基本原理

随机森林使用 Bagging 方式建立一个相互没有关联的决策树森林，这之中的每棵决策树都是基于多个特征进行分类决策，在树的每个结点处，根据特征的表现通过某种规则分裂出下一层的子节点，终端的子节点即为最终的分类结果。由于本文主要是对文本中的留言详情进行一级标签分类，所以当新的输入样本进入时，让森林中的每一颗决策树判断其类别，哪一类被选择最多就将样本预测为那一类。

3.3.3 随机森林算法进行一级分类标签预测测试

将随机森林算法用于上述处理好的文本数据并探究其表现：

模型（3）：基于随机森林算法的一级标签分类模型
总数据集：9055 条留言
训练数据集：7244 条留言
测试数据集：1811 条留言

模型建立步骤：

1 数据预处理：对附件二中的数据进行重复值、x 序列值等进行处理，获取训练集与测试集的 tf-idf 权值向量。

2 参数确定：将参数设置为森林中决策树的个数为 n=100，随机选择的特征个数为 m=sqrt,由于文本数据量较大，max_features 选择 sqrt，取的是特征总数的开方，能有效的减少运行时间。

3 随机森林算法预测：

1）以测试集中留言详情内容进行预处理并转换为权值向量作为模型的输入，得到测试集每条留言详情相对应的一级分类标签值。

2）将随机森林模型预测的一级标签分类值与测试集中真实的一级标签分类值计算 F1-Score。

4 建立一级标签分类预测值与真实值对应的每个类别的混淆矩阵，并绘制每个类别相对应的 ROC 曲线。通过 ROC 曲线下的面积（AUC 值）进一步决定模型的好坏。

AUC 值为 0.5~0.7：效果较差

AUC 值为 0.7~0.85：效果一般

AUC 值为 0.85~0.95：效果较好。

AUC 值为 0.95~1：效果非常好

AUC 值为 1：完美分类器，现实中一般不可能实现。

通过测试数据得到基于随机森林模型一级标签分类的预测值，计算得到“micro”（通过先计算总体的 TP、FN、FP、TN,再计算 F1-Score）和“macro”（分别计算每个类别的 F1 得分，然后根据各类别 F1 的权重相同得出最终的 F1-Score）情况下的 F1-Score 分别为 0.913842，0.902635。如表 4 所示：

表 4 随机森林模型 F1-Score

计算 F1-Score 方法	F1-Score
Micro	0.913842
macro	0.902635

从上表我们可以看出，利用随机森林模型做文本数据的一级标签分类的结果比朴素贝叶斯模型和 CNN 模型准确。因此，本文建立的随机森林模型适合做文本分类。

3.3.5 一级标签类别的 ROC 曲线

利用该模型得到每个类别的 ROC 曲线，从而得到 ROC 曲线下的面积，即：AUC 值。如图 4 得到一级标签为城乡建设的 ROC 曲线：

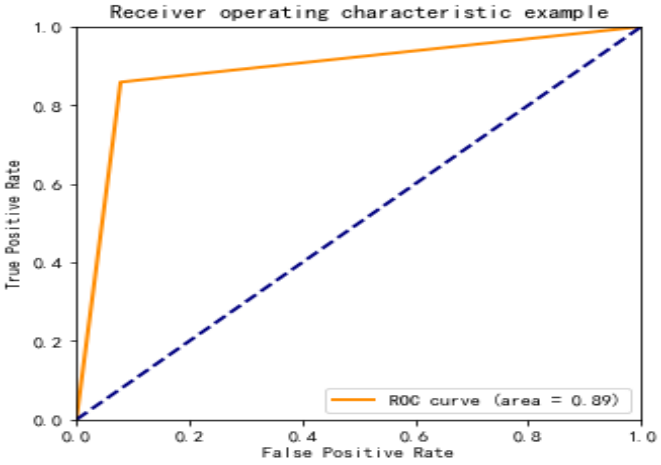

图 4 城乡建设类对应的 ROC 曲线

从图 4 可以看出，城乡建设类的 ROC 曲线下的面积为 0.89，由 AUC 值的分布情况知，该模型对城乡建设类的预测准确率较高。

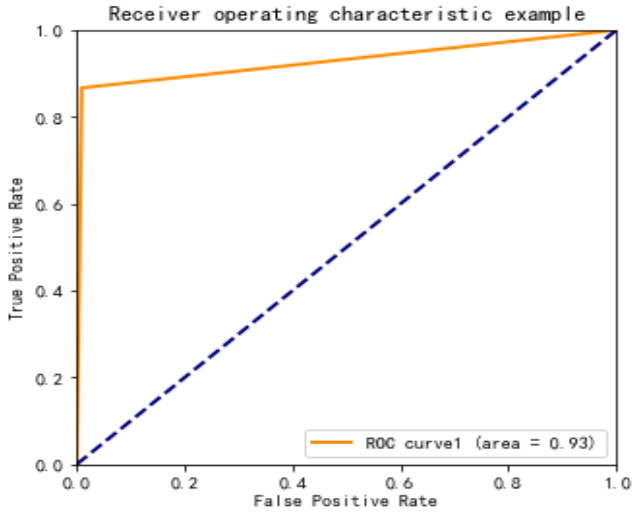

图 5 环境保护类对应的 ROC 曲线

从图 5 可以看出，环境保护类的 ROC 曲线下的面积为 0.93，由 AUC 值的分布情况知，该模型对环境保护类的预测准确率较高。

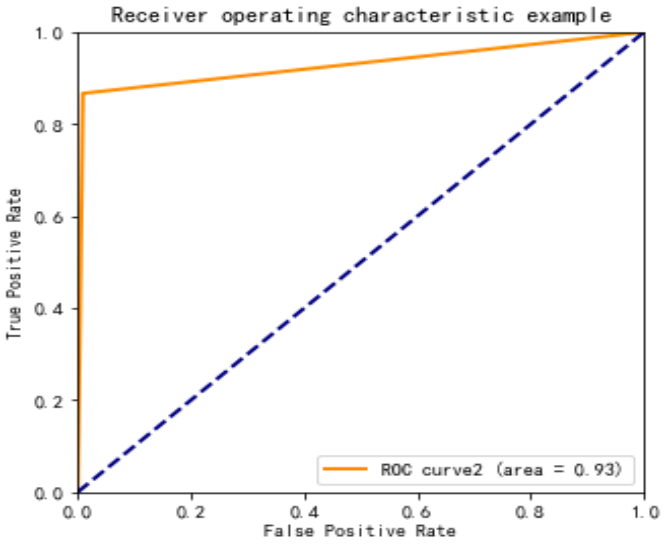

图 6 交通运输类对应的 ROC 曲线

从图 6 可以看出，交通运输类的 ROC 曲线下的面积为 0.93，由 AUC 值的分布情况知，该模型对交通运输类的预测准确率较高。

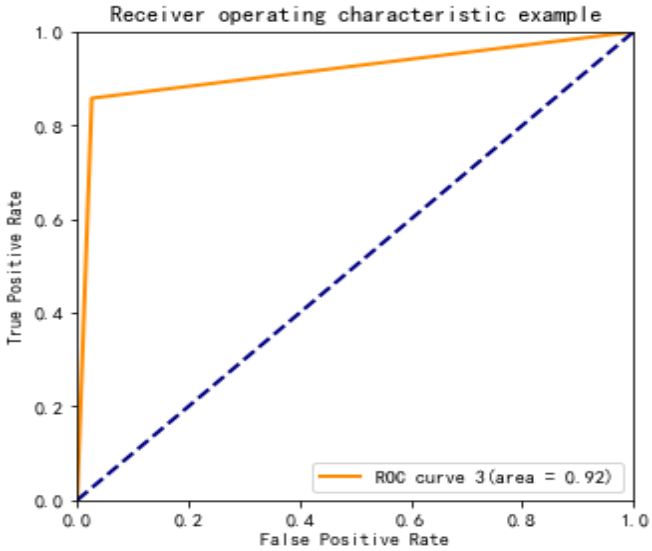

图 7 教育文体类对应的 ROC 曲线

从图 7 可以看出，教育文体类的 ROC 曲线下的面积为 0.92，由 AUC 值得分布情况知，该模型对教育文体类的预测准确率较高。

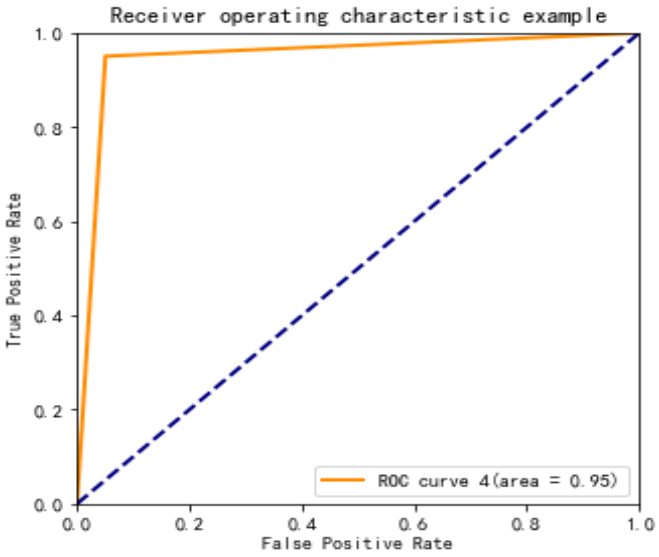

图 8 劳动和社会保障类对应的 ROC 曲线

从图 8 可以看出，劳动和社会保障类的 ROC 曲线下的面积为 0.95，由 AUC 值得分布情况知，该模型对劳动和社会保障类的预测准确率很高。

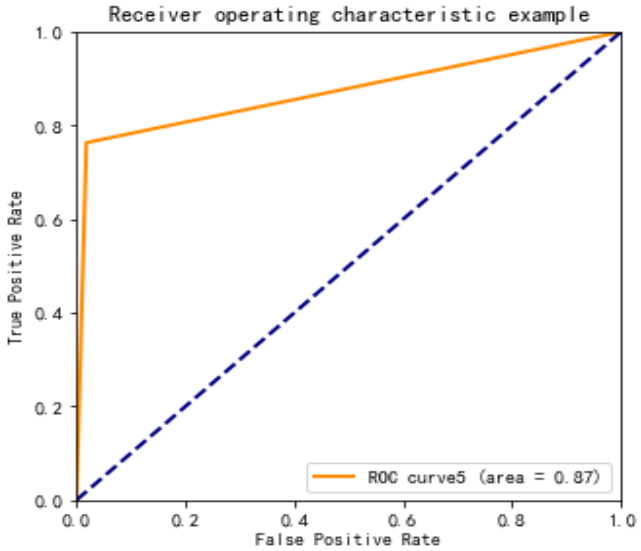

图 9 商贸旅游类对应的 ROC 曲线

从图 9 可以看出，商贸旅游类的 ROC 曲线下的面积为 0.87，由 AUC 值得分布情况知，该模型对商贸旅游类的预测准确率较高。

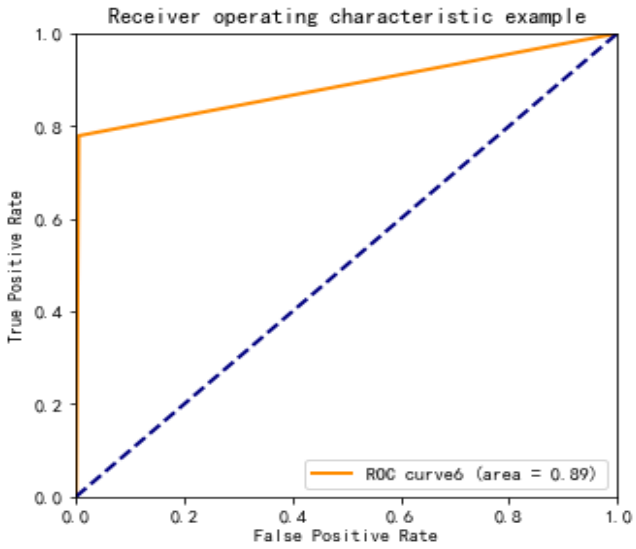

图 10 卫生计生类对应的 ROC 曲线

从图 10 可以看出，卫生计生类的 ROC 曲线下的面积为 0.89，由 AUC 值得分布情况知，该模型对卫生计生类的预测准确率较高。

第四章　附件三文本数据的探索分析

4.1　附件 3 数据预处理

对附件 3 中的文本数据进行数据预处理，减少数据对聚类分析的影响，提高聚类结果的准确度。读取附件 3 表格数据，转化为 DataFrame，后续操作均在 DataFrame 中进行。

4.1.1　去除无效值

将数据中的无效值进行删除。

4.1.2　重复数据的处理

为了使聚类结果更加准确，我们利用程序代码对重复留言进行删除处理。

4.1.3　去除文本数据中的符号

删除文本中的\n、\t 等符号，使文本分析结果更加准确。

4.1.4　统一时间格式

对附件 3 中的“留言时间”格式进行统一，方便计算时长，以及后续表格的制作输出。

4.2　分词

4.2.1 去除停用词

为更好地对文本进行聚类分析，在分词过程中需要对停用词进行剔除，以提高搜索效率。

4.2.2 增加字典

在处理文本数据时，存在许多自定义的地名等文本。为了更加准确地对文本进行处理，需要将自定义的关键词加入字典，避免分词时被剔除。

4.2 .3 结巴分词

利用 Python 中的结巴分词对附件 3 中的“留言主题”和“留言详情”进行分词。

4.2.4 词频统计

将分词结果进行词频统计。

4.2.5 提取关键词

将分词结果中频率较高的单字词进行剔除，作为关键词。

4.2.6 将文本数据转换为向量

将“留言详情”的分词结果转换为 TF-IDF 权值向量。为了提高聚类的效果，提取权值向量中列名属于关键词的向量。

4.3 聚类分析

4.3.1 K-Means 算法定义

K-means 聚类算法也称 k 均值聚类算法，是集简单和经典于一身的基于距离的聚类算法。它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。该算法认为类簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

4.3.2 K-Means 算法过程

1、 首先确定一个 k 值，即我们希望将数据集经过聚类得到 k 个集合。
2、 从数据集中随机选择 k 个数据点作为质心。
3、 对数据集中的每一个点，计算其与每一个质心的距离（如欧氏距离），离哪个质心近，就划分到那个质心所属的集合。
4、 把所有数据归好集合后，一共有 k 个集合。然后重新计算每个集合的质心。
5、 如果新计算出来的质心和原来的质心之间的距离小于某一设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），此时可以认为聚类已经达到期望的结果，算法终止。
6、 如果新质心和原质心距离变化很大，需要重复 3-5 步骤。

4.4 K-Means 聚类结果及分析

4.4.1 k-Means 调参

(1)轮廓系数

轮廓系数（Silhouette Coefficient），是聚类效果好坏的一种评价方式。轮廓系数的值是介于 [-1,1]，越趋近于 1 代表内聚度和分离度都相对较优。

(2)轮廓系数图

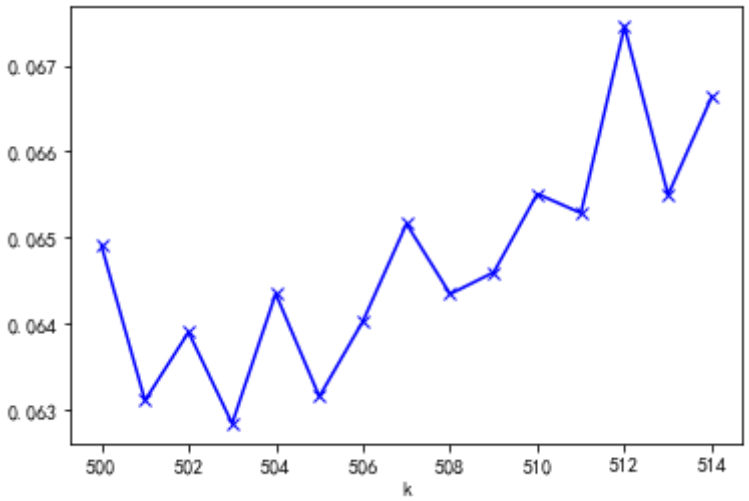

图 11 轮廓系数图

如图，我们可以得出聚类参数 k 等于 512。

4.4.2 K-Means 聚类结果

以下是聚类前 10 的结果：

聚类编号	该聚类的个数
421	112
102	79
69	46
93	36
204	35
35	31
279	29
392	28
67	27
152	25

图 12 部分 K-Means 聚类结果

4.4.3 K-Means 可视化处理

主成分分析的主要原理是将高维数据投影到较低维空间，提取多元事物的主要因素，揭示其本质特征。它可以高效地找出数据中的主要部分，将原有的复杂数据降维处理。可视化结果如下：

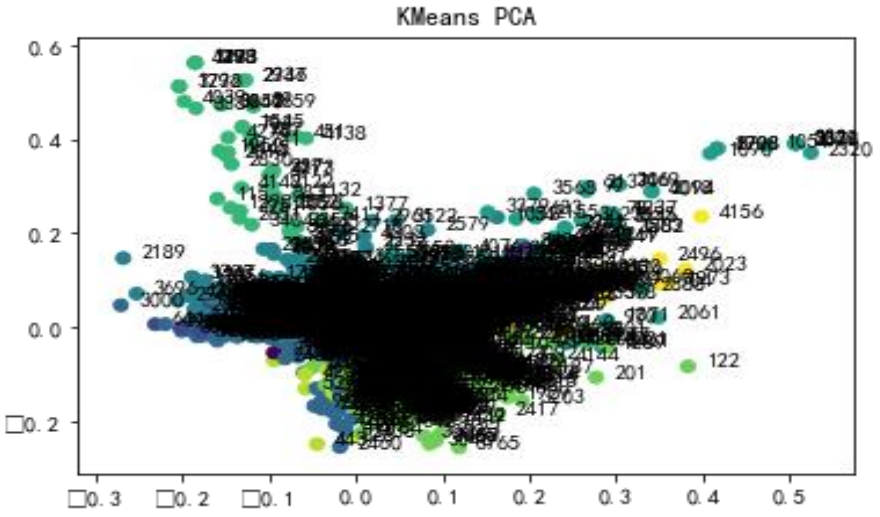

图 13 二维 PCA

如上图所示，由于数据样本量较大，二维 CPA 可视化效果并不理想，所以下面给出三维 PCA 图形。

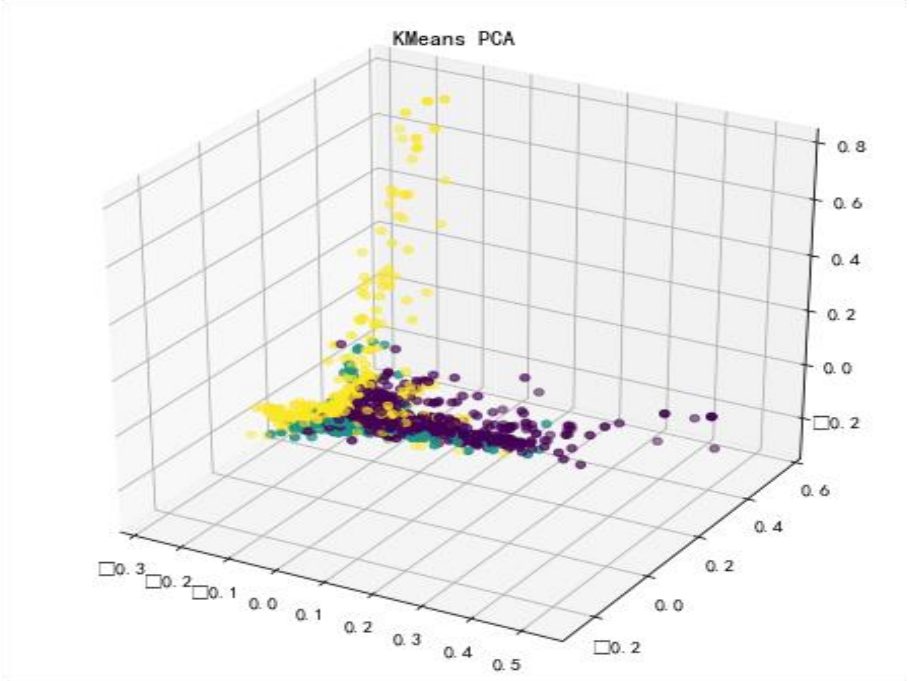

图 14 三维 PCA

由上图可以看出，三维 PCA 可视化效果较好。

第五章　热度评价模型

5.1　综合评价法

5.1.1　RSR 综合评价法[4]定义

综合评价（Comprehensive Evaluation，CE），也叫综合评价方法或多指标综合评价方法，是指使用比较系统的、规范的方法对于多个指标、多个单位同时进行评价的方法。它不只是一种方法，而是一个方法系统，是指对多指标进行综合的一系列有效方法的总称。

5.2 热度评价模型

5.2.1 热度评价指标

表 5 热度评价指标

标签	标签内容	单位
留言时间长度	该留言问题出现的日期长度	天
最新留言时间	最后一次出现该留言问题的日期	年/月/日
最早留言时间	第一次出现该留言问题的日期	年/月/日
留言问题总频次	该问题出现的总频数	条
单位时间出现的问题频数	每一天出现该留言问题的频数	条/天
点赞总数	该问题获得的总点赞数	个
反对总数	该问题出现的总反对数	个

5.2.2　建立热度评价模型过程

1、列出原始数据表并编秩

（1）整次秩和比

将 n 个评价对象的 m 个评价指标排列成 n 行 m 列的原始数据表。编出每个指标各评价对象的秩，其中效益型指标从小到大编秩，成本型指标从大到小编秩，同一指标数据相同者编平均秩。得到秩矩阵，记为

$$R=(R_{ij})_{m\times n}$$

（2）非整次秩和比

用以克服 RSR 法秩次化时易损失原指标值定量信息的缺点。

对于效益型指标：

$$R_{ij}=1+(n-1)\frac{X_{ij}-\min(X_{1j},X_{2j},\cdots,X_{nj})}{\max(X_{1j},X_{2j},\cdots,X_{nj})-\min(X_{1j},X_{2j},\cdots,X_{nj})}$$

对于成本型指标：

$$R_{ij}=1+(n-1)\frac{\min(X_{1j},X_{2j},\cdots,X_{nj})-X_{ij}}{\max(X_{1j},X_{2j},\cdots,X_{nj})-\min(X_{1j},X_{2j},\cdots,X_{nj})}$$

2、秩和比计算

秩和的计算：

$$RSR_{\mathrm{i}}=\frac{1}{n}\sum_{j=1}^{m}w_jR_{ij}$$

其中 w_i 为第 j 个评价指标的权重。

当指标权重相同时，$w_j=\frac{1}{m}$，秩和可以表示为：

$$RSR_{\mathrm{i}}=\frac{1}{mn}\sum_{j=1}^{m}R_{ij}$$

3、RSR 分布

RSR 的分布是指用概率单位表达的值特定的累计频率 。其转换方法为：

（1）编制 RSR 频数分布表，列出各组频数，计算各组累计频数。

（2）确定各组 RSR 的秩次范围及平均秩次。

（3）计算累计频率并修正。

（4）将累计频率换算为概率单位。

4、直线回归方程

（1）以累积频率所对应的概率单位 $Probit_i$ 为自变量，以 RSR 值为因变量，计算直线回归方程，即：

$$RSR=\mathrm{a}+b\times\mathrm{Pr}obit$$

（2）回归方程检验：

a.残差独立性检验：Durbin-Watson 检验

b.方差齐性检验（异方差性）：Breusch-Pagan 检验和 White 检验

c.回归系数的有效性检验：t检验法和置信区间检验法

d.拟合优度检验：(自由度校正)决定系数、Pearson 相关系数、Spearman 秩相关系数、交叉验证法等。

5.2.3 热度评价结果

OLS Regression Results

Dep. Variable:	y	R-squared:	0.920
Model:	OLS	Adj. R-squared:	0.893
Method:	Least Squares	F-statistic:	34.40
Date:	Fri, 08 May 2020	Prob (F-statistic):	0.00989
Time:	09:58:24	Log-Likelihood:	9.1118
No. Observations:	5	AIC:	-14.22
Df Residuals:	3	BIC:	-15.00
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2251	0.142	-1.590	0.210	-0.675	0.225
Probit	0.1538	0.026	5.865	0.010	0.070	0.237

Omnibus:	nan	Durbin-Watson:	2.949
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.718
Skew:	0.857	Prob(JB):	0.698
Kurtosis:	2.286	Cond. No.	35.0

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

回归直线方程为：y = 0.1537585145951022 Probit - 0.22508890877108195

图 15 模型结果

检验统计量较大，说明该热度评价模型的回归系数具有统计学意义，且拟合优度达到了89.3%，拟合效果较好，通过回归检验。

	X1：时间长度	R1：时间长度	X2：最近留言时间	R2：最近留言时间	X3：最早留言时间	R3：最早留言时间	X4：问题频次	\
1	371	5.0	2020/01/08	4.0	2019/01/02	1.0	137	
2	365	4.0	2020/01/08	4.0	2019/01/08	2.0	52	
3	348	2.0	2019/12/26	2.0	2019/01/12	3.0	43	
4	56	1.0	2019/09/01	1.0	2019/07/07	5.0	40	
5	350	3.0	2019/12/31	3.0	2019/01/15	4.0	33	

	R4：问题频次	X5：单位时间出现的问题	R5：单位时间出现的问题	X6：点赞数	R6：点赞数	X7：反对数	R7：反对数	\
1	5.0	0.369272	4.0	132	5.0	11	3.0	
2	4.0	0.142466	3.0	84	4.0	18	5.0	
3	3.0	0.123563	2.0	63	3.0	13	4.0	
4	2.0	0.714286	5.0	25	2.0	1	1.0	
5	1.0	0.094286	1.0	23	1.0	2	2.0	

	RSR	RSR_Rank	Probit	RSR Regression Level
1	0.771429	1.0	6.644854	0.796614 2
2	0.742857	2.0	5.841621	0.673110 3
3	0.542857	3.0	5.253347	0.582658 3
4	0.485714	4.0	4.746653	0.504749 4
5	0.428571	5.0	4.158379	0.414297 4 ,

	f	Σ f	\bar{R} f	\bar{R}/n*100%	Probit
0.428571	1	1	1.0	0.20	4.158379
0.485714	1	2	2.0	0.40	4.746653
0.542857	1	3	3.0	0.60	5.253347
0.742857	1	4	4.0	0.80	5.841621
0.771429	1	5	5.0	0.95	6.644854)

图 16 RSR 模型分析结果报告

将 RSR　Regression 作为热度评价结果，分别为：

0.796614,　0.673110,　0.582658,　0.504749,　0.41429

第六章　答复意见的质量特征提取

6.1 文本预处理相关技术

为保证附件 4 中的数据质量，降低无关数据对结果准确率的影响，需要在评价的初始阶段对文本数据进行预处理，去掉原始数据中的噪声数据，规范数据的格式。预处理的过程如下：

(1)利用 jieba 分词对答复意见以及留言详情进行分词；

(2)为了保证评价的准确率，建立停用词表，删除文本内容中的无用字词以及标点符号，比如：“的”，“了”，“得“及一些无实际意义的副词、介词等。

(3)删除留言详情中的特殊字符

6.2　指标的提取

6.2.1 相关度

留言意见中每一条答复都对应一个留言详情。考虑到留言意见与留言详情之间的相关程度越高，越能表明答复意见的质量越高。计算对应的距离来计算其相关度。

将两列文本数据分词后，把完整的句子根据算法分为独立的词集合，求出两个集合的并集得到一个词包，将各自词集进行向量化，得到 TF-IDF 矩阵，将每一行留言详情与其对应的答复意见量化为二者的余弦相似度。当二者空间距离越近，就说明它们之间的相似性越高。

6.2.2 规范性

数据的完整性是描述信息的完整程度，为了将答复意见进行量化处理，我们将这一指标设立为规范性。数据的规范性是指描述内容的表达形式需要符合国内外的相关规范、有一定的标准。相关政府部门关于群众的反馈，其答复内容需要体现专业以及规范程度，它们都将遵循一个范式。首先建立一个规范的答复模板，计算答复意见中是否满足这一规范。将不符合模板的视为完整性不高。规范性的计算公式为：

数据集中符合规范的数量/数据集中记录的总数

指标提取量化的过程中，将答复意见的模板以及答复意见进行分词、去停用词等操作，遍历答复意见中的每一条内容，用 Sklearn 库中的 CountVectorizer 来计算句子的 TF 矩阵，然后利 Numpy 来计算二者的交集和并集，最后计算每一条答复意见与答复模板之间的 Jaccard 相似系数。Jaccard 相似系数越大，样本的相似度越高，说明答复意见的规范性越强。

6.2.3 可读性[5]

答复意见的可解释性是指人们理解的难易程度。由于文本深度代表了撰写者答复内容的深入性，而答复意见的深度外在表现即答复长度。因此可将答复的可读性用答复意见的长度来衡量。答复文本的长度的越长，答复意见中含有有效且可解释的内容可能越多，因此文本的长度视为可读性的量化指标是合理的。

在该指标提取过程中，对每一条答复意见分别提取他们的字符数，考虑到将文本长度作为特征权重时，一般的特征权重值大于 1 的可能性较小，因此将长度视为权值在[0,1]之间的特征项，对答复的长度作归一化处理，

其公式如下所示：

$$l=\frac{x-\min(d)}{\max(d)-min(d)}$$

其中，x代表每一行字符的长度，$\min(d)$代表出现的文本中字符长度的最小值，$\max(d)$出现的文本中字符的最大值，l代表处理后的数据。

6.3 指标特征提取结果展示

由表 6 展示指标特征提取后的部分结果。详细的结果请看数据附件 5。

表 6 部分指标数据

留言用户	相关性	规范性	可读性
A00045581	0.61927836	0.209835302	0.5
A00023583	0.31598499	0.391222614	0.25
A00031618	0.470987568	0.359873538	0.25
A000110735	0.279219821	0.202662477	0.25
A0009233	0.496202852	0.270211782	0.25
A00077538	0.103920592	0.140245059	0.25
A000100804	0.373019689	0.169396917	0.25
UU00812	0.373019689	0.181968375	0.5

第七章 评价模型的构建

在构建评价模型过程中，根据获得的评价指标建立质量评价模型，然后将答复意见的数据分为训练集和测试集，并利用训练集中的数据获得各个评价指标的权重和利用测试集来验证模型的性能。

7.1 广义线性模型

广义线性模型也叫做广义线性回归模型，它是一般线性模型的扩展，在我们的评价中，我们利用多元线性回归模型，即有两个或者两个以上的影响因素作为自变量来解释因变量的变化。若设 Y 为因变量，$X_1,X_2,X_3\cdots X_k$ 为自变量，则多元线性回归模型可表示为：

$$Y=b_0+b_1X_1+\cdots+b_kX_k+e$$

其中 b_0 为常数项，$b_1\cdots b_k$ 为回归系数。 在本次构建评价模型中，我们利用广义线性模型对答复意见的质量进行回归分析。

7.2 确定指标的权重

7.2.1 确定权重的方法

表 7 指标介绍

	指标的介绍
指标	说明
相关性	衡量留言问题与答复的相关程度
规范性	衡量留言回复意见是否满足某些特定的规范
可读性	量化答复意见的长度

关于评价文本质量模型的研究中，不同指标的权重将会得到完全不同的结论，因此选择权重会直接影响最终模型的成败。权重的正确建立，应该是因素客观信息的反应，但也是决策者主观判断的结果。

在确定权重时，应用了决策矩阵信息以及数学理论相结合的方法，采用层次分析法得到 W_1 与信息量权数法得到的权重 W_2，将二者相结合取平均值得到最终的权重 W，最后与因子分析法得到的权重 W_3 作比较。下面介绍三个权重的理论及得到的结果

7.2.2 特定主观确定确定的权重w_1-层次分析法

首先构建如下表所示的层次分析结构模型：目标层为答复的质量，决策层为三个指标；结构模型如下图：

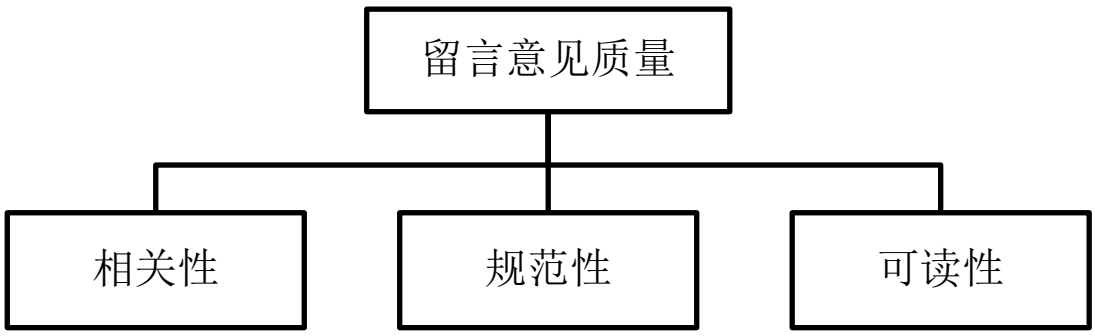

图 17 层次结构模型

其次将决策层中各个元素相对于目标层进行两两重要性比较，并将比较的结果构造成一个比较矩阵：

$$A=\begin{bmatrix} 1 & 2 & \frac{4}{3} \\ \frac{1}{2} & 1 & \frac{3}{2} \\ \frac{3}{4} & \frac{2}{3} & 1 \end{bmatrix},$$

利用 MATLAB 软件计算该矩阵的最大特征根及其对应的归一化向量，最后的到各个指标所对应的权重：

表 8 层次分析法确定权重

层次分析法确定权重w_1	
指标	权重
相关性	0.4489
规范性	0.2941
可读性	0.2569

为了验证该权重确定的排序顺序是否合理，对该对指标层进行了一致性检验，一致性检验的结果如下：

表 9 一致性检验

一致性检验汇总				
最大特征根	CI 值	*RI* 值	CR 值	一致性检验结果
3.0735	0.0368	0.52	0.0707	通过

由CR 值小于 0.1，则上述的比较矩阵满足一致性检验。故指标的权重排序也较为合理。

7.2.3 特定客观确定权重W_2-信息权数法

根据三个评价指标所包含的信息来确定权数。在该过程中采用变异系数法，计算各个指标的变异系数。变异系数越大，所赋给该指标的权数越大。将CV作为权重的分值，再将其进行归一化处理得到四个信息量的权重系数。其中变异系数计算公式为：

$$CV=S_i/x_i$$

其中S为各个指标的标准差 x_i为各个指标的平均值。利用 SPSS 软件得到三个量化指标的平均值以及标准差，结果如下图：

描述统计

	个案数	最小值	最大值	平均值	标准差
相关性	2816	.00000000	.93195579	.348492367	.163293197
规范性	2816	.00000000	.53877611	.2218497090	.0705547629
可读性	2816	.00	1.00	.3631	.29818
有效个案数	2816				

图 18 指标的描述性统计

利用得到的结果计算 CV 值，最后对权值进行归一化得到指标权重的结果为：

表 10 信息权数法确定的权重

信息权数法确定权重W_2	
指标	权重
相关性	0.2935
规范性	0.2144
可读性	0.4921

7.2.4 层次分析法与信息权数法的结合

为了降低质量评价过程中的主观性，采用层次分析法和信息权数法相结合的方式构造权重。并将二者单独的权重进行测试，最终和相结合得到的权重建立的评价模型相比对。相结合得到的权重如下：

表 11 层次分析法与信息权数法确定的权重

相结合得到权重$W_{合}$	
指标	权重
相关性	0.3712
规范性	0.2543
可读性	0.3735

7.2.5 因子分析法确定的权重W_3

根据数理统计中因子分析的方法，对每个指标计算共性因子的累积贡献率来定权。贡献率越大，说明该指标对共性因子的作用越大，所定权数也越大。

利用 SPSS 软件对量化后的数据进行因子分析，首先对标准化的数据进行因子分析，使用方差最大化进行旋转。得到各个指标的相关矩阵、因子得分以及每个因子的贡献率，最后确定该方法下指标的权重W_3

利用 SPSS 软件得到旋转后的成分矩阵为：

旋转后的成分矩阵[a]

	成分	
	1	2
相关性	.815	.242
规范性	.042	.980
可读性	.863	-.131

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

a. 旋转在 3 次迭代后已收敛。

权重计算时用第i主成分上每一个指标对应的得分除以该主成分上的得分之和即得到对应的权重。第i个指标的权重可定义为：

$$w_i=\frac{|a_{i1}|}{\sum_{k=1}^{3}|a_{k1}|}+\frac{|a_{i2}|}{\sum_{k=1}^{3}|a_{k2}|}$$

最终计算的权重结果为：

表 12 因子分析法确定的权重

因子分析法确定权重w_3	
指标	权重
相关性	0.328
规范性	0.377
可读性	0.295

7.3 构建答复意见质量评价模型

根据上述的指标特征，用广义线性模型构建该问题的质量评价模型。*Quality* 表示答复意见的质量，得分记为Q，建立如下的回归模型：

$$Q=w_1a_1+w_2a_2+w_3a_3$$

其中，$w_i=(w_1, w_2, w_3)$表示各指标的权重，a_1表示相关性；a_2表示完整性；a_3表示可读性。

7.4 不同权重下的质量评价结果

7.4.1 层次分析法

由上述层次分析法确定的权重w_1得到评价模型：

$$Q=0.4489a_1+0.2941a_2+0.2569a_3$$

将量化的数据代入该模型，得到留言意见的质量得分。部分结果如下表，全部详细数据请看附件 6

表 13 层次分析法得到的质量评分

相关性	规范性	可读性	质量得分
0.61927836	0.05726257	0.5	0.569033
0.31598499	0.038344337	0.25	0.348749
0.470987568	0.044946673	0.25	0.411608
0.279219821	0.038979177	0.25	0.291818
0.496202852	0.020060945	0.25	0.403704
0.103920592	0.029075673	0.25	0.199743
0.373019689	0.030726257	0.25	0.326792
0.373019689	0.078847131	0.5	0.452513
0.332894311	0.063737938	0.5	0.443277

7.4.2 信息权数法

由信息分析法确定的权重w_2得到评价模型：

$$Q=0.2935a_1+0.2144a_2+0.4921a_3$$

将量化的数据代入该模型，得到留言意见的质量，部分结果如下表：

表 14 信息权数法得到的质量评分

相关性	完整性	可读性	信息加权法得分
0.619278	0.209835	0.5	0.472796888
0.315985	0.391223	0.25	0.299644723
0.470988	0.359874	0.25	0.338416738
0.27922	0.202662	0.25	0.248426852
0.496203	0.270212	0.25	0.326593943
0.103921	0.140245	0.25	0.183594234
0.37302	0.169397	0.25	0.268824978
0.37302	0.181968	0.5	0.394545298
0.332894	0.222905	0.5	0.391545232

7.4.3 层次分析法与信息权数法的结合

由上述二者结合得到的评价模型：

$$Q=0.3712a_1+0.2543a_2+0.3725a_3$$

将指标的量化结果代入模型，得到部分留言质量的得分：

表 15 综合法得到的质量得分

相关性	规范性	可读性	综合
0.619278	0.209835	0.5	0.461115
0.315985	0.391223	0.25	0.294297
0.470988	0.359874	0.25	0.345112
0.27922	0.202662	0.25	0.240222
0.496203	0.270212	0.25	0.335249
0.103921	0.140245	0.25	0.161769
0.37302	0.169397	0.25	0.267909
0.37302	0.181968	0.5	0.363729
0.332894	0.222905	0.5	0.357611

7.4.4 因子分析法

由因子分析法确定的权重w_3得到评价模型：

$$Q=0.328a_1+0.377a_2+0.295a_3$$

将量化的数据代入该模型，得到留言意见的质量，部分结果如下表：

表 16 因子分析法得到的质量得分

相关性	完整性	可读性	因子得分
0.6192784	0.209835302	0.5	0.429731211
0.315985	0.391222614	0.25	0.398634002
0.4709876	0.359873538	0.25	0.437656246
0.2792198	0.202662477	0.25	0.315487855
0.4962029	0.270211782	0.25	0.338374377
0.1039206	0.140245059	0.25	0.160708341
0.3730197	0.169396917	0.25	0.259963096
0.3730197	0.181968375	0.5	0.412202535
0.3328943	0.222904625	0.5	0.414474378

第八章　留言的质量分布

8.1 因子分析法确定的评价模型 1 下的质量分布[6]

在留言数据中，有效数据为 2816 条数据，利用 spss 软件得到文本质量的描述性统计：

表 17 对因子分析法得分的描述性统计

统计		
因子		
个案数	有效	2816
	缺失	0
平均值标准误差		.002497065930
中位数		.30215030200
标准差		.132509292000
方差		.018
范围		.693968961
最小值		.000000000
最大值		.693968961

根据所有留言质量的数据,利用SPSS绘制直方图,观察其分布情况。

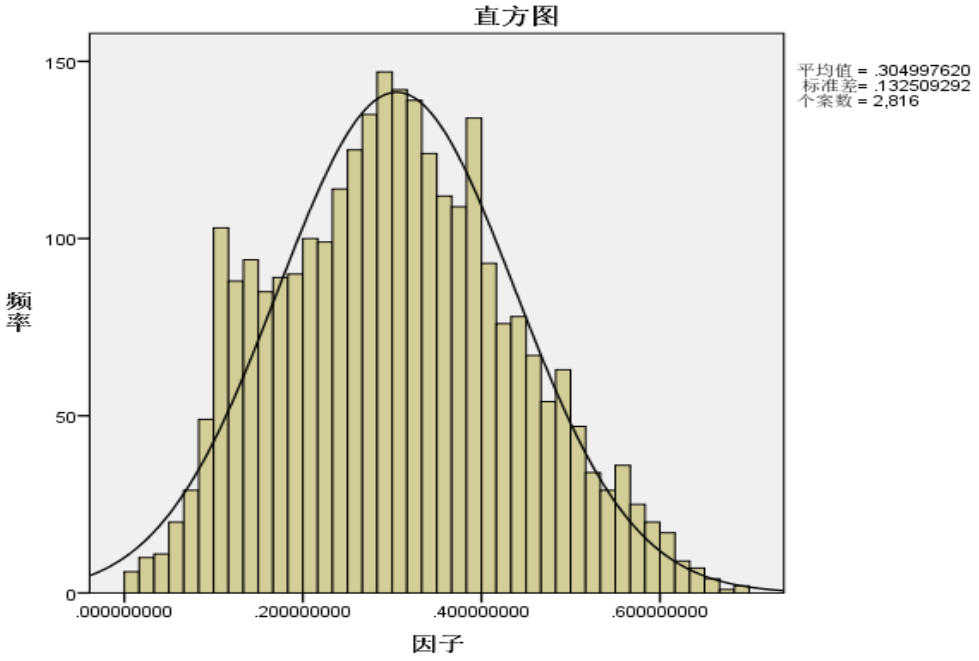

图 19 因子分析法确定模型的质量分布

模型1的质量水平呈正态分布，说明低质量的答复和高质量的答复相差不大，而质量大多分布在0.3-0.4左右。

8.2 综合法确定的评价模型 2 下的质量分布

根据层次分析法和信息权数法确定的模型利用 spss 软件得到文本质量的描述性统计：

表 18 综合法模型下质量的描述性统计

统计		
综合		
个案数	有效	2816
	缺失	0
平均值标准误差		.002976596850
中位数		.31361048600
标准差		.157956078000
方差		.025
范围		.782367724
最小值		.000000000
最大值		.782367724

根据所有留言质量的数据,利用SPSS绘制直方图,观察其分布情况。

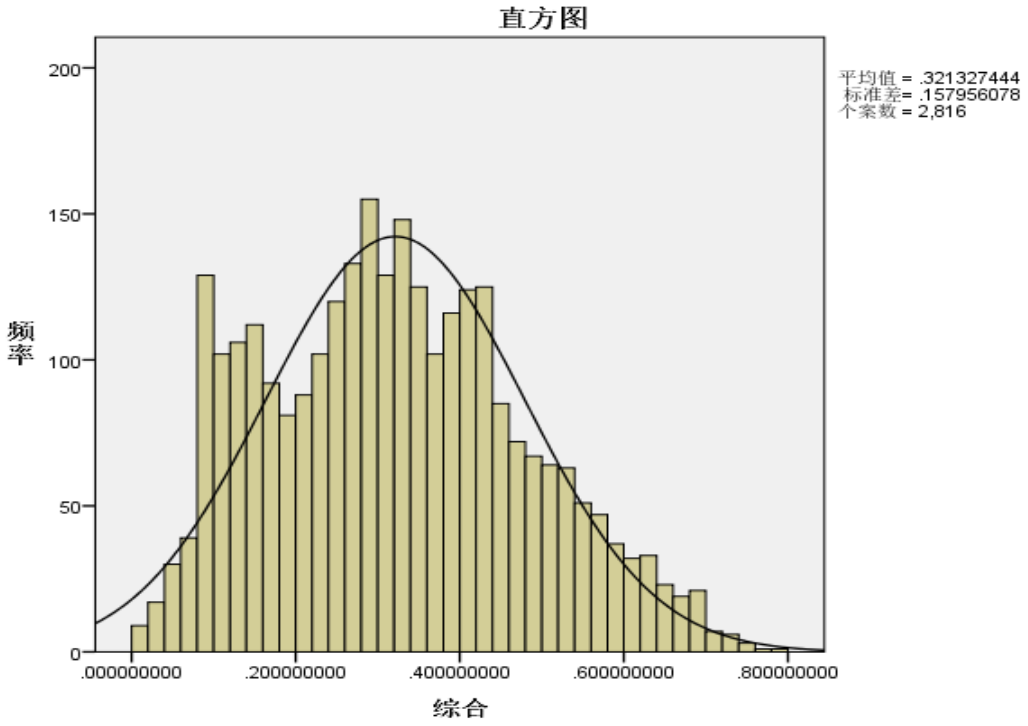

图 20 层次与信息权数法确定模型的质量分布

模型2的质量也是呈偏正态分布，质量大多分布在0.3-0.5左右。相对低质量的文本较多，但高质量与低质量的差距不是特别大。

8.3 模型训练权重

采用后200个数据作为训练集，在训练集上利用广义线性模型进行分析，初步地得到各个指标的权重，W=（0.522 0.207 0.271），本模型变为：

$$Q_1=0.522a_1+0.207a_2+0.271a_3$$

在该模型下质量的分布情况如下图：

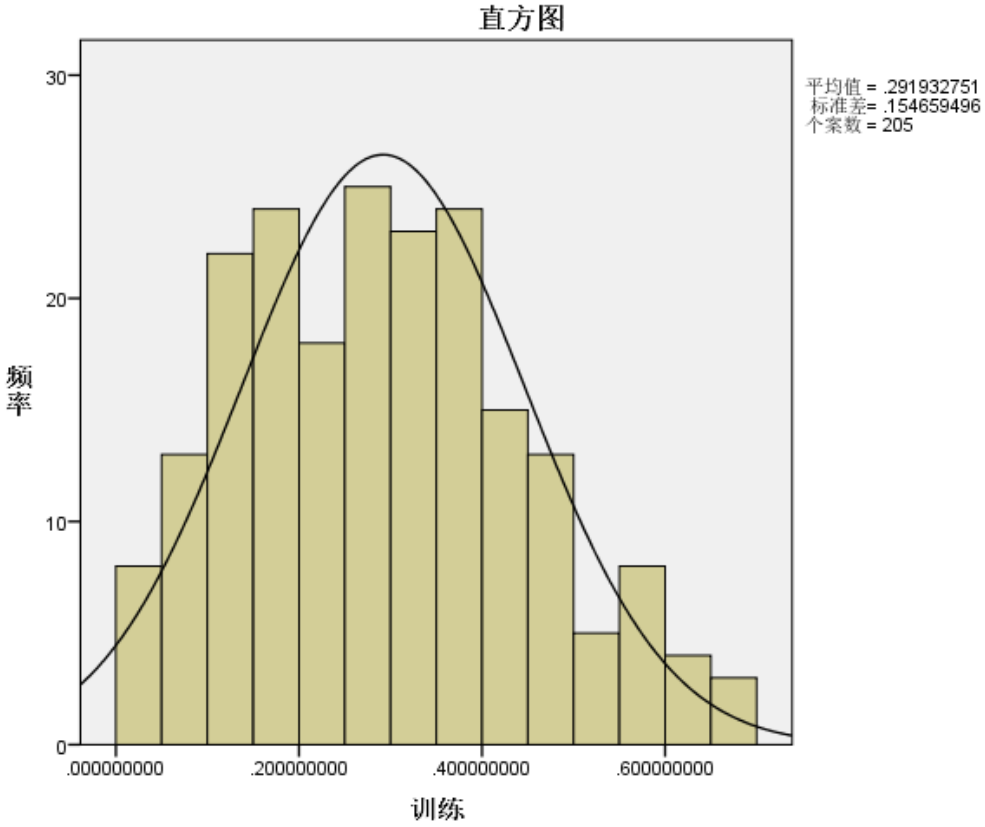

图 21 测试集下的质量分布

在该模型中的质量分布图与层次分析法和信息权数法确定的模型的质量分布图相似，都是呈偏正态分布， 说明基于主客观相结合得到的质量评价模型有较高的准确率。因此在该评价模型中我们最终选取了综合法的评价模型：

$$Q=0.3712a_1+0.2543a_2+0.3725a_3$$

参考文献

[1]张航.基于朴素贝叶斯的中文文本分类及 python 实现[D].山东师范大学,2018.

[2]万磊，张立霞，时宏伟，基于 CNN 的多标签文本分类与研究[J].现代计算机，2020.

[3]张鑫. 随机森林算法的优化研究及在文本并行分类上的应用[D].南京邮电大学,2018.

[4]陈园园,朱滨海.应用 RSR 法综合评价某院近十年的医疗质量[J].南京医科大学学报(社会科学版),2015,15(03):216-219.

[5]郭银灵. 基于文本分析的在线评论质量评价模型研究[D].内蒙古大学,2017.

[6]邹沁含,庞晓阳,黄嘉靖,刘司卓.交互文本质量评价模型的构建与实践——以 cMOOC 论坛文本为例[J].开放学习研究,2020,25(01):22-30.