

第八届“泰迪杯”全国数据挖掘挑战赛论文报告

所选题目：C 题：

基于自然语言处理技术的智慧政务系统

综合评定成绩：

评委评语：

评委签名：

基于自然语言处理技术的智慧政务系统

摘要

随着网络问政平台的不断进步和发展，各类社情民意留言数据的规模越来越大，这给传统的人工留言划分和热点整理工作带来了极大的困难与挑战。本文通过建立基于自然语言处理技术的智慧政务系统，采用机器学习的方法实现对网络问政平台上留言划分和热点整理的智能化处理。

针对问题一：我们首先对留言数据进行分词并剔除停用词，然后将其转换成词频向量并通过 TF-IDF 权重算法转化为权重矩阵，最后借助多项式朴素贝叶斯模型建立关于留言内容的一级标签分类模型。利用 Python 对分类模型进行训练，训练后的分类模型在测试集上的 F1-score 值均在 0.9 左右，这说明该分类模型是有效的。

针对问题二：我们首先对留言内容采用了 jieba 分词和词频统计方法进行数据处理并构建 LdaModel 来对关键词进行检索。然后在多次检索后得到热点问题对应的诸多关键词，并通过对关键词的拼接得到一系列热点问题。最后对每一个热点问题的总数据中出现的频次进行统计，并将其在总数据中所占比例作为热度评价指标，从而得到热点问题表以及热点问题留言明细。排名前五的热点问题为：伊景园车位捆绑销售问题、魅力之城小区临街餐饮店噪音油烟扰民问题、经济学院强制学生定点实习问题、人才购房补贴问题和地铁扫码问题。

针对问题三：我们对相关部门给出的答复意见数据进行分析，筛选出合适的评价指标，建立基于层次分析法的质量评价模型。结果表明：在答复内容相关性方面，意见的可解释程度与内容一致性对其相关性评价影响较大；在表述的完整性方面，意见的语言表达与礼貌性回答对其完整性评价影响较大；在回答的可解释性方面，意见的理由充分程度与意见的可接受程度对其可解释性评价影响较大。

综上所述，本文所建立的基于多种自然语言处理技术的智慧政务系统，可以显著提高政府了解社情民意、处理民生问题的效率。

关键词：自然语言处理；文本分类；智慧政务系统；Lda Model；层次分析法

Abstract

With the continuous progress and development of the online questioning platform, the scale of various social opinion and public opinion message data is increasing, which brings great difficulties and challenges to the traditional manual message division and hotspot sorting work. In this paper, through the establishment of a smart government system based on natural language processing technology, the machine learning method is used to realize the intelligent processing of message division and hotspot sorting on the network inquiry platform.

For problem one: we first segment the message data and remove the stop words, then convert it into a word frequency vector and convert it into a weight matrix through the TF-IDF weight algorithm. Class label classification model. Using Python to train the classification model, the F1-score values of the trained classification model on the test set are all around 0.9, which shows that the classification model is effective.

In response to question two: we first used jieba word segmentation and word frequency statistical methods for data processing and constructed LdaModel to retrieve keywords. Then, after many searches, many keywords corresponding to the hotspot questions are obtained, and a series of hotspot questions are obtained by stitching the keywords. Finally, the frequency of occurrence of each hot issue in the total data is counted, and its proportion in the total data is used as the heat evaluation index, so as to obtain the hot spot question table and the hot spot message details. The top five hotspot issues are: Yijingyuan parking lot bundling sales issues, the noise and smoke disturbance of street restaurants in the charming city community, the mandatory internships of students in the School of Economics, talent subsidies for housing purchases and subway code scanning.

In response to question three: We analyze the response data given by the relevant departments, select suitable evaluation indicators, and establish a quality evaluation model based on AHP. The results show that: in terms of the relevance of the content of the reply, the interpretability of the opinion and the consistency of the content have a greater impact on its relevance evaluation; in terms of the completeness of the expression, the language expression of the opinion and the polite response have a greater impact on its completeness evaluation Large; in terms of the interpretability of

the answer, the adequacy of the reason for the opinion and the acceptability of the opinion have a greater impact on its interpretability evaluation.

In summary, the smart government system based on a variety of natural language processing technologies established in this paper can significantly improve the efficiency of the government in understanding social conditions and public opinions and dealing with people's livelihood issues.

Keywords: natural language processing; text classification; smart government system; Lda Model; AHP

目 录

1	挖掘目标.....	1
1.1	研究背景.....	1
2	分析方法与过程.....	2
2.1	问题分析.....	2
2.2	总体流程图.....	2
3	研究假设.....	3
4	关于留言内容的一级标签分类模型.....	3
4.1	数据预处理.....	4
5	热点问题挖掘.....	8
5.1	数据预处理.....	8
5.2.1	模型概述	10
5.2.2	模型使用	10
5.3	数据导出.....	11
6	基于层次分析法的答复意见质量评价模型.....	11
6.1	留言答复评价判定.....	12
6.2	答复意见质量评定指标.....	12
6.3	构造比较判断矩阵并求指标的权重.....	13
7	不足与展望.....	17
	参考文献	18
	附录	19

1 挖掘目标

1.1 研究背景

近年来，伴随着计算机技术的高速发展与互联网的普遍应用，人与人之间开始习惯使用网络进行更快捷地交流与获取信息，这些信息通常是以文本为载体。第 45 次《中国互联网络发展状况统计报告》便指出，截至 2020 年 3 月，全国网民规模为 9.04 亿，互联网普及率达到 64.5%。在这样庞大的网民数量下，网络文本资源将以指数级增长。

同时，通过微信、微博、市长信箱等渠道建立网络问政平台在不断进步与完善下，各类型社情民意相关文本数据量逐渐增大，这给传统的人工进行留言划分和热点整理的工作带来了极大的困难与挑战。

因此，建立基于自然语言处理技术的智慧政务系统迫在眉睫，这对提升政府处理与反馈民意、汇聚民智、凝聚民气具有极大的推动作用。

1.2 目标任务

题中给出了互联网公开来源的群众问政留言记录、群众留言的标签体系，及相关部门对部分群众留言的答复意见，团队需利用自然语言处理和文本挖掘的方法解决下面三个问题：

问题一：根据附件 2 给出的有关群众问政留言记录的数据，建立关于留言内容的一级标签分类模型。

问题二：根据附件 3 将某一时段内反映特定地点或人群问题的留言归类，定义合理的热度评价指标，并得出评价结果。

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，并尝试实现。

2 分析方法与过程

2.1 问题分析

针对问题一：根据附件 2 所给出的数据，首先对数据进行预处理：去除表格中不需要的部分，对留言进行结巴分词、去除停用词，及对预处理的数据进行函数封装。之后将文本数据进行向量化表示，得到 TF-IDF 权值向量，构建关于留言内容的一级标签分类模型。

针对问题二：根据附件 3 所给出的数据，首先对数据进行清洗，其次对留言进行 jieba 分词，接着再去除停用词，之后进行词频统计，宏观通过对词频的观察了解热点关键词大致情况，构建 LdaModel 来对关键词进行检索，多次检索后得到出现频率最高的词，此时得到的就是相应的热点问题关键词之一，最后对关键词进行进一步拼接得到热点问题。最终导出热点问题表以及热点问题留言明细表。

针对问题三：根据附件 4 相关部门对留言所给出的答复意见，选择合适的评价指标，建立一套基于层次分析法的质量评价方案，在此基础上，结合所给出的留言信息与答复意见，建立关于答复意见的相关性、完整性、可解释性的质量评价模型。

2.2 总体流程图

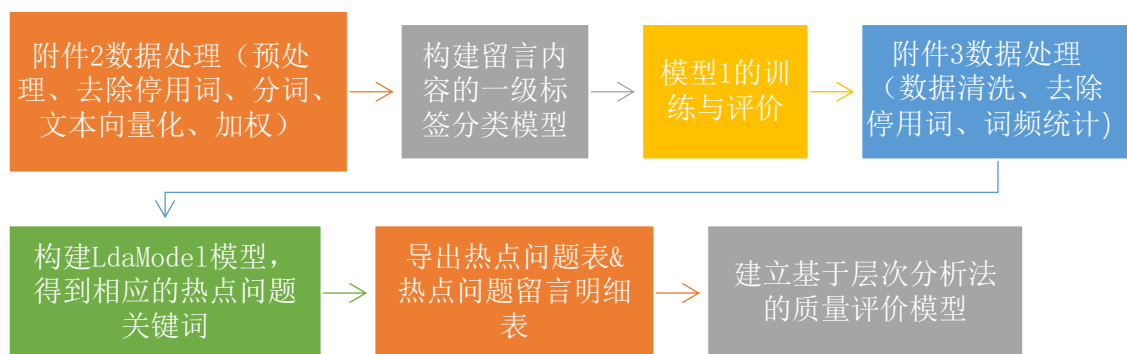


图 2-1 总体流程图

3 研究假设

(1) 假设我们所抽取的留言内容和答复意见的样本为简单随机抽样，具有典型性和普遍性，基本上能够集中反映市民答复意见的相关性、完整性、可解释性。

(2) 在数据计算过程中，假设误差在合理范围之内，对数据结果的影响可以忽略。

(3) 假设模型中各级指标分析因素能代表多数市民对答复意见满意度的指标。

4 关于留言内容的一级标签分类模型

文本分类的基本步骤为：（1）将已有数据分为训练集与测试集；（2）对训练集与测试集进行预处理：数据增强、数据清洗、分词、去停用词等；（3）使用文本分类算法进行模型构建，选择文本特征、计算特征项的权值并进行向量空间化；（4）利用文本分类方法构建分类器，训练分类器模型；（5）将测试集放入训练后的分类器模型中进行分类。分类流程如图所示：

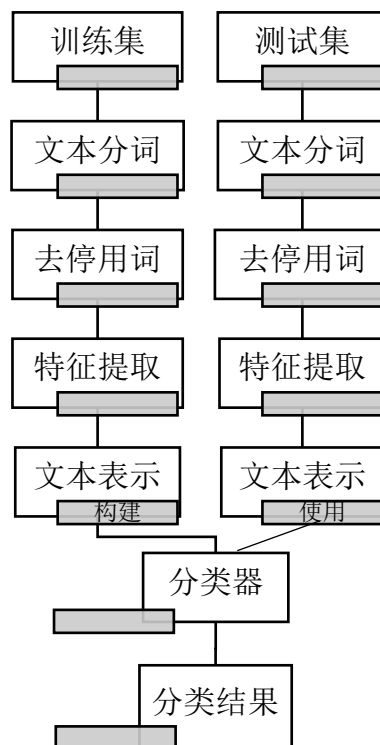


图 4-1 分类流程图

4.1 数据预处理

附件 2 中的数据包括留言编号、用户、主题、时间、详情以及对应的一级分类标签，在建立留言内容的一级标签分类模型时，需要留言主题、详情和对应的一级标签，则将其余无用的数据删除。将留言主题与详情列合并，则文本数据变为包含标签与留言内容两列多行的矩阵。

(1) 数据增强

给出的数据中包含了七类一级标签及对应的留言内容，它们分别是：城乡建设（2009 例）、环境保护（938 例）、交通运输（613 例）、教育文体（1589 例）、劳动和社会保障（1969 例）、商贸旅游（1215 例）和卫生计生（877 例），由此可见数据虽然无缺失值，但存在分布不均衡现象，本文中采用回译法（将目标语言翻译成另外一种语言再翻译回来），借助百度的通用翻译 API 进行数据增强，增强后的数据每一类均为两千例左右，共有数据 14000 行。

(2) 数据清洗

处理文本数据前需要对文本数据进行数据清洗、分词和添词典去停用词等处理。本文中主要借用正则表达式进行对数据的替换、删除或添加，数据清洗则包括去除文本中的空格、软换行、脱敏处理后的*字符串以及文本去重。

(3) jieba 分词

中文文本中词与词之间不像英文一样有明显的分割符，为了有效地处理高频词语，第一步需要对留言内容进行分词。由于中文词在语句中的前后关系较为负载，相同的词在不同的语境中也会有不同的语义，因此中文分词的难度比英文大。本文中的分词方法采用了目前被广泛使用的基于 Python 的中文分词工具：jieba 分词，jieba 分词支持简体、繁体分词，支持自定义词典。根据相关资料显示，jieba 中文分词能够有效地解决未登陆词和会产生歧义的词语，准确率高达 97%，召回率高达 90%。^[1]

(4) 去除停用词

中文表达中最常用的功能词时限定词，如“的”、“一个”、“这”、“那”等，这些词语的使用频率很高，但没有太多的实际含义，将类似于这些词的噪音词作为停用词去除，可以降低句子噪音，减少特征词的数量以此来提

高文本分类的精度。本文中采用常见的停用词表，并根据文本内容增加与删减停用词语。

4.2 文本模型构建

因为计算机只能对于二进制形式的数值进行识别和运算，因此我们需要将文本转化为计算机可以进行数值操作的形式。文本模型的构建在解决文本分类问题上起到了至关重要的作用。比较常用的文本模型有向量空间模型和概率模型。^[2]

本文中采用了向量空间模型来将处理好的分词进行转化，每一条留言内容被看作是含有的特征项所组成的 n 维向量空间，每一个特征项都互不干涉，并且每一个特征值都一一对应到一个权重值，该模型的每个特征项是相互独立的。

4.3 文本特征的选择

特征选择是指在不改变原始空间属性的前提之下，通过特定的方法构造出一个能够表达原始的文本中最小且最佳特征项集合的过程。本文中采用频率统计的方法进行文本特征的选择。

(1) 词频 (Term Frequency, 简称 TF) 即词条在某一文档中出现的频率若大于事先设定的最小频率值，便对该词条予以保留，作为该文档信息的特征项。

(2) 文档频率 (Document Frequency, 简称 DF) 即设定一个最小阈值和一个最大阈值，作为选择特征项的条件，统计所有数据集中出现的同一个词的数量，将其除以文档中词的总数，若频率在最小阈值与最大阈值之间，便作为特征值，反之删除。

4.4 文本特征的加权

当进行完特征选择后，我们获得了文本向量空间的各个特征项，此时我们需要对这些特征项进行加权，特征项的权重可以作为衡量其对文本分类的贡献程度。本文中选择 TF-IDF 权重算法。

权重策略文档中的高频词应具有表征此文档较高的权重，除非该词也是高文档频率词。

TF: Term Frequency, 即关键词词频, 指一篇文档中关键词出现的频率。N 为单词在某文档中的频次, M 为该文档的单词数。

$$TF = \frac{N}{M} \quad (4.1)$$

IDF: Inverse Document Frequency, 指逆向文本频率, 适用于衡量关键词权重的指数。在统计学语言表达下, 最常见的词赋予最小的权重, 较常见的词赋予较小的权重, 而很少见的词给予较大的权重。一个词的逆向文本频率的大小与它的常见程度呈反比。D 为总文档数, D_w 为出现了该单词的文档数。

$$IDF = \log\left(\frac{D}{D_w}\right) \quad (4.2)$$

计算公式为:

$$TF - IDF = TF \times IDF \quad (4.3)$$

可以看出, TF-IDF 与一个词在文档中出现的次数成正比, 与这个词在整个语言中出现的次数成反比。所以一个词在文章中越重要, 那么它的 TF-IDF 值就越大。

4.5 文本分类方法

朴素贝叶斯法分为高斯朴素贝叶斯、伯努利朴素贝叶斯和多项式朴素多项式法, 多项式模型将单词作为粒度, 是一种以词频为基础模型; 而伯努利模型是以文件为粒度的一种基于文档的模型。^[1]

由于多项式朴素贝叶斯法比较适合于离散型数据, 比如文本分类, 所以本文采用了多项式朴素贝叶斯方法作为文本分类方法。

设训练集分为 k 类, 记为 $S = \{S_1, S_2, \dots, S_k\}$, 则每类 S_i 的先验概率为 $P(S_i), i = 1, 2, \dots, k$,

$$P(S_i) = \frac{S_i \text{ 类的样本数}}{\text{训练集总样本数 } n} \quad (4.4)$$

对于测试样本 d , 其中属于 S_i 类的条件概率是

$$P(S_i|d) = \frac{P(d|S_i)P(S_i)}{P(d)} \quad (4.5)$$

条件概率和先验概率在计算中都使用了最大似然估计法，计算出来的是相对频率值，以此实现训练集出现的概率最大的目标。

$P(d)$ 在所有类别的情况下都是常数，可以忽略，则式可简化为

$$P(S_i|d) = P(d|S_i)P(S_i) \quad (4.6)$$

为了避免 $P(S_i) = 0$ ，采用拉普拉斯平滑（加 1 平滑）公式进行估计^[3]

$$P(S_i) = \frac{1 + |D \times S_i|}{|S| + |D \times S|} \quad (4.7)$$

其中 $|S|$ 为训练集中类的数量， $|D \times S_i|$ 为训练集中 S_i 类的留言数， $|D \times S|$ 为训练集中包含的总留言数。若训练集中各大类的样本数相等时，则类的先验概率都相等。

4.6 模型训练与评价

本文中对模型的训练使用基于 Python 的机器学习库 sklearn^[4]，之后用训练后的分类器对测试集进行检验，并用 F1-score 值进行评价。

首先定义三个统计值：TP：True Positive，指预测答案正确的数量；FP：False Positive，指错将其他类预测为本类的数量；FN：False Negative，指将本类标签预测为其他类标的数量。

之后根据这三个统计值计算每个类别下的精确度和召回率，精确度（precision）：指被分类器判定正例中正确样本的比例，精确度越高，模型某类的分类效果就越好。

$$Precision = P = \frac{TP}{TP + FN} \quad (4.8)$$

召回率（recall）：指的是被预测为正例的占总的正样本的比重：

$$Recall = R = \frac{TP}{TP + FN} \quad (4.9)$$

F-Measure（又被称为 F-Score）综合考虑精确度和召回率，他是精确度和召回率的加权调和评价。

$$F = \frac{(\alpha^2 + 1) \times PR}{\alpha^2 \times (P + R)} \quad (4.10)$$

当 $\alpha = 1$ 时，就是最常见的 F_1 值：

$$F_1 = \frac{2PR}{P + R} \quad (4.11)$$

采用训练后的多项式朴素贝叶斯模型，对测试集进行分类，并进行评价计算，我们得出模型的置信度为 92.06%，分类指标的文本报告为（表中的标号 1~7 分别对应一级标签城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生）：

表 4-1 分类指标的文本报告

	Precision	Recall	F1-score	Support
1	0.88	0.84	0.84	381
2	0.91	1.00	0.95	379
3	0.92	0.95	0.93	392
4	0.94	0.91	0.93	402
5	0.89	0.93	0.91	390
6	0.94	0.97	0.90	427
7	0.95	0.95	0.95	426
accuracy			0.92	2797
Macro avg	0.92	0.92	0.92	2797
Weighted avg	0.92	0.92	0.92	2797

由表 4-1 可知，在此学习后的模型中，精确度、召回率和 F_1 值均较为理想。 F_1 值在 1 时达到最佳值（完美的精确度和召回率），最差为 0，七类一级标签的 F1-score 值都在 0.9 左右，则说明本模型是有效的。

5 热点问题挖掘

文本分类的基本步骤为：（1）将附件 3 数据进行预处理：数据清洗、分词、去停用词等；（2）将已经预处理后的数据带入 LdaModel 中进行关键词分析，多次代入后得到出现频率最高的关键词；（3）对关键词进行拼接得到相关热点问题语句；（4）导出“热点问题表.xls”以及“热点问题留言明细表.xls”。

5.1 数据预处理

（1）数据清洗

由于使用 Python 读取数据时会默认读取文件的换行符、制表符等特殊符号，其次在文本中也会出现大量特殊表情符号等无用信息，因此需要对数据进行清洗得到相关便于后续操作的数据。在此我们采用了正则表达式，对读取的结果进行了数据清洗。

```
#去除文件中的\t,\n, x序列(正则表达式)
data1 = data['留言主题']
data_clear = data1.dropna().apply(lambda x: re.sub('\t', '',x))
data_clear2 = data_clear.dropna().apply(lambda x: re.sub('\n', '',x))
data_clear3 = data_clear2.dropna().apply(lambda x: re.sub('x', '',x))
data['留言主题'] = data_clear3
```

图 5-1 数据清洗代码

(2) jieba 分词:

由于读取到的文本是以字符串的形式存储在对象当中的，因此要处理时我们无法准确处理每一句文本，此时我们通过导入 Python 的 jieba 库来对每一句数据进行分词（jieba 分词的优势见问题一阐述），得到一个个的词语字符串便于我们后续的词频统计以及其他的处理。

```
#jieba分词
jieba.load_userdict('dictionary.txt')
data_cut = data['留言主题'].apply(lambda x:jieba.lcut(x))
data['留言主题'] = data_cut
```

图 5-2 jieba 分词代码

(3)去停用词:

在进行 jieba 分词之后虽然我们得到了一个个词汇，但是其中仍然掺杂了许多干扰我们判断的词语，如：哈哈，的，啊等。因此为了进一步处理并确保我们之后使用的数据足够精确，我们选取了特定的停用词表并进行了相关补充来对数据进行更深层次的处理。

```
#去除停用词
stopWord = pd.read_csv('呆萌的停用词表.txt',sep = 'hahaha')
stopWord = [' ','.',',','A市','A[1-9]'] + list(stopWord.iloc[:,0])
data_sw = data['留言主题'].apply(lambda x:[i for i in x if i not in stopWord])
data['留言主题'] = data_sw
```

图 5-3 去停用词代码

5.2 LdaModel 的构建

5.2.1 模型概述

LDA (Latent Dirichlet Allocation) 是一种由文档主题生成的模型, 也称为三层贝叶斯概率模型, LDA 包含词、主题和文档三层结构。所谓生成模型, 就是我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题, 并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布, 主题到词服从多项式分布。^[5]

我们首先定义文档集合为 D , 主题集合为 T 。

D 中的每个文档 d 我们可以把它看成一个单词序列 $\langle w_1, w_2, \dots, w_n \rangle$, w_i 表示第 i 个单词, 我们假设 d 有 n 个单词:

D 中的所有不相同的单词组成 VOC;

对每个 D 中的 d , 对应到不同主题的概率为 $\theta_d = \langle \theta_{d1}, \dots, \theta_{dt} \rangle$, θ_{di} 表示 d 对应 T 中第 i 个主题的概率, 计算公式为

$$\theta_{di} = n_{di} / n$$

其中 n_{di} 表示 d 中对应第 i 个主题词的数目, n 是 d 中所有词的总数;

对每个 T 中的主题, 生成不同的单词概率为 $\phi_t = \langle \phi_{t1}, \dots, \phi_{ti} \rangle$, 其中, ϕ_{ti} 表示 t 生成 VOC 中第 i 个单词的概率, 计算公式为

$$\phi_{ti} = \frac{N_{ti}}{N} \quad (5.1)$$

其中 N_{ti} 表示对应到主题 t 的 VOC 中第 i 个单词的数目, N 表示所有对应到主题 t 的单词总数。

核心公式:

$$p(w|d) = p(w|t) * p(t|d) \quad (5.2)$$

通过 θ_d 和 ϕ_t 给出文档 d 中出现单词 w 的概率。其中 $p(t|d)$ 利用计算 θ_d 得到, $p(w|t)$ 利用 ϕ_t 计算得到^[6]。

5.2.2 模型使用

首先我们通过对词频进行了统计, 对出现频率最高的热点词进行了宏观了解, 为下一步调用 LdaModel 进行铺垫。

```
#词频统计
word_fre = {}

for i in data['留言主题']:
    for j in i:
        if j not in word_fre.keys():
            word_fre[j] = 1
        else:
            word_fre[j] += 1
```

图 5-4 词频统计代码

接着我们采用了 LdaModel 对关键词进行检索，通过对相关词语词频的宏观掌握以及对 LdaModel 的多次调用得到的输出结果，两者进行拟合得到了附件 3 中热点问题的关键词汇。

```
#调用LdaModel模型
data_sw.columns = ['comment']
mid = list(data_sw)
dictionary = Dictionary(mid)
bow = [dictionary.doc2bow(comment) for comment in mid]
model = LdaModel(corpus = bow, id2word = dictionary, num_topics = 5)
for i in range(5):
    print(model.print_topic(i))
```

图 5-5 LdaModel 使用代码

5.3 数据导出

在得到热点问题关键词汇后，我们对相关关键词进行组合拼接得到了相关的热点问题。随后我们将热点问题按照出现频次进行排序汇总得到了“热点问题表.xlsx”，在已经建立好热点问题表的基础上，我们检索了符合热点问题主题的每一条留言并将其归类至“热点问题留言明细表.xlsx”中，最终得到结果。

6 基于层次分析法的答复意见质量评价模型

我们首先根据数据中给出的答复意见，建立留言意见评价指标，结合留言内容的阐述，建立关于答复意见的相关性、完整性、可解释性的质量评价模型。

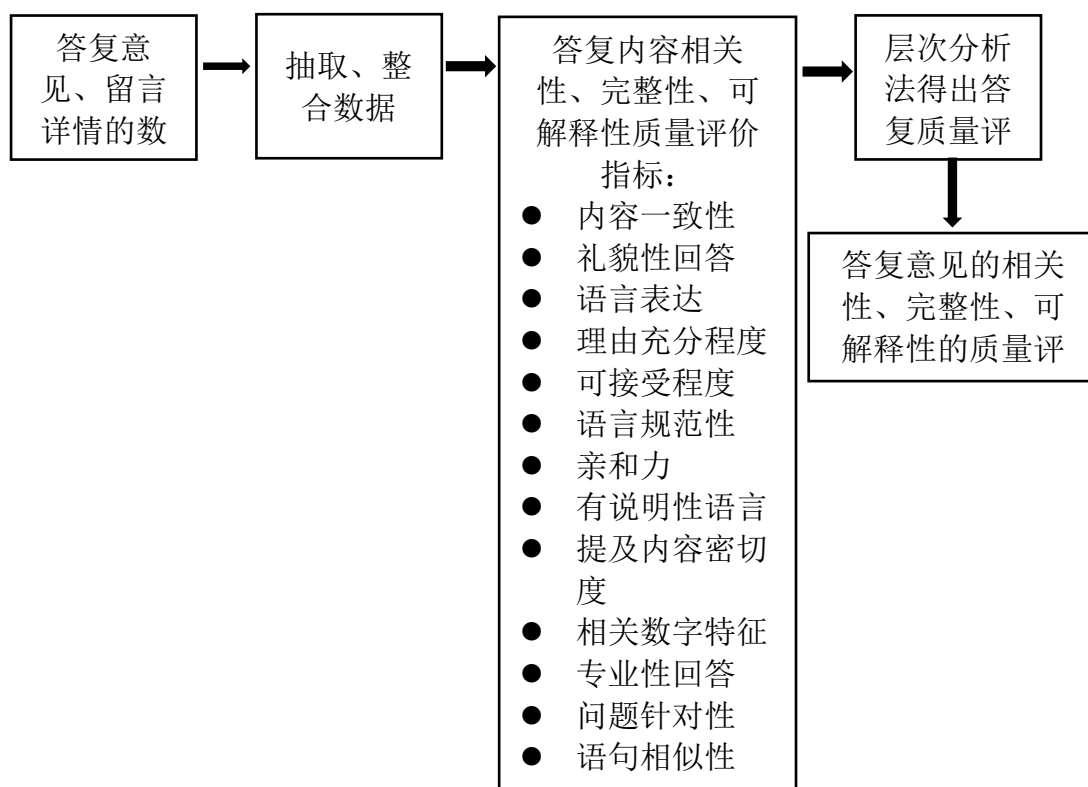


图 6-1 答复意见的相关性、完整性、可解释性的质量评价模型

6.1 留言答复评价判定

我们根据答复意见质量的要求，确保回复内容严格依据现行的方针政策和工作实际，对需要调查核实的留言，办理部门要坚持做到充分了解情况，不得采用利益相关单位的答复。回复要做到语言朴实、简洁、亲切、既要表现出个性化，更要展示出人性化。

本论文计算和提取 13 项答复意见质量评价指标，其中相关性指标、完整性指标、可解释性指标均为 5 项，包括：内容一致性、语言规范性、亲和力、有说明性语言、提及内容密切度、相关数字特征、专业性回答、问题针对性、语句相似性、礼貌性回答、语言表达、理由充分程度、可接受程度。

6.2 答复意见质量评定指标

网络问政平台逐步成为政府了解民意、凝聚民气的重要渠道，市民的留言体现在城乡建设、环保卫生、交通运输、教育文化、劳动与社会保障等方面，而在“智慧政务”系统上为市民答疑解惑成为提升政府管理水平的关键一步。

通过对市民的留言内容以及相应的答复意见的分类处理，可以对留言内容去除表格中不必要的空格、符号，我们对答复意见质量指标进行分级。

在数据整合后，我们针对这些答复意见得出了以下三级评价指标。

表 6-1 为答复意见质量指标

一级评价指标	二级评价指标	三级评价指标
答复内容的相关性	内容一致性	词语间相似度
	语言规范性	称谓、问候语
	亲和力	语言朴实简洁
	有说明性语言	相关数字特征
	可接受程度	解决问题的能力
表述的完整性	提及内容密切度	词语间相似
	礼貌性回答	语言规范性
	语言表达	亲和力
	相关数字特征	有说明性语言
	可接受程度	相关政策的落实
回答的可解释性	专业性回答	专业性术语表达
	问题针对性	有强烈说服力
	语句相似性	有效性解释
	理由充分程度	有说明性语言
	可接受程度	解决问题的能力

6.3 构造比较判断矩阵并求指标的权重

根据表 1 中三个级别的评价指标，利用 1~9 标度法（表 2），通过对同种级别评价指标重要程度的相互比较，构建出对应的比较判断矩阵。用 a_{ij} 表示第 i 个因素相对于第 j 个因素的比较结果，则

$$A = (a_{ij})_{n \times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \quad (6.1)$$

A 被称为成对比较矩阵；我们认为一级评价指标相比有同等重要性，所以三者权重各占 1/3。

表 6-2 标度法

重要性标度	含义
1	表示两个元素相比，具有同等重要性
3	表示两个元素相比，前者比后者稍重要
5	表示两个元素相比，前者比后者明显重要
7	表示两个元素相比，前者比后者强烈重要
9	表示两个元素相比，前者比后极端重要
2, 4, 6, 8	表示上述的中间值
倒数	若元素 i 与元素 j 的重要性之比为 a_{ij} , 则元

得到的二级指标的判断矩阵：

$$B_1 = \begin{pmatrix} 1 & 4 & 3 & 2 & 1/2 \\ 1/4 & 1 & 1/2 & 1/3 & 1/3 \\ 1/3 & 2 & 1 & 1/2 & 1/4 \\ 1/2 & 3 & 2 & 1 & 2 \\ 2 & 3 & 4 & 1/2 & 1 \end{pmatrix} \quad (6.2)$$

$$B_2 = \begin{pmatrix} 1 & 1/2 & 1/3 & 2 & 4 \\ 2 & 1 & 1 & 3 & 2 \\ 3 & 1 & 1 & 4 & 3 \\ 1/2 & 1/3 & 1/4 & 1 & 2 \\ 1/4 & 1/2 & 1/3 & 1/2 & 1 \end{pmatrix} \quad (6.3)$$

$$B_3 = \begin{pmatrix} 1 & 3 & 2 & 1/3 & 1/2 \\ 1/3 & 1 & 1/2 & 1/4 & 1/3 \\ 1/2 & 2 & 1 & 1/3 & 1/3 \\ 3 & 4 & 3 & 1 & 2 \\ 2 & 3 & 3 & 1/2 & 1 \end{pmatrix} \quad (6.4)$$

利用规范列平均法求权重，首先对 B_1, B_2, B_3 矩阵的列向量归一化得到对应矩阵 C_1, C_2, C_3 ，再对其进行算术平均，得到特征向量： $\omega_1, \omega_2, \omega_3$ 。

$$\omega_1 = (0.2846, 0.0686, 0.1013, 0.2530, 0.2925)^T \quad (6.5)$$

$$\omega_2 = (0.1871, 0.2785, 0.3439, 0.1042, 0.0863)^T \quad (6.6)$$

$$\omega_3 = (0.1690, 0.0722, 0.1097, 0.3918, 0.2574)^T$$

(6.7)

对构造的判断矩阵进行一致性检验，用来确定权重分配是否合理。计算一致性比例 CR，并对值进行判断：

$$CR = \frac{CI}{RI}$$

(6.8)

$$CI = \frac{\lambda - n}{n - 1}$$

(6.9)

其中，CI 为一致性指标：λ为最大特征根，n 为判断矩阵阶数，平均随机一致性指标 RI 值见表 6-3

表 6-3 平均随机一致性指标 RI 表
(1000 次正互反矩阵计算结果)

矩阵阶数	1	2	3	4	5	6	7	8
RI	0	0	0.52	0.89	1.12	1.26	1.38	1.41
矩阵阶数	9	10	11	12	13	14	15	
RI	1.46	1.49	1.52	1.54	1.56	1.58	1.59	

当 CR<0.1 时，认为判断矩阵的一致性在可接受范围内，CR>0.1 时，则判断矩阵不符合一致性要求，对该判断矩阵进行重新修正。

计算可得出： $CI_1=0.0975$ ， $CI_2=0.0675$ ， $CI_3=0.0294$ ， $CR_1=0.0871$ ， $CR_2=0.0603$ ， $CR_3=0.0262$ ，因为得出的一致性均小于 0.1，所以权重值合理。

根据上述计算，可以得到答复意见的相关性、完整性、可解释性的质量评价指标权重，表 6-4 所示。

表 6-4 答复意见质量评价指标权重

一级指标权重		
答复意见的相关性权重	表述的完整性	回答的可解释性
0.33	0.33	0.33
(二级) 答复内容的相关性权重		
答复内容的相关性指标	权重	
内容一致性	0.284589483	

语言规范性	0.068582292
亲和力	0.101342911
有说明性	0.25298748
可接受程度	0.292497834
(二级) 表述的完整性	
表述的完整性	权重
提及内容密切度	0.187101977
礼貌性回答	0.278524032
语言表达	0.343867947
相关数字特征	0.104197718
可接受程度	0.086308326
(二级) 回答的可解释性	
回答的可解释性	权重
专业性回答	0.16896604
问题针对性	0.072186454
语句相似度	0.109714428
理由充分程度	0.391685485
可接受程度	0.257447594

答复意见的相关性、完整性、可解释性评价得分和综合得分的计算公式为:

$$Y = f(x, k) = \sum_{i=1}^n x_i \times k \quad (6.10)$$

其中, x 为单项指标得分, k 为此项指标对应权重。

由此可知, 在答复内容的相关性方面, 答复意见的可解释程度与答复内容一致性对答复内容的相关性影响较大; 在表述的完整性方面, 答复意见的语言表达与礼貌性回答对答复内容表述的完整性评价影响较大; 在回答的可解释性方面, 答复意见的理由充分程度与意见可接受程度对回答的可解释性评价影响较大。

7 不足与展望

(1) 在关于留言内容的一级标签分类模型中, 由于文本是离散的, 而图像是属于连续空间的, 所以文本数据增强不同于图像领域的的数据增强, 在目前的的研究中, 文本数据增强存在一定的难度, 我们在利用回译法时可能会增大文本数据的噪音。如果可以增加机器学习的训练集样本量, 那么测试结果的置信度会大大提高。随着文本数据的不断增加, 对于文本分类的研究也变得更加重要, 通过 TF-IDF 算法、朴素贝叶斯模型可以精确地对文本进行分类, 这极大优化了人工分类的繁琐, 因此更加需要精确可靠的训练集样本和算法的支持。

(2) 本次选用的LDA模型偏向于主题词在文档中出现的频率, 其返回的结果是大量的词语。因此可能存在的问题是, 该模型检索出的是全部数据中的高频词, 其准确程度与数据清洗的结果直接相关。与此同时由于返回结果是一个个高频词语, 在把词语拼接为成句子时计算量会较大, 同时拼接出来的句子语义也会存在部分问题, 因此需要大量训练才能得出较为准确的结果。

(3) 在基于层次分析法的质量评价模型中, 因为留言和答复意见的数据较多, 每个市民对自己所得到的答复意见都有不同的评判标准, 在模型中就会存在定量数据较少, 定性成分多, 不易令人信服的问题; 如果指标过多时, 数据统计量大, 一般情况我们对层次分析法的两两比较是用 1 至 9 说明其相对重要性, 指标增多, 我们对两指标之间的重要程度的判断就可能会出现困难, 权重就会难以确定。

层次分析方法把研究的答复意见评价作为一个系统, 按照分解、比较判断、综合的思维方式进行决策。今后我们可以对市民所关心的问题分层和分类, 对他们所想要得到的意见答复定性分析, 把多目标、多准则又难以全部量化处理的决策问题化为多层次单目标问题, 有针对性的解答不同类型的留言问题并给予相关建议, 以便于快速有效提升政府管理水平。

参考文献

- [1]石凤贵.基于 TF-IDF 中文文本分类实现[A].安徽省教育厅科学研究项.2020.2
- [2]朱英龙.基于机器学习的文本分类[D].陕西.西安科技大学.2019
- [3]罗海飞,吴刚,杨金生.基于贝叶斯的文本分类方法[A].计算机工程与设计.Vol.27.
No.24.2006
- [4]黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[A].软件工程. Vol.273.
No.3.2020.03
- [5]杨潇,马军,杨同峰,杜言琦,邵海敏.基于主题模型 LDA 的多文档自动摘要[A]第五届全国信息检索学术会议论文集[C]上海.2009(11)
- [6]基于信息融合的多文档自动文摘技术[J]. 徐永东,徐志明,王晓龙. 计算机学报. 2007(11)

附录

答复意见质量评价指标权重计算结果：

	相关性	准则层特征向量
内容一致性特征向量	0.284589483	0.449028656
语言规范性特征向量	0.068582292	0.190517379
亲和力特征向量	0.101342911	0.360453965
有说明性语言特征向量	0.25298748	
可接受程度特征向量	0.292497834	
最大特征值 λ	5.39	
一致性指标CI	0.0975	
随机一次性指标RI	1.12	
一次性指标率CR	0.0871	

	完整性	准则层特征向量
提及内容密切度特征向量	0.187101977	0.449028656
礼貌性回答特征向量	0.278524032	0.190517379
语言表达特征向量	0.343867947	0.360453965
相关数字特征特征向量	0.104197718	
可接受程度特征向量	0.086308326	
最大特征值 λ	5.2687	
一致性指标CI	0.0675	
随机一次性指标RI	1.12	
一次性指标率CR	0.0603	

	可解释性	准则层特征向量
专业性回答特征向量	0.16896604	0.449028656
问题针对性特征向量	0.072186454	0.190517379
语句相似度特征向量	0.109714428	0.360453965
理由充分度特征特征向量	0.391685485	
可接受程度特征向量	0.257447594	
最大特征值 λ	5.1174	
一致性指标CI	0.0294	
随机一次性指标RI	1.12	
一次性指标率CR	0.0262	