题目：基于自然语言处理的“智慧政务”分析

2020 年 4 月

基于自然语言处理的“智慧政务”分析

摘要：随着我国网络问政平台的不断普及，传统问政方式无法满足于大幅增长的政务问答文本数据量。大数据、人工智能技术的发展能够为解决这一困难提供有效的帮助，本文基于自然语言处理（NLP）技术，解决政务留言类别划分、热点挖掘以及政务问答评价等一系列问题。具体过程如下：

首先，对群众留言分类文本进行数据清洗，中文分词以及去停用词等预处理，利用 TF-IDF 技术进行词向量转化。基于训练集（70%）数据使用交叉验证方法评估各种分类模型的平均 F-score，选取得分最高的支持向量机（F-score=0.90）进行训练，引入测试集（30%）数据进行预测，最终得到 F-score 为 0.91。

其次，对热点问题进行挖掘，采用命名实体识别技术基于正则表达式对留言主题的地点标签进行提取，并对留言较多的地点分别建立潜在语义分析（LSA）主题模型，提取主题关键词对全部留言进行匹配归类，并对同类热点问题进行总结，并对每个热点问题包含的留言个数进行归一化，得到热度指标与热点问题表。

最后，基于相关性、及时性、完整性与可解释性建立评价指标，提出政务答复意见的评价方法。

关键词：政务分析；TF-IDF；支持向量机；命名实体识别；LSA 主题模型

Analysis of "intelligent government" based on natural language processing

Abstract: Along with our country network asks the political platform unceasing popularization,The traditional way of asking politics cannot be satisfied with the large increase of the text data of government q&a.The development of big data and artificial intelligence technology can provide effective help to solve this difficulty. This paper, based on natural language processing (NLP) technology, solves a series of problems such as categorization of government comments, hot spot mining and q&a evaluation of government affairs.The specific process is as follows:

Firstly, the classified text of public comments was cleaned with data, Chinese word segmentation and stop words were preprocessed, and tf-idf technology was used for word vector transformation.Based on the training set (70%) data, the cross-validation method was used to evaluate the average f-score of various classification models. The support vector machine with

the highest score (f-score =0.90) was selected for training, and the test set (30%) data was introduced for prediction. The final f-score was 0.91.

Second, hot issues, excavation the named entity recognition technology based on regular expressions to extract of message theme the location of the label, and the message more sites respectively establish a model of latent semantic analysis (LSA) theme, extract topic keyword to match all the message category, and to summarize the similar issues, and each hot issue contains the message number is normalized, and hot issues of heat index table.

Finally, based on the relevance, timeliness, integrity and interpretability, the evaluation index is established, and the evaluation method of government response is put forward.

Key words: Government affairs analysis,TF-IDF, LSI, SVM,Identity of named entity

目 录

1 引 言

1.1 问题背景

2016 年，习近平总书记在网络安全和信息化工作座谈会上指出，要“分级分类推进智慧型城市建设”。近年来，我国的智慧城市建设进程逐步加快，各地政府都积极进行智慧政务政务平台建设，截至 2019 年 5 月，我国 36 个主要城市（直辖市、省会、副省级城市）中，已有 14 个城市基本建成政务大数据平台，大部分城市仍处于正在建设阶段[1]。现如今，传统人工政务问答方式已经不足以匹配大幅增长的政务问答数据量，结合大数据与人工智能技术的不断发展与创新，建设高效的智慧政务平台成为了各地政府在智慧城市建设进程中的主要任务之一。

目前，我国的问政方式仍存在一些不足之处。首先，在传统的网络问政方式中，问政平台工作人员需要人工对群众留言进行分类，并将留言分配到各相关部门，这种方式的工作量较大，并且分配精度也得不到有效保障。其次，传统的人工问答方式无法对特定地点或群体在某一时段所提出的相似问题进行有效归纳，这极大增加了相关工作人员的工作量，降低了工作效率。此外，当前传统政务问答平台中，难以对相关答复进行有效性评价，这也导致了群众对相关职能部门答复不满意的现象时有发生。

自然语言处理（以下称 NLP）技术能够为上述问题提供有效的解决方案。基于 NLP 技术所构建的文本分类模型，能够有效对群众留言进行类别划分，并准确分配到相关职能部门，在提高问题转送效率的同时也优化了分类准确率。而 NLP 中的文本聚类技术则能够有效地对相似问题进行归纳，便于工作人员制定统一答复，减轻不必要的重复性工作量。

综上所述，本文基于上述问题，利用词向量化（TF-IDF）、多种文本分类模型（Logistic 回归、朴素贝叶斯、随机森林、支持向量机）、文本语义索引（LSA 主题模型）等工具对政务留言文本数据进行分析，基于F1-Score、混淆矩阵等方式对模型进行选择与评价，得到较为理想的效果。

1.2 工作环境

本文基于 Windows10 操作系统（64 位）进行工作，GeForce940M 显卡搭配 16G 运行内存以及 1TB 机械硬盘，CPU 为 i5-4120M，以 Python3 为主要编程工具，Excel 为辅助工具进行操作。由于设备限制，本文所涉及的代码运行时间用于相同环境下的不同模型进行相对比较。

2 群众留言分类

2.1 分析框架

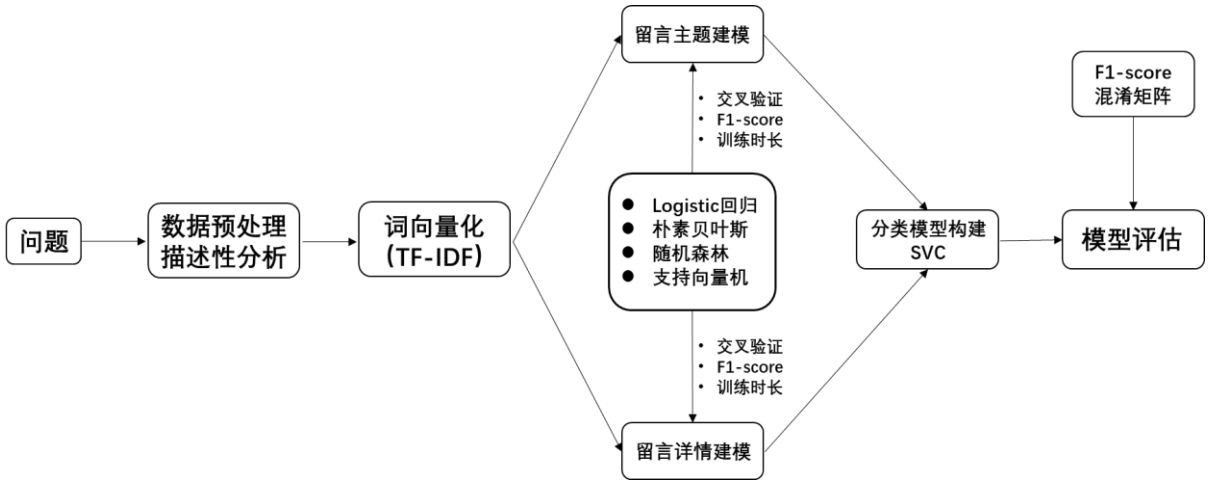

图 1 群众留言分类流程图

2.2 数据预处理与描述性分析

为了保证模型的高效与准确，需要在分析之前对文本数据进行预处理，将非结构化的中文文本转化为计算机可以识别的结构化数据。在进行模型构建之前，我们需要对原始数据进行整体的认知，为后续操作过程提供初步导向，减轻不必要的工作量。本过程主要内容为：数据预处理、描述性分析与词向量化。

2.2.1 数据预处理

首先，我们将附件 2 另存为 csv 文件并设置为‘UTF-8’编码格式，以便使用 Python 中的 pandas 库进行读取。数据预处理过程如下：

步骤一：去除冗余文本。通过对文本数据的初步浏览，我们发现“留言主题”、“留言详情”中存在一些与分类过程无关的信息，这些信息会对分类模型的构建产生一定程度上的影响，需要进行剔除。本文采用正则表达式对上述信息进行识别与剔除。具体信息如下表所示：

表 1 去除冗余文本示例表

示例文本	冗余信息	正则表达式
A3 区大道西行便道……强烈请求文明城市 A 市，尽快整改这个极不文明的路段。	“A3 区”、“A 市”等泛化地点名词	[A-Z][0-9]?[0-9]?[市|县|区]
我记得第一年（2013 年～2014 年）的费用是拖了两个月时间才付，……第三年（2015 年～2016 年）就至今没有付给我了（而且每次要提供相应金额的发票才能办付款手续）。	“2013” “2014”等数字	\d+

步骤二：中文分词。分词是自然语言分析的重要步骤，在英文中，词与词之间通常以空格连接，因此分词处理起来较为方便。对于中文的分词，主要有机械分词与基于统计（机器学习）分词等方法，其中，机械分词主要分为最大正向匹配、最大逆向匹配与双向匹配等，这些方法简单实用，但过多依赖于现有词典，难以处理未知词汇。基于统计（机器学习）分词方法主要为最大概率发于基于隐马尔科夫链（HMM）分词法，它们能够基于上下文环境进行分词，以优化分词效果。Python 中的 jieba 工具是当前最为流行的分词工具，它可以添加用户自定义词典，保证更高的准确率。

步骤三，去停用词。自然语言中存在一些常用的功能词，例如：语气助词、副词、介词以及连接词等等，这些词单独分开时并无实际意义，还可能会对分析结果造成影响，因此需要进行剔除。

以上步骤的效果如下所示：

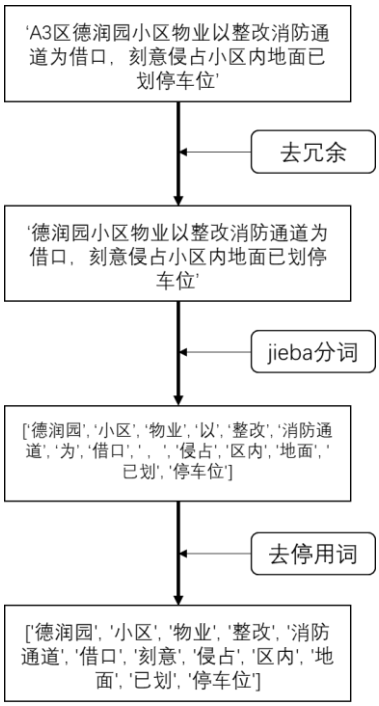

图 2 数据预处理示例

2.2.2 描述性分析

首先我们对分类标签进行统计，探究数据集中分类标签的比例，绘制如下饼状图：

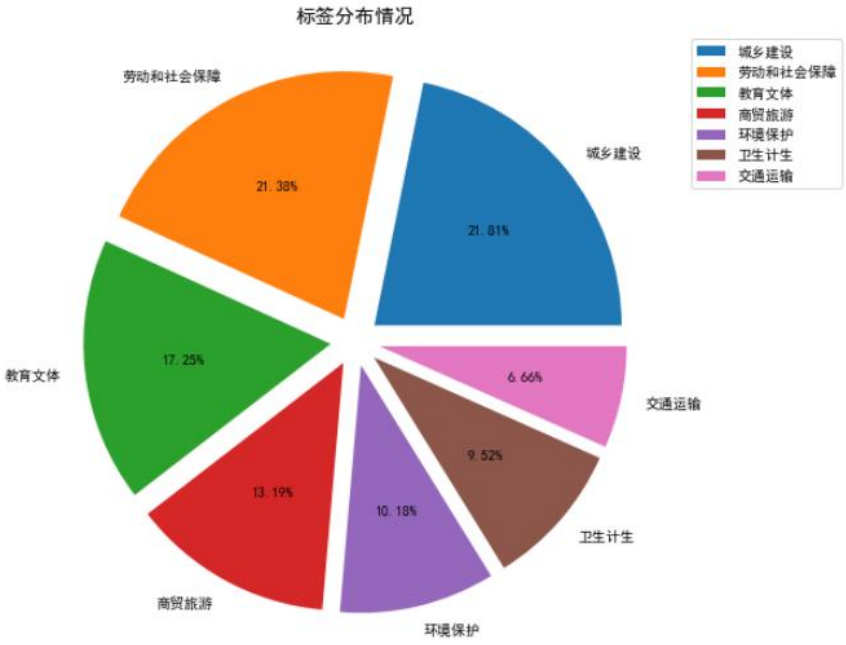

图 3 分类标签分布

可以看到，附件 2 中的留言分类标签共有 7 个，分别为城乡建设（21.81%）、劳动和社会保障（21.38%）、教育文体（17.25%）、商贸旅游（13.19%）、环境保护（10.18%）、卫生计生（9.25%）以及交通运输（6.66%），附件 1 中的一级分类标签并未全部出现，并且从比例来看，各分类标签的留言数量分布较为均匀，并未出现样本量极端不平衡的情况。

根据上文得到的分词结果，我们可以对数据做出进一步的描述性分析，对分词结果进行整体的认识。在附件 2 的数据中，留言主题与留言详情两部分数据问文本，我们分别对它们分词后的文本长度分布进行整体认识，使用 matplotlib 工具绘制分词长度条形图，结果如下所示：

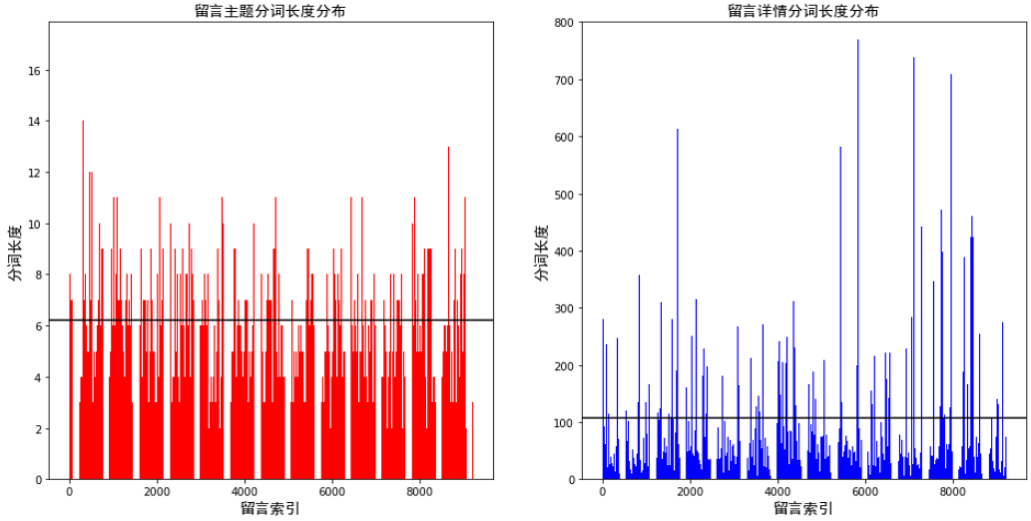

图 4 分词长度分布

可以看到，在一共 9210 条记录中，留言分词主题长度的分布较为均匀，其均值约为 6.2 个词。而留言详情分词长度的分布较为离散，其均值主要受较少部分长文本影响，约为 100 个词。

为了能够更直观地对出现较为频繁的词进行发掘，考虑到留言主题的文本较为精炼，我们绘制了关于留言主题的词云图片，具体效果如下：

图 5 “留言主题”词云图

由图可知在留言主题中，“西地省”、“小区”、“教师”、“公司”、“医院”等等词汇被提及次数较多，由此可见在群众反映问题主要为社区、企业、医院、教职工以及福利待遇等领域。

2.2.3 词向量化

上文所得分词文本并不是真正意义上的结构化数据，我们需要将分词结果进行数字化，将文本转化为计算机可识别的形式，因此需要对文本进行向量化表示。在进行词向量化之前，我们首先以 7:3 的比例将数据集随机划分为训练集与测试集，并基于训练集的词典进行向量化。

自然语言处理中，One-Hot 和 TF-IDF 是目前最为常见的用于提取文本特征的方法。其中，One-Hot 将词汇用二进制表示并构成稀疏向量来进行表示，词袋大小即为向量维度，所有维度中只有一个元素为 1，它并不能很好地体现出文本中具有代表性关键词的信息。

TF-IDF（Term Frequency–Inverse Document Frequency）是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。它主要分为两个部分：词频（TF）与逆概率（IDF）。

词频（TF）主要反映词的出现频率，计算公式如下：

$$TF = \frac{N}{M}$$

其中，N 表示一个词在某文档中出现的频次，M 则表示该文档。

逆概率（IDF）主要衡量一个词在某文档中的重要性（权重），计算公式如下：

$$IDF = log(\frac{D}{D_w + 1})$$

其中，D 为文档总个数，D_w为出现词 w 的文档个数，IDF 取值大小与该词的常见程度成反比。

最终，TF-IDF 值的计算公式如下：

$$TF - IDF = TF \times IDF$$

它的主要原则为：在词频（TF）统计的基础上，所有文本中越常见的词赋予越小的权重（逆概率 IDF）。

在本分类问题中，TF-IDF 方法较为适用。其原因主要为：在群众留言中，关键词与类别标签的联系较为紧密，例如：文本中频繁出现“排污”一词，那么该文本很大可能属于“环境保护”类别，同理，“医疗”、“药品”等词语与“卫生计生”这一类别联系较为紧密；反之，“请问”、“解决”等词在各留言中较为常见，但它们对分类过程并不能起到有效的

帮助作用。TF-IDF 方法能够很好地处理关键词信息，因此，本文选择该方法，以 7:3 的比例将数据集划分为训练集与测试集，使用 sklean 中的 CountVectorizer 与 TfidfTransformer 工具，基于训练集词典分别对训练集、测试集进行词向量化。

2.3 分类模型训练与分析

2.3.1 模型选择

首先，我们需要确定模型的选择标准。

本文中，本文主要基于交叉验证方法结合 F-score 对模型进行评估。交叉验证的基本思想是把在某种意义下将原始数据进行分组，一部分做为训练集，另一部分做为验证集，首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型以此来做为评价分类器的性能指标。F-score 的计算公式如下所示：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i})$$

其中，P_i为第 i 类的查准率, R_i为第 i 类的查全率。

针对“留言主题”、“留言详情”文本数据向量化，结合 10 折交叉验证建立不同分类模型（随机森林、多分类 Logistic 回归、朴素贝叶斯与支持向量机），提高分类的预测准确性同时顾及分类的灵敏度。我们得到的模型结果表明，针对“留言详情”数据使用支持向量机模型分类效果最佳。具体内容如下：

1、随机森林

在决策树算法中，分类树被用于预测定性变量。对于分类树而言，给定观测值被预测为它所属区域内的训练集中最常出现的类。普通决策树通常存在着高方差的劣势，这意味着如果将训练集随机分成两半，对这两个子集分别建立回归树，可能得到截然不同的两棵树。为了减少这种统计学习方法的方差从而增加预测准确性，有一种很自然的方法：从总体中抽取多个训练集，对每个训练集分别建立预测模型，再对由此得到的多个预测值求平均。这便是装袋法的主要思想。而随机森林方法通过在每个分裂点所用的预测变量作抽样来对书作去相关，实现对装袋法的改进。

2、多分类 Logistic 回归模型

决策树在解释性方面强于线性回归方法，遗憾的是树的预测准确性一般无法达到其他回归和分类方法的水平。接下来我们先考虑回归方法再考虑其他分类方法。

3、朴素贝叶斯

和决策树模型相比，朴素贝叶斯分类器(Naive Bayes Classifier 或 NBC)发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC 模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，当属性之间相互独立，NBC 模型与其他分类方法相比具有最小的误差率。

4、支持向量机

通常来说，如果数据可以被超平面分割开，那么事实上存在无数个这样的超平面，因为可以在不接触这些数据点的情况下对这些超平面进行上下移动或是旋转。一个很自然的方法是找到最大间隔超平面使训练观测到分割超平面的隔间达到最大。最大间隔超平面通常由观测的一个小子集决定，我们把这部分观测成为支持向量。这也是支持向量分类器的主要思想。支持向量分类器产生的是线性分类边界，对于多分类问题准确率不太理想。因此我们考虑支持向量机方法。支持向量机是支持向量分类器的一个扩展，扩展的结果是支持向量机使用了一种特殊的方式，即核函数来扩大特征空间，能够产生非线性分类边界。

针对“留言主题”数据各分类方法的比较，从下图中可以发现，支持向量机模型在模型得分以及模型数据处理效率方面均优于其他几个分类模型，交叉验证平均得分达到了0.86。

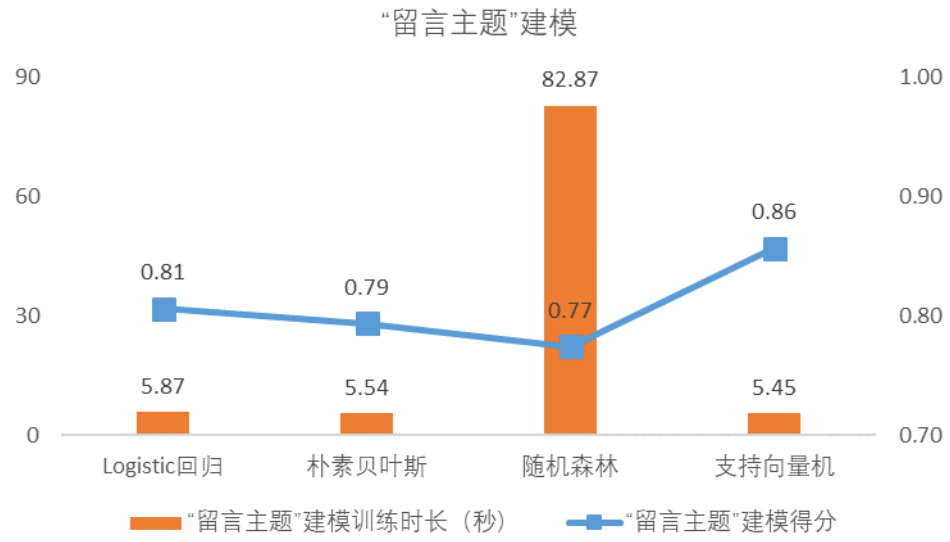

图 6 “留言主题”建模比较

针对“留言详情”数据（X_tr.csv）各分类方法进行比较，绘制图 8：

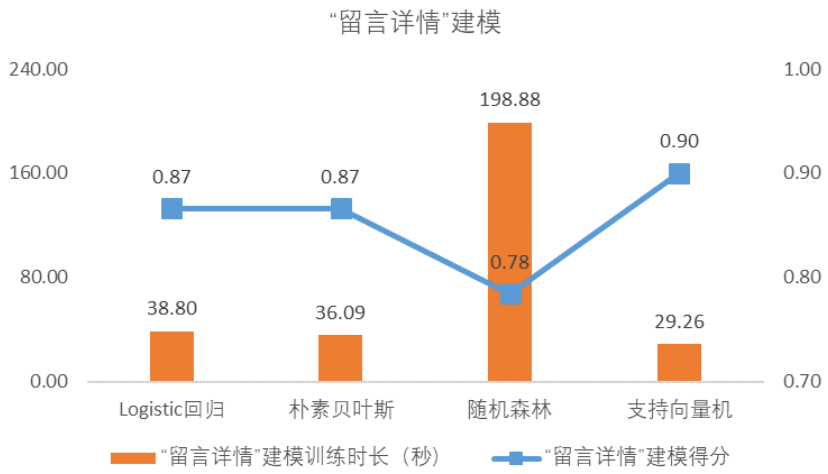

图 8 “留言分词”建模比较

从图 8 中可以发现，基于“留言详情”进行建模，虽然运行时间增加，各个模型的准确率均有所提高，其中支持向量机模型的得分达到了 0.9 以上，并且运行时间依旧是所有模型中最短的。

基于上述比较，本文选择支持向量机模型基于“留言详情”词向量化后的数据进行建模分析。

2.3.2 支持向量机模型构建与评估

基于上文分析，本文采用“留言详情”进行支持向量机分类器训练，并对测试集进行预测与评估，最终得到 F-score 约为 0.9052，结果较为理想。为了验证模型对各类别的预测效果，我们对每类别的精确率（precision）、召回率（Recall）以及 F-score 进行统计，得到如下结果：

	precision	recall	f1-score	support
0	0.94	0.93	0.94	461
1	0.89	0.80	0.84	183
2	0.94	0.95	0.94	601
3	0.88	0.83	0.86	364
4	0.93	0.91	0.92	251
5	0.93	0.92	0.93	289
6	0.85	0.91	0.88	614
accuracy			0.91	2763
macro avg	0.91	0.89	0.90	2763
weighted avg	0.91	0.91	0.91	2763

图 9 各类别评估指标统计

由图 9 可以看出，模型对各类别标签预测的 F-score 均达到 0.84 以上，并未出现极端情况，说明模型不管从整体角度还是具体类别，都能达到理想的效果。

为了更加直观地展现预测效果，我们绘制了预测精度混淆矩阵如图 10 所示：

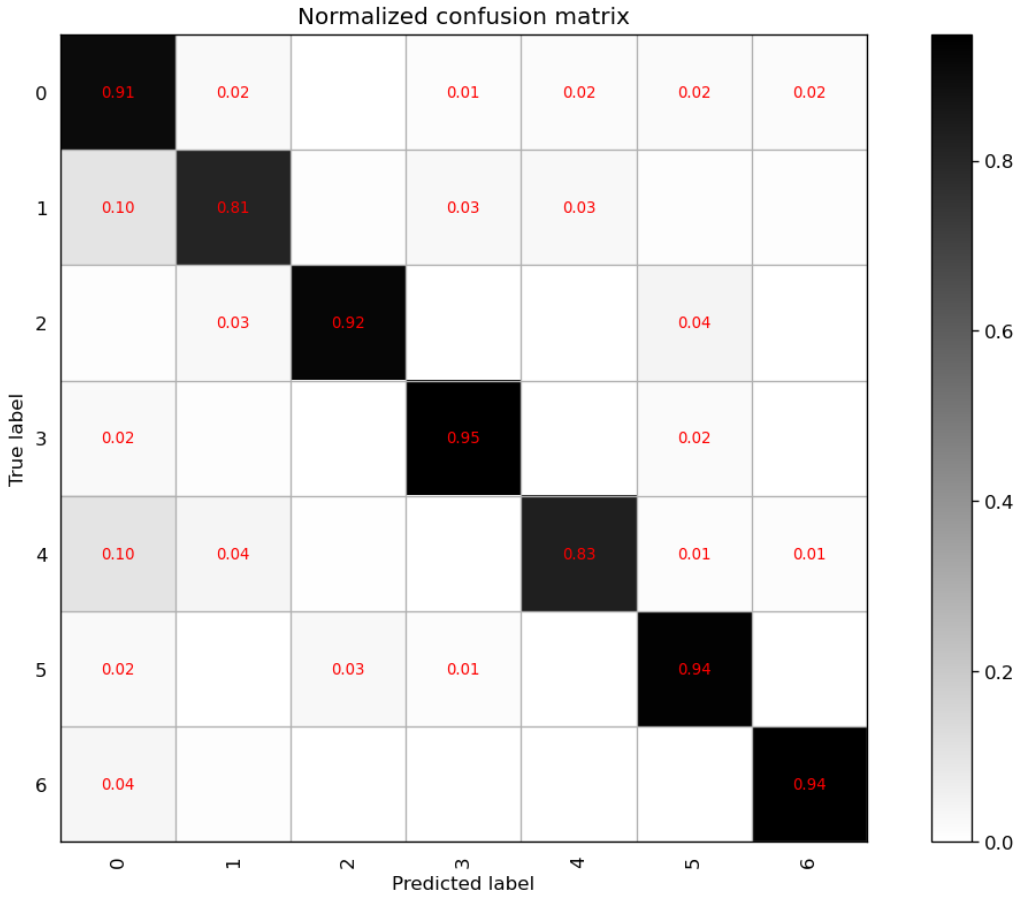

图 10 基于测试集预测混淆矩阵

从生成的混淆矩阵图 10 中我们可以发现，应用支持向量机模型分类的总体预测准确度较好的同时，分类灵敏度也表现出色。

3 热点问题挖掘

3.1.1 热点问题挖掘简介

“热点问题”在人们日常生活中，往往被理解为造成一定社会影响，并受较广大群众关注的新闻事件或者社会焦点问题。在本文中，结合题目实际情况，我们将它定义为某一时段内群众集中反映的某一问题。根据定义，我们将一个热点问题分为三个主要的要素：地点、反映的对象（人群）、造成的影响。

在本题中，我们需要最终给出排名前 5 的“热点问题表”，同时给出“热点问题留言明细表”——列出每个热点对应的留言编号内容。因此，首先我们需要定义合理的热度评价指标，考虑到热点的热度评价指标总是与反映该热点的样本量成正比关系，所以我们将热点问题归类的过程转化为根据热点问题的要素特征，匹配样本信息，得到该类样本量的过程。

根据解题思路，我们再进一步利用自然语言处理工具和数据挖掘算法优化过程，得到了热点问题挖掘模型。

3.1.2 热点问题挖掘模型概况

为了热点问题挖掘模型能够正确的提取出热点，并且最终能输出“热点问题表”、“热点问题留言明细表”，我们提出了一种基于命名实体识别和 LSA 主题模型的热点问题挖掘模型。该模型的核心思想是提取出热点主题词匹配样本留言归类成热点，主要包括四个步骤：数据预处理、留言样本集进行初步分类、提取并确定热点（主题词）、匹配留言样本。

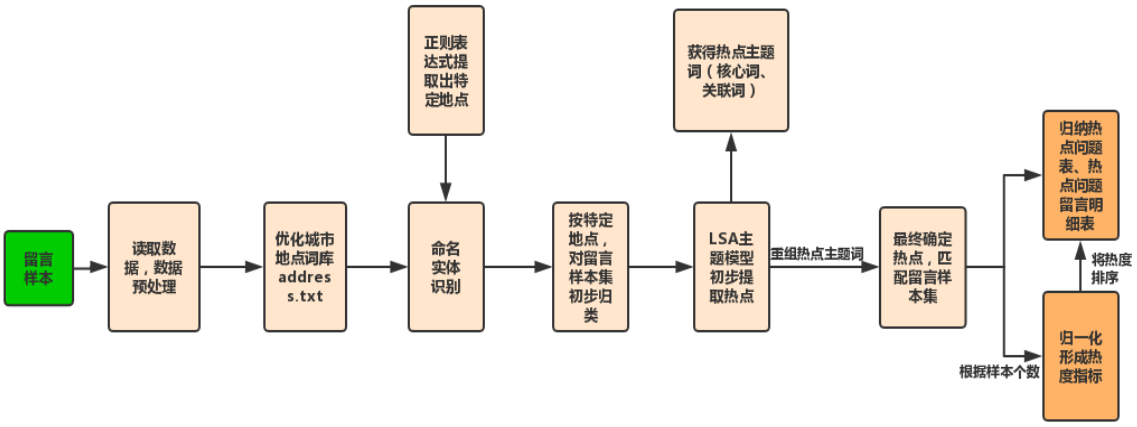

图 11 热点问题挖掘流程

第一步：数据预处理。对问题给出的数据集，首先我们需要进行思考它需要怎样的处理，得到我们能够进行下一步算法分析的数据，好的数据预处理往往能改进数据的质量,从而有助于提高其后的挖掘过程的准确率和效率。在数据预处理过程中，我们优化分词词库，得到我们想要的分词结果。

第二步：留言样本集进行初步分类。由于热点问题挖掘难度较大，很难直接从大量的数据集中挖掘得到较理想的结果，我们考虑将留言样本分块处理。我们观察到留言样

本的地点命名特殊性，利用命名实体识别去提取留言样本的特定地点，将这些特定地点作为依据，对留言样本集初步归类。

第三步：提取并确定热点（主题词）。我们在进行对已经分类的数据集做初步提取热点。这里，我们比较了 LSA 主题模型与 LDA 主题模型的优劣，最终选择了 LSA 主题。提取的热点主要由主题词构成，我们依据留言样本能提取出一个核心词和诸多关联词之一来归类于某一热点问题的原则，对主题词进行重组，确定最终的热点主题。（我们将主题词分为核心词和关联词：核心词是每一个主题必不可少的词，一个主题只有一个核心词；关联词是每一个主题中可或缺的词，它与）

第四步：匹配留言样本。在已知分类的结果下，通过提取出的每个热点主题，去匹配类中的留言样本，便可以比较精确的得到归于同一个热点的留言了。我们将基于热点的热度评价指标总是与反映该热点的样本量成正比这一关系，通过归一化样本量得到热度评级指标，并将热点问题按热度降序，最终，我们会通过程序输出“热点问题表”、“热点问题留言明细表”。

3.2 模型解答实际过程

3.2.1 数据预处理过程

在实际的数据预处理过程中，我们将读取的数据依此进行缺失数据处理（采用过滤掉缺失数据的办法）、去重、去除无用字符等操作。初步处理后我们得到 4326 条有效样本。

之后我们需要对留样样本数据的“留言主题”、“留言详情”进行 jieba 分词处理。jieba 分词在中文文本处理中不仅有很多强大的功能，而且还有精度高、效率快的优点。在为了使得分词效果良好，我们通过爬虫技术，从网页上了爬取了题目中出现的相关地点词库加入 jieba 分词词库中，得到了优化后的城市地点词库。这样得到的分词结果有我们需要的一些特定地点名出现，以便于之后获得热点的地点要素。

经过了分词、去停用词、分词重组连接等步骤，最终我们得到了 4326 条留言样本详细分词结果。数据预处理结果如表 2 所示：

表 2 数据预览

	留言编号	留言主题	留言时间	留言详情	反对数	点赞数	留言主题（分词）	留言详情（分词）	留言主题（分词连接）	留言详情（分词连接）
0	188006	一米阳光婚纱艺术摄影是否合法纳税了	2019/2/28 11:25:05	座落在联丰路米兰春天G栋，一家名叫一米阳光婚纱艺术摄影的影楼，据说年单这一个工作室营业额就上...	0	0	[一米阳光, 婚纱, 艺术摄影, 合法, 纳税]	[座落在, 联丰路, 米兰春天, G, 栋, 一家, 名叫, 一米阳光, 婚纱, 艺术摄影,...	一米阳光 婚纱 艺术摄影 合法 纳税	座落在 联丰路 米兰春天 G 栋 一家 名叫 一米阳光 婚纱 艺术摄影 影楼 年单 工作室 ...
1	188007	咨询道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	道路命名规划已经初步成果公示文件，什么时候能转化成为正式的成果，希望能加快完成的路名规范，给...	0	1	[道路, 命名, 规划, 初步, 成果, 公示, 城乡, 门牌]	[道路, 命名, 规划, 初步, 成果, 公示, 文件, 转化, 正式, 成果, 希望, 加...	道路 命名 规划 初步 成果 公示 城乡 门牌	道路 命名 规划 初步 成果 公示 文件 转化 正式 成果 希望 加快 路名 规范 道路 安...
2	188031	反映春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	本人系春华镇金鼎村七里组村民，不知是否有相关水泥路到户政策和自来水到户政策，如政府主导投资村...	0	1	[春华镇, 金鼎村, 水泥路, 自来水, 到户]	[系, 春华镇, 金鼎村, 七里, 组, 村民, 不知, 相关, 水泥路, 到户, 政策, ...	春华镇 金鼎村 水泥路 自来水 到户	系 春华镇 金鼎村 七里 组 村民 不知 相关 水泥路 到户 政策 自来水 到户 政策 政府...
3	188039	黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	靠近黄兴路步行街，城南路街道、大古道巷、一步两搭桥小区（停车场东面围墙外），第一单元一住户卫...	0	1	[黄兴路, 步行街, 大古道巷, 住户, 卫生间, 粪便, 外排]	[靠近, 黄兴路, 步行街, 城南路, 街道, 大古道巷, 一步, 两, 搭桥, 停车场, ...	黄兴路 步行街 大古道巷 住户 卫生间 粪便 外排	靠近 黄兴路 步行街 城南路 街道 大古道巷 一步 两 搭桥 停车场 东面 围墙 外 第一 ...
4	188059	中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	中海国际社区三期四期中间，即蓝天璞和洲幼儿园旁边那块空地一直处于三不管状态，物业不管城管不管...	0	0	[中海国际社区三期, 四期, 空地, 夜间, 施工, 噪音, 扰民]	[中海国际社区三期, 四期, 蓝天, 璞, 洲, 幼儿园, 旁边, 块, 空地, 处于, 三...	中海国际社区三期 四期 空地 夜间 施工 噪音 扰民	中海国际社区三期 四期 蓝天 璞 洲 幼儿园 旁边 块 空地 处于 三不管 状态 物业 城管...

3.2.2 留言样本集进行初步分类过程

热点问题挖掘在多样本的情况下难度很大，不仅是因为样本情况的复杂度，还考虑到文本数据挖掘的高维性。因此，我们对数据集做初步分类处理来达到降低挖掘热点难度的目的。在方法的选择中，我们考虑利用命名实体识别的方法提取“留言主题”中的特定地点来作为分类的依据。

命名实体识别（Named Entity Recognition，简称 NER）又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。在本题中，我们观察到留言样本里基本都会出现诸如“A 市”、“A7 县”、“A4 区”此类的特定地点，因此我们选择提取留言样本中的特定地点。基于正则表达式有强大能力提取特定格式的文本内容，我们采用正则表达式去提取特定地点。

“留言主题”内容相对“留言详情”内容呈现出精简的特点，便于提取特定信息，且在实际解题过程中确实效果更佳。在这一系列操作之后，得到的结果如下表 3 所示：

表 3 提取特定地点后留言数据表

	留言编号	留言主题	留言时间	留言详情	反对数	点赞数	地点标签	留言主题（分词）	留言详情（分词）	留言主题（分词连接）	留言详情（分词连接）
0	188006	一米阳光婚纱艺术摄影是否合法纳税了	2019/2/28 11:25:05	座落在联丰路米兰春天G栋，一家名叫一米阳光婚纱艺术摄影的影楼，据说年单这一个工作室营业额就上...	0	0	A3区	[一米阳光, 婚纱, 艺术摄影, 合法, 纳税]	[座落在, 联丰路, 米兰春天, G, 栋, 一家, 名叫, 一米阳光, 婚纱, 艺术摄影,...	一米阳光 婚纱 艺术摄影 合法 纳税	座落在 联丰路 米兰春天 G 栋 一家 名叫 一米阳光 婚纱 艺术摄影 影楼 年单 工作室 ...
1	188007	咨询道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	道路命名规划已经初步成果公示文件，什么时候能转化成为正式的成果，希望能加快完成的路名规范，给...	0	1	A6区	[道路, 命名, 规划, 初步, 成果, 公示, 城乡, 门牌]	[道路, 命名, 规划, 初步, 成果, 公示, 文件, 转化, 正式, 成果, 希望, 加...	道路 命名 规划 初步 成果 公示 城乡 门牌	道路 命名 规划 初步 成果 公示 文件 转化 正式 成果 希望 加快 路名 规范 道路 安...
2	188031	反映春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	本人系春华镇金鼎村七里组村民，不知是否有相关水泥路到户政策和自来水到户政策，如政府主导投资村...	0	1	A7县	[春华镇, 金鼎村, 水泥路, 自来水, 到户]	[系, 春华镇, 金鼎村, 七里, 组, 村民, 不知, 相关, 水泥路, 到户, 政策, ...	春华镇 金鼎村 水泥路 自来水 到户	系 春华镇 金鼎村 七里 组 村民 不知 相关 水泥路 到户 政策 自来水 到户 政策 政府...
3	188039	黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	靠近黄兴路步行街，城南路街道、大古道巷、一步两搭桥小区（停车场东面围墙外），第一单元一住户卫...	0	1	A2区	[黄兴路, 步行街, 大古道巷, 住户, 卫生间, 粪便, 外排]	[靠近, 黄兴路, 步行街, 城南路, 街道, 大古道巷, 一步, 两, 搭桥, 停车场, ...	黄兴路 步行街 大古道巷 住户 卫生间 粪便 外排	靠近 黄兴路 步行街 城南路 街道 大古道巷 一步 两 搭桥 停车场 东面 围墙 外 第一 ...
4	188059	中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	中海国际社区三期四期中间，即蓝天璞和洲幼儿园旁边那块空地一直处于三不管状态，物业不管城管不管...	0	0	A3区	[中海国际社区三期, 四期, 空地, 夜间, 施工, 噪音, 扰民]	[中海国际社区三期, 四期, 蓝天, 璞, 洲, 幼儿园, 旁边, 块, 空地, 处于, 三...	中海国际社区 三期 四期 空地 夜间 施工 噪音 扰民	中海国际社区三期 四期 蓝天 璞 洲 幼儿园 旁边 块 空地 处于 三不管 状态 物业 城管...

上表中，地点标签一栏就是我们所提取出的特定地点，一共有 39 个特定地点。我们将对同一地点在留言样本的“留言主题”中出现，且次数大于 40 的直接归为一类，否则，归入其他类。例如，留言样本的“留言主题”中出现“A3 区”432 次，则我们将这 432 个样本归入“address_A3”类中；再如留言样本的“留言主题”中出现“B 市”5 次，则将这 5 个样本归入“others”类中。最终我们得到“address_A”、“address_A1”、“address_A2”、“address_A3”、“address_A4”、“address_A5”、“address_A6”、“address_A7”、“address_A8”、“address_A9”、“xds”、“others”共十二类数据用于接下来的分析。

3.2.3 提取并确定热点（主题词）过程

接下来，我们在进行对已经分类的留言样本数据集提取热点。在这一实际过程中，我们主要比较了 LSA 主题模型与 LDA 主题模型哪一模型更适合对“留言主题（分词）”提取热点主题。

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，在业界也被称为三层贝叶斯概率模型，这里的三层结构主要包括词、主题和文档。而该模型之所以被称为一个生成模型，我们认为一段文字中的每个词都是可以通过“这段文字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。

LSA(Latent semantic analysis)同样也是主题模型。最初用来解决一词多义、一义多词

的问题。LSA 的基本思想就是把高维的文档降到低维空间，那个空间被称为潜在语义空间。这个映射必须是严格线性的，并且是基于共现表的奇异值分解。LSA 方法可以很好的得到热点主题，但是 LSA 缺点就是目前没有很好的统计解释。

最终我们选择了 LSA 主题模型，发现它更适应于短文本热点主题提取。提取的热点主题主要由主题词构成，我们将主题词分为核心词和关联词：核心词是每一个主题必不可少的词，一个主题只有一个核心词；关联词是每一个主题中可或缺的词，一个主题至少有一个关联词。如果某个留言样本的“留言主题”中能提取出一个核心词和该核心词相匹配的诸多关联词之一，就将该样本归类于某一热点问题。

但是 LSA 主题模型得到的热点主题还需要加工处理，将热点主题词重组，得到我们想要的热点主题。在重组过程中，我们选择将核心词在原始数据中匹配，再在此中寻找关联词，这一过程较为繁琐。最终我们得到的主题有 125 个，部分热点主题词结果如下表 4 所示：

表 4 热点主题词表（部分）

地点标签	主题	核心词	关联词 1	关联词 2	关联词 3
address_A3	address_A3_1	西湖	茶场	五组	
address_A3	address_A3_2	扰民	夜间	施工	
address_A3	address_A3_3	兰亭湾畔	违规	餐厅	
address_A3	address_A3_4	梅溪湖	润芳园	油烟	
address_A3	address_A3_5	西海岸	保利	幼儿园	
address_A3	address_A3_6	青青家园	兰亭湾畔	幼儿园	
address_A3	address_A3_7	楚仪	车辆	停	
address_A3	address_A3_8	惟盛园	安置小区	街道	
address_A3	address_A3_9	盛世耀凯	杜鹃路	渣土	

得到热点主题词之后，我们会用到这些主题词去匹配留言样本信息，将留言样本归类于这些热点。

3.2.4 匹配留言样本过程

匹配过程中的方法是将一个主题中的核心词和关联词在留言样本的“留言主题”中遍历，如果某个留言样本的“留言主题”出现这个核心词和与它相关的关联词之一，就将这个样本归类于这一热点主题中，通过这样的方式我们可以得到将留言样本归类于热点的基本正确结果。

但是 125 个热点中有一些热点的主题词极为接近，我们需要合并一些类似的热点留言样本编号，并删去一些热点。详情见表 5：

表 5 热点留言编号详情表（部分）

	地点标签	主题	核心词	关联词1	关联词2	关联词3	留言编号列表	留言个数
45	address_A	address_A_1	滨河	景园	车位	nan	[264944, 225217, 258242, 244512, 235362, 25978...	39
35	address_A2	address_A2_1	丽发新城	搅拌站	扰民	nan	[264944, 225217, 244512, 258242, 235362, 25978...	35
47	address_A	address_A_3	购房	人才	补贴	nan	[189180, 199435, 203760, 205771, 212128, 21902...	16
85	address_A5	address_A5_1	魅力之城	油烟	夜宵摊	扰民	[360101, 360100, 236798, 360102, 205168, 18938...	13
49	address_A	address_A_5	经开区	规划	公园	nan	[226408, 233542, 238692, 239670, 256358, 26162...	11

我们发现得到的结果效果很理想，在留言样本的留言主题中可以归入“滨河+景园 or 车位”热点的留言样本达到了 39 个，是所有热点中热度最高的。

对于热点的热度评价指标，我们是基于热点的热度评价指标与反映该热点的样本量成正比这一关系，通过归一化样本量的方法得到热度评级指标，并将热点问题按热度评级指标降序，标上热度排名，再描述出每个热点的时间范围、地点/人群、问题描述，详情见表 6。

表 6 热点问题表

主题	热度排名	问题 ID	热度指标	时间范围	地点/人群	问题描述
address_A2_1	1	1	1	2019/11/2 至 2020/1/25	A2 区丽发新城小区	小区附近修建搅拌厂噪音扰民、灰尘污染
address_A_1	2	2	0.897435897	2019/7/7 至 2019/9/1	A 市伊景园滨河苑	强制捆绑销售车位
address_A_3	3	3	0.41025641	2018/11/15 至 2019/12/2	A 市引进人才	人才购房补贴申请不成功
address_A_5	4	4	0.282051282	2019/1/2 至 2019/8/26	A 市经开区	区内若干项目规划优化咨询建议
address_A3_1	5	5	0.230769231	2019/1/9 至 2019/9/12	A3 区西湖街道茶场村	拆迁规划咨询

最终，通过编写程序输出《热点问题留言明细表（初步）》（表 7），供我们查找高热度热点中的留言样本详情。

表 7 热点问题留言明细表（部分）

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
address_A_1	264944	A0004260	A2 区丽发新城附近修建搅拌厂噪音、灰尘污染	2019/11/2 14:23	A 市 A2 区丽发新城小区附近，作为长株潭绿心地带，还是人口众多的小区，竟然堂而皇之修建搅拌厂，请问环保部门、城建部门为什么可以审批通过，如何获得施工手续的？值得深思，请政府重视！	0	0
address_A_1	225217	A909223	A2 区丽发新城附近修建搅拌厂严重影响睡眠	2019/11/15 9:17	我已经好久没睡过安稳觉了,A 市暮云街道丽发新城小区开发商在小区附近建一搅拌站,每天尘土飞扬,噪音嗡嗡嗡嗡嗡,晚上睡不着,白天开车在路上很容易出问题的,请关注!	0	0
address_A_1	258242	A909220	A 市暮云街道丽发新城社区搅拌站灰尘，噪音污染严重	2019/12/2 12:23	A 市暮云街道丽发新城小区附近的搅拌站灰尘、噪音污染严重，严重影响附近居民休息，使其不能以最佳状态投入到建设祖国的伟大事业中去。	0	0

最后将表 6 与表 7 分别整理成题目要求格式（“热点问题表.xls”与“热点问题留言明细表.xls”）。

4 答复意见评价方案

4.1 答复意见评价问题简介

智慧政务中提高对于人们提问的答复意见质量是智慧政务的一大关键，同时，评价答复意见又是一大难题，本题第三问需要我们从答复的相关性、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。对于答复的不同性质角度我们将找出不同的指标去量化描述，构建评价指标体系，建立答复意见综合评价模型，最终以综合得分的形式对答复意见质量给出综合评价。

评价指标体系是指由表征评价对象各方面特性及其相互联系的多个指标，所构成的具有内在结构的有机整体。它具有系统性原则，及各指标之间要有一定的逻辑关系，各指标之间相互独立，又彼此联系，共同构成一个有机统一体；具有典型性原则，及务必确保评价指标具有一定的典型代表性，尽可能准确反映出特定答复意见的综合特征；具有可量化原则，指标选择上，特别注意在总体范围内的一致性，指标选取的计算量度和计算方法必须一致统一。具有侧重性原则，评价指标体系的设置、权重在各指标问的分配及评价标准的划分都应该与答复意见某个侧重点的重要程度相适应。

基于以上指标选择的要求，我们对于答复意见的性质角度做出删选，将选择及时性指标、相关性指标、完整性指标、可解释性指标去评价答复意见。为了得到综合评价得分，我们会对各个指标的重要性作出衡量，赋予不同的权重。对于一个答复意见样本，量化出各个指标的得分，带入综合评价模型中，即得综合得分。

4.2 指标选择

4.2.1 相关性指标

相关性是评价答复质量的一个关键点，这一性质可以避免答非所问（问与答主题词不匹配、问与答总体内容不匹配）的情况。因此，我们将相关性指标选入指标体系中。

在量化相关性指标的过程中，我们选用文本匹配工具，包括模糊匹配特征、关键词匹配特征等。

文本匹配是自然语言处理中的一个核心问题，很多自然语言处理的任务都可以抽象成文本匹配问题，例如信息检索可以归结成查询项和文档的匹配，问答系统可以归结为问题和候选答案的匹配，对话系统可以归结为对话和回复的匹配。针对不同的任务选取合适的匹配模型，提高匹配的准确率成为自然语言处理任务的重要挑战。

对于文本匹配，我们采取综合考虑模糊匹配特征与关键词匹配特征的方法。首先进

行模糊匹配特征，我们将对于问答主体内容全部提取特征，计算出文本相似度；其次进行关键词匹配特征，我们对于提问的文本内容用 CRF 工具提取特定的地点、任务等主题词，通过在答复中遍历，查看主题词存在个数确定关键词匹配特征得分。通过以上双匹配结果，综合考量问答相关性，计算出问答相关性匹配得分用于描述相关性指标。

$$Cor_i = \frac{\text{模糊匹配特征得分} + \text{关键词匹配得分}}{2}$$

4.2.2 及时性指标

及时性是反应答复意见效率的一个很好的性质，答复意见的质量在很大方面通过答复的快慢来体现，答复越快就越能体现智慧政务系统的工作效率高。我们经过考虑后，将及时性纳入评价指标体系。

将及时性选入评价指标体系后，我们将量化及时性，通过及时性指标计算得分公式来计算每个样本的及时性指标分数：

$$Time_i = \frac{t_{i,answer} - t_{i,question}}{\max(t_{i,answer} - t_{i,question})}$$

$t_{i,\ answer}$表示的是每个样本的回答时间，$t_{i,\ question}$表示的是每个样本的提问时间。这一计算公式呈现出计算效率快、统一得分的优点。观察数据后，我们发现对于每一个留言样本，并无缺失现象，我们都可以得到有效的及时性得分。

4.2.3 完整性指标

完整性是评价答复质量不可或缺的组成部分，这一指标用于衡量政务回复过程中针对提问者所提的问题回答的完整性情况，是有效改进回复遗漏、回复残缺情况的重要依据。

在量化完整性指标的过程中，主要做两方面的考虑，一是对回复内容的长度的量化处理及量化得分，二是同一提问者问题个数与回复中的分点个数的匹配情况的量化处理。

回复内容长度的量化得分，采用归一化方法，先对语句长度进行量化处理，再将所得到的数值进行标准化处理，所得到的值作为该语句长度的量化得分。对于同一提问者问题个数与回复中的分点个数的匹配情况，先对问题语句中问题个数做量化提取，再对回复内容中分点个数做量化提取，后者与前者差值为非负则认为回复完整，差值为负则认为回复不完整，差值绝对值越大回复越不完整。以差值的绝对值进行标准化处理作

为问题个数与回复中的分点个数的匹配情况的得分。

对回复内容长度的量化得分及问题个数与回复中的分点个数的匹配情况的得分进行赋权，各自赋予 0.5 的权重作为完整性指标的最后得分。

4.2.4 可解释性指标

首先，我们对答复意见的可解释性进行分析，我们将可解释性定义为答复意见对留言详情的解释程度，亦即“答”是否为“所问”。对于这个问题，我们选择“留言详情”“答复意见”标签匹配的方法。步骤如下：

步骤一，“留言主题”与“答复意见”标签预测。在第一题中，我们所构建的留言分类模型 F-score 达到 0.9 以上（基于一级标签），预测效果较为理想，因此，我们基于该模型分别对“留言详情”、“答复意见”进行类别预测。

步骤二，预测标签匹配。对上一步的预测结果进行匹配，若同一条留言中，“留言主题”与“答复意见”的预测标签一致，则判定为“可解释”，赋值为 1，否则赋值为 0，得到可解释性指标。

$$\mathit{I}\text{nterpretative}_{\text{i}} = \begin{cases} 0, y_{1i} \neq y_{2i} \\ 1, y_{1i} = y_{2i} \end{cases}$$

其中，y_{1i}、y_{2i} 分别表示“留言主题”与“答复意见”的预测标签。

此外，在数据允许的情况下，可以对三级标签进行建模分类，这样得到的效果会更加理想。

4.3 建立答复意见综合评价模型

完成确立相关性指标、及时性指标、完整性指标、可解释性指标量化公式之后，模型需要将这四个一级指标按重要性程度赋予权重。我们采取专家调查权重法。

专家调查权重法是一个较科学合理的方法，依据"德尔菲法"的基本原理，选择社会各方面的专家、学者或政务工作人员，采取独立填表选取权数的形式，然后将他们各自选取的权数进行整理和统计分析，最后确定出各因素，各指标的权数。集合了各专家、学者、政务人员的智慧和意见，并运用数理统计的方法进行检验和修正。

在经过该方法调查后，我们小组最终采用将相关性指标和可解释性指标赋予 0.3 权重，及时性和完整性赋予 0.2 权重。及建立答复意见评价指标模型：

$$\mathit{A}\text{nswer-}\mathit{S}\text{core}_{\text{i}} = 0.3\mathit{C}\text{or}_{\text{i}} + 0.2\mathit{T}\text{ime}_{\text{i}} + 0.2\mathit{C}\text{ompleteness}_{\text{i}} + 0.3\mathit{I}\text{nterpretative}_{\text{i}}$$

参考文献

[1] 2019-2025 年中国智慧政府行业深度调研与发展趋势研究报告

[2]杜鹏.国内政府智慧服务模型构建研究[J].经济研究导刊,2020(07):174-177.

[3]石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[4]火善栋.用 AdaBooster 算法实现中文文本分类问题[J].现代计算机(专业版),2016(30):3-6.

[5]许高建,胡学钢,王庆人.文本挖掘中的中文分词算法研究及实现[J].计算机技术与发展,2007(12):122-124+172.

[6]杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术与信息化,2020(03):146-148.

[7]CSDN《使用 python 绘制混淆矩阵》https://blog.csdn.net/qq_36982160/article/details/80038380

[8]曹春萍,崔海船.基于 LSA 和结构特性的微博话题检测[J].计算机应用研究,2015,32(09):2720-2723.