“智慧政务”中的文本挖掘应用

摘 要

自 2019 年底，新冠疫情开始蔓延，全世界都在与新冠疫情作斗争，全国人民积极抗疫。在我国疫情常态化防控的背景下，我国游客对于短程旅游的关注度上升，使得本地周边游更受人们欢迎。

本文利用赛题所提供数据，从大量的文本之中进行旅游要素抽取，挖掘相关数据，计算旅游产品的相关热度，寻找旅游产品之间的关联，分析当地周边游产品的需求变化，并撰写信件，向旅游行业发展提供建议。

针对问题一，使用爬虫技术，进行微信公众号文章爬取，共计 3162 篇文章，按照 9:1 的比例划分为训练集、验证集。使用 BERT+DNN 构建文本分类模型，并进行评估，结果取三次平均，Loss=0.0519，Accuracy=0.9790。将文章标题，与正文拼接生成新文本，使用文本分类模型，进行文章分类。考虑到文章标题的概括性以及其可能更具代表性，在拼接生成新文本时，其系数设置为 2。

针对问题二，使用百度飞桨深度学习平台，从飞桨提供的 66 种实体类型中选取 9 个相关性较高的实体类型，筛选过滤无关实体类型后，进行命名实体识别，提取出本文所需实体，从而实现旅游产品提取；对于产品热度的计算，本文统计旅游产品出现频次，结合产品发表时间，采用指数方法，进行产品热度计算，依照产品发表年度进行旅游产品热度排序，并对周边游产品进行热度分析。

针对问题三，使用 Apriori 算法寻找数据之间的关联度，进行产品强关联规则的挖掘。本文使用置信度，支持度两个指标计算关联度 Conviction 参数，对问题二得到的旅游产品，利用关联度 Conviction 参数进行关联分析，并对旅游产品的关联进行可视化操作。

针对问题四，由于 OTA、UGC 数据内容分散及碎片化，故本文利用挖掘赛题提供数据得到的相关有用信息，详细分析在新冠疫情前后广东茂名市的旅游产品的变化情况，并结合对茂名当地的旅游产品变化的分析结果，向有关部门提出旅游行业发展的建议。

关键词:文本分类; 命名实体识别; 产品热度分析; 产品关联

目录

1. 挖掘目标

1.1. 目标概述

本次建模的目标是利用收集自互联网公开来源的群众问政留言记录，以及相关部门对部分群众留言的答复意见，通过 jieba 中文分词工具对留言主题进行分词、卷积神经网络方法等方法，达到以下三个目标：

1) 利用数据字典和卷积神经网络方法对非结构化的数据进行文本挖掘，根据已知数据结果，结合网络论政平台的留言划分标准，建立关于留言的一级分类模型。
2) 根据收集的群众留言数据，在合理的热度评价指标下，对某一特定的的地点或特定的人群的留言进行归类后进行热度评价。统计相关数据，分析当下热点问题，了解群众所需，提升相关部门的服务效率。
3) 根据相关部门的留言答复意见，借助服务质量评价模型与理论，建立答复质量评价模型，进行留言答复意见质量的实证研究，寻找提升解决民生问题的服务水平与质量的路径与策略，

1.2. 数据条件

附件 1 给出了一种三级标签体系，包含了 15 类一级标签，102 类二级标签以及 390 类三级标签；附件 2,3,4 给出了收集自互联网公开来源的群众问政留言记录，包含了留言用户的 ID、留言主题、留言详情、留言时间等主要信息；其中附件 2 还包括了留言信息的分类标签、附件 3 给出了该条留言信息的点赞数和反对数、附件 4 给出了有关部门对留言信息的回复内容。

2. 分析方法与过程

本用例包括如下步骤：

步骤一：数据预处理，在题目所给出的数据中，出现了很多用户发布的留言诉求，在问题处理的基础上，在原始的数据上进行所需单元的选择和留言数据得去空去重等处理方式，在处理后的文件基础上进行模型的建立以及相应数据的分析。

步骤二：数据分析，在对处理好的文件做相应的问题处理，具体问题具体分析，找到问题关键，获取问题所需。

步骤三：数据筛选，统计相关数据，分类筛选汇总，总结问题结果

步骤四：得出问题结论，给出问题改进的意见和建议。

2.1. 问题一分析方法与过程

2.1.1. 流程图

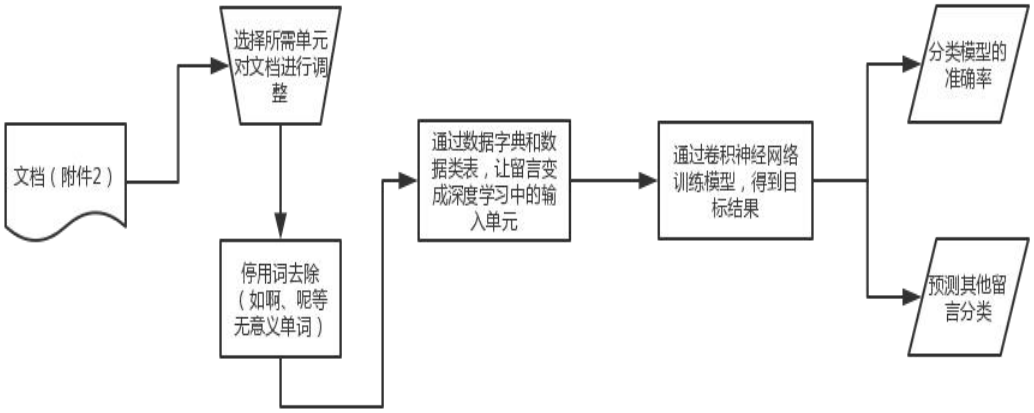


图 1 问题一处理流程图

2.1.2. 数据预处理

2.1.2.1. 数据字典

数据字典是一个集合，其中包括键值对，通过键索引然后关联到相对的值。数据字典的实现实际是基于哈希算法原理。

哈希表是根据关键值进而直接进行访问的一种数据结构。通过把键和值映射到表中某个位置来访问记录，这种方式使得查询速度与更新速度非常快，这个映射函数就叫做哈希函数，把存放值的数组称为哈希表。

2.1.3. 卷积神经网络[1, 2]

卷积神经网络主要由卷积层、池化层以及全连接层组成。卷积层用来实现提取特征，池化层则压缩结果、保留重要的特征，在全连接层将文本数据映射到样本标记空间。每一层网络有多个神经元结构，上一层的神经元通过激活函数映射到下一层神经元，每个神经元之间有相对应的权值，输出即作为我们的分类类别。卷积神经网络模型基本架构如图 2 所示。

2.1.3.1. 卷积层

卷积层是卷积神经网络的核心内容，卷积操作其实就是卷积核矩阵和对应的输入层中一小块矩阵的点积相乘，卷积核通过权重共享的方式，按照步幅上下左右的在输入层滑动提取特征，以此将输入层做特征映射作为输出层。每一个卷积核可以看做一个特征提取器，不同的卷积核就负责提取不同的特征，在句子里面提取有利于分类的特征需要从词语或者级别去提取，也就是卷积宽口的宽度应该覆盖完全单个词向量，也就是说卷积神经网络的卷积核宽度必须要等于词向量的维度。

2.1.3.2. 池化层

池化是一种非线性采样形式，主要作用是通过减少网络的参数来减小计算

量，并且能够在一定程度上控制过拟合[3]。文本处理中的池化层一般采用最大池化，它将卷积层的每一个通道得到的向量进行最大池化，得到一个标量，因此可以看到简单的卷积网络只能提取到某个句子之中是否有某个向量 n-a ，它并不能得到这个 n-a 在语句中出现的位置，更不能提取到此 a 和其他 a 之间的关联依赖关系。这样，卷积核有多少个就会有多少个最大标量，然后将这些标量拼接起来成一个向量；所有大小的卷积核所得到的向量再次拼接起来，得到一个最终的一维向量，并将最终的向量传到全连接层或 softmax 层进行分类。

2.1.3.3. 全连接层

卷积层和池化层将文本数据映射到特征空间，在全连接层将卷积层和池化层处理后得到的局部特征通过权值矩阵进行整合处理，并映射到样本标记空间。即全连接层作为“分类器”在卷积神经网络中承担重要作用。

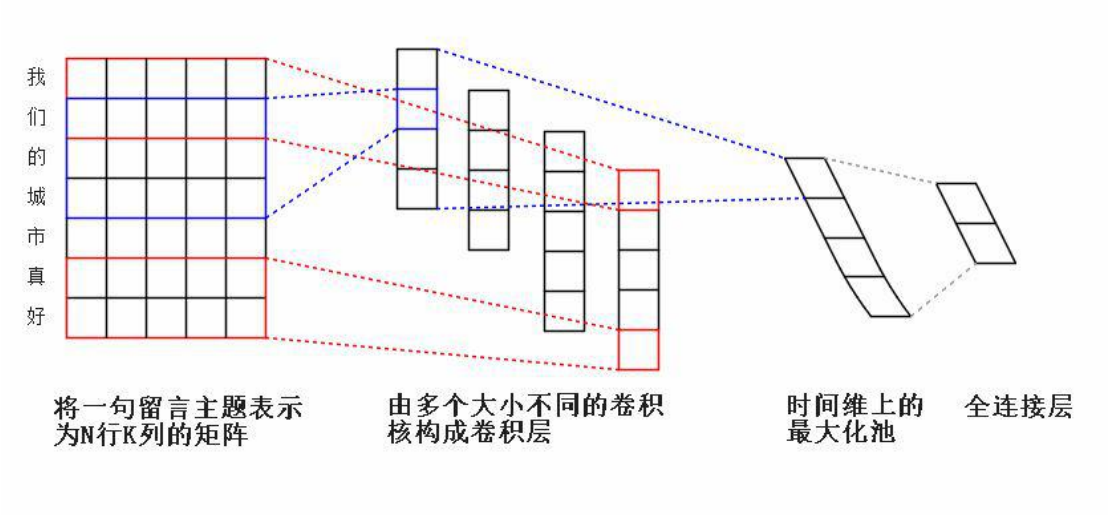


图 2 卷积神经网络模型结构图

2.1.4. 分类模型评价

2.1.4.1. 损失函数

损失函数是估量模型的预测值与真实值的不一致程度的函数，它是非负实值函数，损失函数越小，模型的鲁棒性就越好[5]。

对于本题我们采用交叉熵损失函数[4,5]来进行平均损失的计算，该函数计算方式如下：

$$L=\frac{1}{N}\sum_{i}L_{\mathrm{i}}=\frac{1}{N}\sum_{i}-\sum_{c=1}^{M}y_{ic}\log\left(p_{ic}\right)$$

其中：

M——类别的数量；

y_{ic}——指示变量（0 或 1），如果该类别和样本 i 的类别相同就是 1，否则是 0；

p_{ic}——对于观测样本 i 属于类别 c 的预测概率。

2.2. 问题二分析方法与过程

2.2.1. 流程图

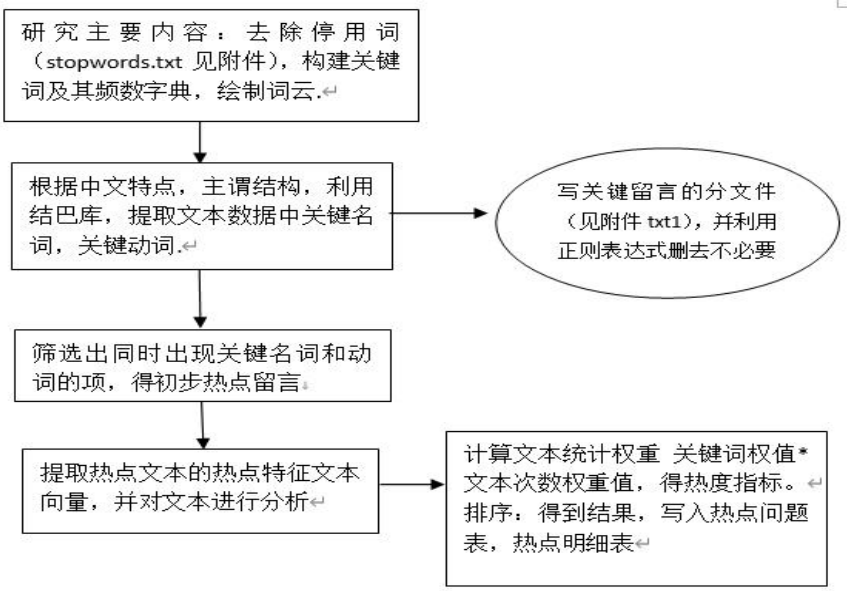


图 3 问题二处理流程图

2.2.2. 数据预处理

2.2.2.1. 对留言信息进行去重、合并

在题目给出的数据中，出现了很多重复的留言数据。例如同一用户利用相似留言主题、相似留言内容在一段时间内重复留言，同时这种留言的点赞数会分散。考虑到这样的重复留言会影响到热点问题的判断，因此需要去掉重复记录，并且合并重复记录的点赞数量。同时，在留言主题描述为空的记录，干扰了问题的分析，直接采用滤过方法，从文本中删除。

2.2.2.2. 对留言信息进行中文分词

在对留言信息进行数据挖掘分析之前，先要把非结构的文本信息转换为计算机能够识别的结构化信息。在附件二中以中文文本的方式给出了数据。为了方便转换，先要对这些留言数据信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。Jieba 分词是基于前缀词基于前缀词典实现词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），采用动态规划查找最大概率路径，找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。

2.2.2.3. TF-IDF 算法

在对留言信息进行分词后，需要把文中的词语变成向量以供之后我们进行文本处理、数据分析所使用。这里采用 TF-IDF 算法，将留言内的关键信息转换为有权重的向量，TF-IDF 的具体步骤如下：

1) 计算词频：词频 = 某个词在文章中出现的总次数

为了消除不同留言大小之间的差异，便于不同留言之间的比较，我们在此标

准化词频:词频 = 某个词在文章中出现的总次数/文章的总词数

或：词频 = 某个词在文章中出现的总次数/文章中出现次数最多的词个数

2） 计算逆文档频率

在此，首先需要一个语料库来模拟语言的使用环境。

逆文档频率（IDF） = log（词料库的文档总数/包含该词的文档数+1）

为了避免分母为 0，所以在分母上加 1.

3） 计算 TF-IDF 值

基于之前的分析了解，有：TF-IDF 值 = TF * IDF。

在此有：TF-IDF 值与该词的出现频率成正比，与在整个语料库中的出现次数成反比，符合之前的分析。

```
[[0.0032989  0.00659779 0.02639116 0.00659779 0.0032989  0.00659779
  0.0032989  0.0032989  0.0032989  0.03958674 0.0032989  0.02199263
  0.0032989  0.01099632 0.0032989  0.0032989  0.0032989  0.0032989
  0.02199263 0.01099632 0.01099632 0.01099632 0.01099632 0.02199263
  0.01099632 0.01099632 0.03298895 0.04398527 0.01099632 0.01099632
  0.01099632 0.02199263 0.01099632 0.07697422 0.02199263 0.01099632
  0.01099632 0.01099632 0.01099632 0.01099632 0.08797053 0.01099632
  0.01099632 0.01099632 0.01099632 0.01099632 0.05278232 0.01099632
  0.01319558 0.03958674 0.10556464 0.01099632 0.02199263 0.00989669
  0.02199263 0.0032989  0.0032989  0.01099632 0.01099632 0.12095949
  0.07917348 0.01099632 0.01319558 0.01099632 0.01099632 0.01099632
  0.02199263 0.01099632 0.01099632 0.01319558 0.01319558 0.21112928
```

图 4 语料库次数比

4） 求出关键字

计算出留言中每个词的 TF-IDF 值之后，进行排序，选取其中值最高的几个作为关键字。

5） 计算文章的相似性

计算出每段留言的关键词，从中各选取相同个数的关键词，合并成一个集合，计算每段留言对于这个集合中的词的词频，生成每段留言各自的词频向量，进而通过欧氏距离或余弦距离求出两个向量的余弦相似度，值越大就表示越相似。

2.2.3. 提出热度评价假设

本题在现有的研究基础上，根据实际生活，对群众留言热度评价中的影响因素做出相应的研究假设。

2.2.3.1. 信息源对留言热度评价影响的假设[6]

多数的热度评价都是从事件发生的主客体及作用力三个角度各自包含的影响因素进行评价。但忽视了事件问题自身所具有的特征。如事件本身所具有的危害度。因此，本题在原有的研究基础上，增加了话题事件自身对问题热度评价的影响因素。

基于此做出如下假设：

H1：话题敏感度对留言问题热度评价存在的影响；

H2：事件所属类型对留言问题热度评价存在的影响

2.2.4. 热度评价构建

现主要研究留言内容：去除停用词（stopwords.txt 见附件），构建关键词及其频数字典，绘制词云（大致了解分词统计结果如图 5 所示），根据这些研究内容构建热度评价模型。

图 5 关键词分布图

2.3. 问题三分析方法与过程

2.3.1. 流程图

图 6 问题三分析流程图

2.3.2. 数据预处理

2.3.2.1. 评价数据处理

附件中，添加了本次预处理的.py 文件，我们采用对停用词表的循环处理方式，将文本中的空格以及***等脱敏符号之类的特殊符号进行转换，使其具有可读性。再读区与导出 csv 格式文件时，采用“utf-8-sig”的格式，将文本中出现的乱码\ufeff 去除，以免影响对数据的分析。最后使用.replace()将读取值中的\n, \t, \xa0, \u3000 等 csv 格式乱码去除。

2.3.3. 答复意见质量评价模型设计

留言答复意见多为网络在线的方式，对其质量的评价既要考虑到服务对象-公众的期望与实际满意度，也要结合在线方式对答复意见质量的支撑作用。因此基于传统的服务质量评价模型和信息答复的特征，进行答复意见质量评价模型。

2.3.3.1. 答复意见质量评价模型构建的依据[10]

各个行业和领域对质量评价的研究增多。芬兰 Gronroos 教授提出“顾客感知服务质量”的概念，就是说服务质量是通过顾客的主观感知来进行评价的[11]。则留言的答复意见质量是由群众的主观感知来进行评价。基于 SERVQUAL 模型对服务质量评价进行定量计算。形成了包含感知性、可靠性、响应性、保证性、以及移情性五个维度下二十二项指标的模型量表。本题在此基础上灵活的调整维度和指标，进行答复意见质量评价模型的设计。

2.3.3.2. 答复意见质量评价模型的设计[10]

考虑到留言答复的回复者是有关部门的工作人员，其公信力保证了信息来源的权威性以及回复的内容是真实可靠的；同时考虑到留言之间的时间差，在线平台系统等硬件或软件的服务有形性对答复意见质量评价影响甚微，可忽略。所以本题建立的答复意见质量评价模型在原模型下修改为相关性、完整性、可解释性这 3 个指标。模型如图 7 所示。

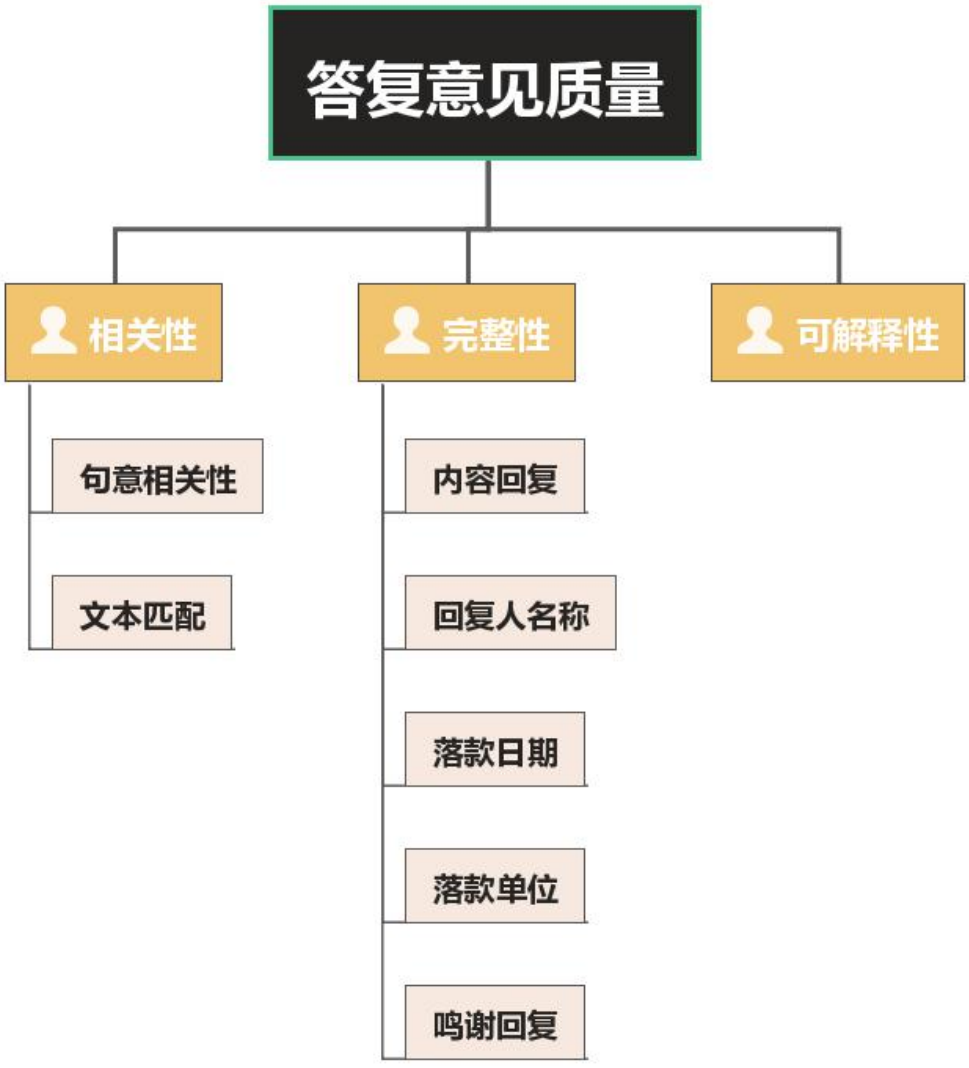


图 7 答复意见质量评价模型

1) 相关性

在相关项指标中，我们可以构建两类模型：一类是有正误数据标签数据的可训练的模型；一类是使用已训练好词向量的模型做匹配分析。

其中，**第一类模型**是基于 BERT 的句意相关性分析。具体做法可以如 https://github.com/lonePatient/bert-sentence-similarity-pytorch 中所示。

但是我们认为有一些更好的办法来判断句子的相似程度，先将文本中长度不一的句子，统一映射到相同维数，即句子标准化。如果直接生成文本词频向量，用词频来代替句子，这样做会产生近义词信息、语义信息、大量文本下运算等诸多问题。如果两段很长的文本进行比较（比如上万字的文章），岂不是维度要扩增很多倍？而且矩阵会非常稀疏，就是很多取值都是 0，计算开销大且效率低。

对 UCI-News Aggregator Dataset 数据集的分析如图 8 所示。

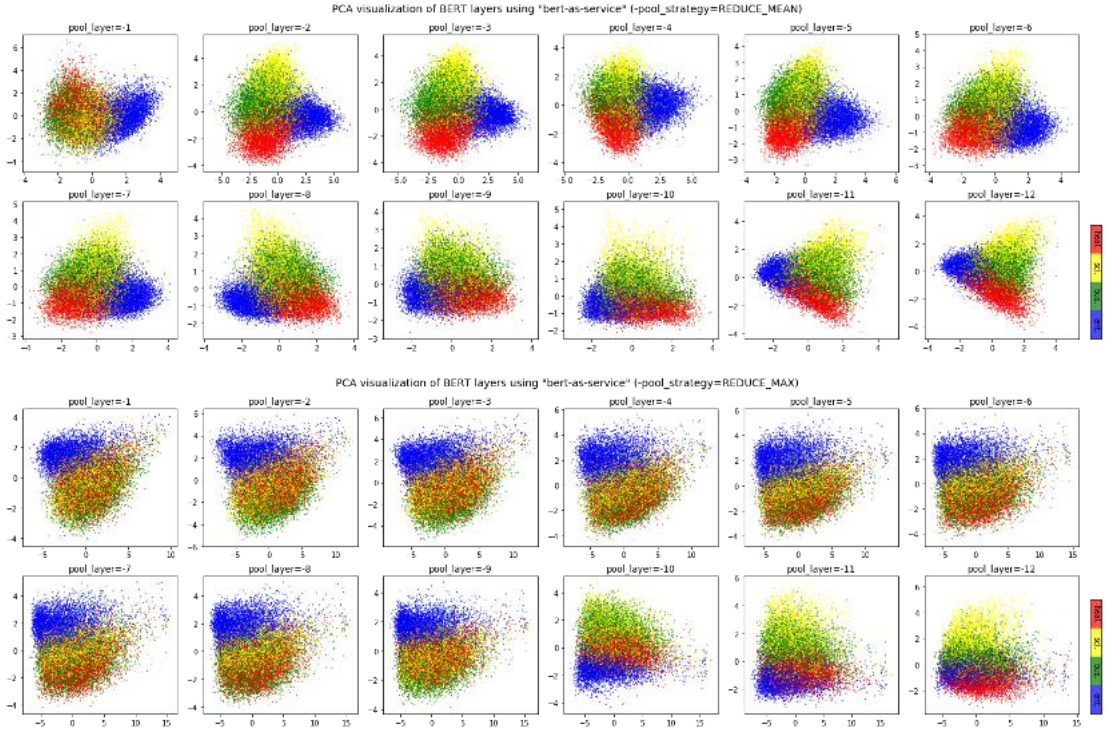


图 8 UCI-News Aggregator Dataset 数据集结果

直观上，pooling_layer = -1 接近训练输出，因此它可能会偏向训练目标。如果不对模型进行微调，则可能会导致不良表示。 pooling_layer = -12 接近词向量嵌入，可以保留非常原始的单词信息（没有花哨的添加等）。 另一方面，仅使用词嵌入即可达到相同的性能。 就是说，在[-1,-12]之间的任何事物都是一个权衡。

接下来，在利用相同向量文本的优势，用 BERT 的方法，将二者的近似度求出来。最基础的方法是 NNLM 模型，在開始之前，引入模型复杂度，定义例如以下：

$$O = E * T * Q$$

当中，E 表示训练的次数，T 表示训练语料中词的个数，Q 因模型而异。E 值不是我们关心的内容，T 与训练语料有关，其值越大模型就越准确。

在 NNLM 模型中，从隐含层到输出层的计算时主要影响训练效率的地方，CBOW 和 Skip-gram 模型考虑去掉隐含层。实践证明新训练的词向量的准确度可能不如

NNLM 模型（具有隐含层），但能够通过添加训练语料的方法来完好。

Word2vec 包括两种训练模型。各自是 CBOW 和 Skip_gram(输入层、发射层、输出层)，例如图 9 所示：

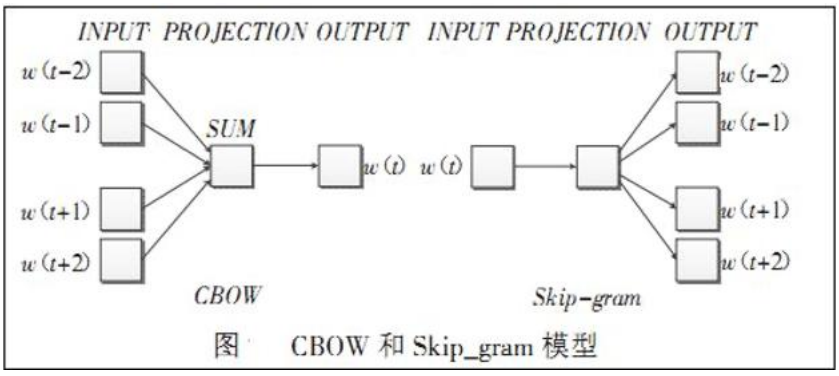


图 9 Word2vec 两种训练模型

此时，使用 Skip-gram 模型同意某些词被跳过，因此可组成 3 元词组。由上述表达可知：一方面 Skip-gram 反映了句子的真实意思，在新组成的这 18 个 3 元词组中，有 8 个词组可以正确反映例句中的真实意思；还有一方面，扩大了语料。3 元词组由原来的 4 个扩展到了 18 个。语料的扩展可以提高训练的精确度。获得的词向量更能反映真实的文本含义。

最后在 word2vec 中提供了 distance 求词的 cosine 相似度。并排序。也能够在训练时，设置-classes 参数来指定聚类的簇个数。使用 k-means 进行聚类。

第二类方法是：使用，基于 LDA 模型的文本匹配。通过文本之间主题分布的距离来评估文本之间的语义相似性。利用这种语义的相似性筛选出答复价值质量不高的答复信息。这种语义相似性可以进一步作为机器学习模型的特征。

测量两种局部分布的距离度量包括 Hellinger 距离（HD）和 Jensen-Shannon 发散（JSD）。Hellinger 距离的正式定义如下：

$$HD(P,Q)=\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{K}\left(\sqrt{pi}-\sqrt{qi}\right)^2}$$

在这里，pi 和 qi 是相应分布的第一个元素。Jensen-Shannon 发散的定义如下：

$$JSD(P\|Q)=\frac{1}{2}\left(KLD(P\|M)+KLD(Q\|M)\right)$$

$$M=\frac{1}{2}(P+Q)$$

$$KLD(P\|M)=\sum_{i=1}^{K}\mathrm{pi}\ln\frac{pi}{mi}$$

其中 KLD 代表 Kullback-Leibler 发散。

2) 完整性

在完整性的指标中，我们确定了几个抬头。例如，回复里开头的问候语，以及是否包含留言网友的昵称，“您的留言已收悉。现将有关情况回如下：“，以及回复里结尾是否有“感谢您对我们工作的支持、理解与监督！”、以及回复出自某个部门的部门落款以及日期落款，之类的字眼，包含了就是这些表达，我们认为具有框架完整性。

更重要的是，是否有对评论的回复，在 1)中的相似度处理中，有些留言只有开头和结尾，没有对留言内容的具体回复，我们认为这样的回复是残缺的。在指标中，我们定义内容如图 10 所示。

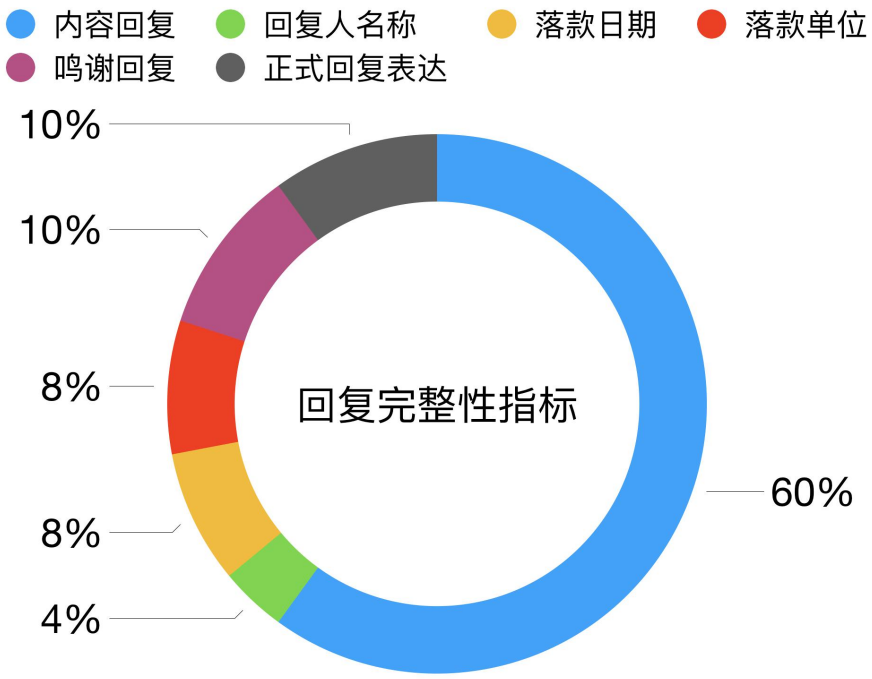


图 10 完整性指标内容比率图

基于我们给出的评价指标，我们量化我们的评价分级如表 1 所示：

表 1 评级分级总评表

不完整	<60
基本完整	>60
比较完整	>72
完整	>82
非常完整	>90

*只要正则式符合相关特征条件，即认为满足所对应的完整性条件

*评价指标总打分为 100 分

3) 可解释性

我们可以利用在问题 2 中的方法，提取每条留言以及对应回复，在 tf-idf 下的特征关键词提取。在对应的每条回复——留言中，将属于分类的特征词一一对

应。正式定义如下：

$$Weight\_h = \frac{\text{当前词匹配数}}{\text{匹配词总个数}}$$

$$CRI = \frac{\text{留言中的特征个数}}{\text{改良后的总解释个数}} = \frac{\text{留言中被解释的个数}+\text{已匹配答复中的解释个数}}{\text{原有的特征个数}+\text{已匹配个数}}$$

并给出我们认为的评价指标如表 2 所示：

表 2 可解释性指标内容

不完整	<0.5
基本完整	0.5-0.6
比较完整	0.6-0.7
完整	0.7-0.85
非常完整	>0.85

*评价指标总打分为 1 分

3. 结果分析

3.1. 问题一结果分析

3.1.1. 数据集

本题实验数据来源于附件 2，其中包括了城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生等 7 个类别，共计 9210 个文本数据，通过数据预处理形成数据列表和数据字典。详细的样本类别以及数据的统计信息如图 11 所示。

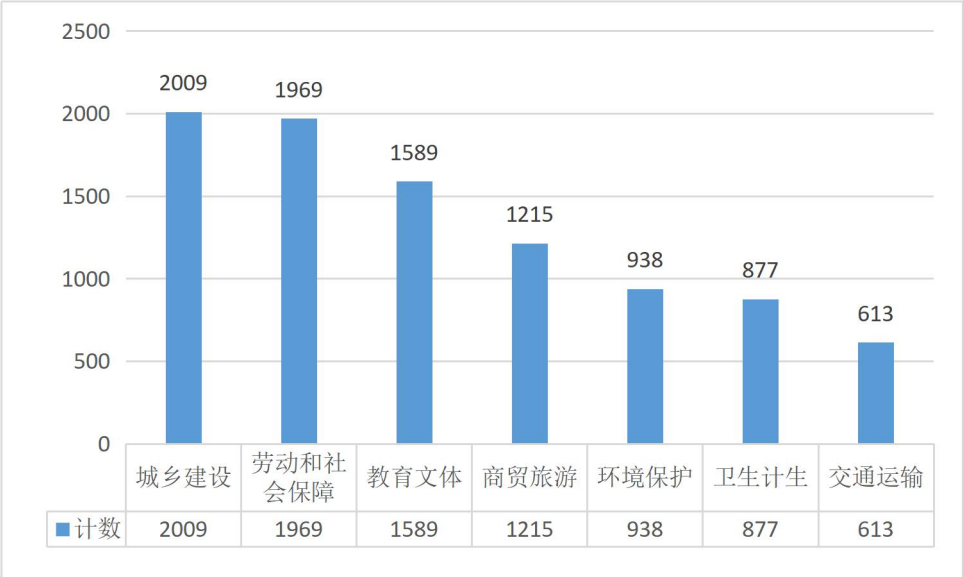


图 11 样本类别及数据统计信息

3.1.2. 数据准备

创建好的数据字典如图 12 所示，储存在 dict.txt 文件中。

{'黄': 0, '曲': 1, '谓': 2, '裴': 3, '援': 4, '八': 5, '室': 6, '审': 7, '钻': 8, '渝': 9, '窑': 10, '鑫': 11, '神': 12, '佛': 13, '划'
'命': 113, '商': 114, '空': 115, 'K': 116, '永': 117, '抖': 118, '烟': 119, '语': 120, '婴': 121, '县': 122, '浦': 123, '蝈'
215, '侍': 216, '鑫': 217, '琵': 218, '缺': 219, '梁': 220, '乌': 221, '泷': 222, '云': 223, '里': 224, '浪': 225, '衍': 22
'P': 318, '澹': 319, '酬': 320, '洽': 321, '门': 322, '体': 323, '钰': 324, '缓': 325, '瓦': 326, '盾': 327, '斤': 328, '亨'
420, '抑': 421, '疆': 422, '耳': 423, '注': 424, '讲': 425, '谊': 426, '贻': 427, '痕': 428, '泵': 429, '庵': 430, '拐': 43
'芳': 523, '鼓': 524, '叉': 525, '武': 526, '檬': 527, '首': 528, '桨': 529, '圃': 530, '军': 531, '航': 532, '昂': 533, '球
625, '松': 626, 'U': 627, '戎': 628, '哺': 629, '完': 630, '.': 631, '薪': 632, '厘': 633, '款': 634, '及': 635, '啦': 636,
'旦': 728, '唱': 729, '背': 730, '智': 731, '磁': 732, '湄': 733, '肢': 734, '夫': 735, '麟': 736, '桑': 737, '凹': 738, '折
830, '·': 831, '看': 832, '忘': 833, '内': 834, '午': 835, '引': 836, '苦': 837, '挡': 838, '凿': 839, '总': 840, '牌': 841,
'玲': 933, '矛': 934, '但': 935, '腹': 936, '日': 937, '相': 938, '允': 939, '0': 940, '沓': 941, '塔': 942, '舜': 943, '疽'
': 1032, '洒': 1033, '效': 1034, '龄': 1035, '簿': 1036, '滋': 1037, '笔': 1038, '似': 1039, '逻': 1040, '成': 1041, '沁
: 1125, '拍': 1126, '肝': 1127, '秘': 1128, '诸': 1129, '鲤': 1130, '纬': 1131, '围': 1132, '必': 1133, '颔': 1134, '鸟
1218, '够': 1219, '季': 1220, '囍': 1221, '选': 1222, '劫': 1223, '连': 1224, '判': 1225, '周': 1226, '乃': 1227, '贴':

图 12 数据字典

抽取 945 条数据作为测试数据储存在 text_list.txt 文件中，其余 8503 条数据作为训练数据储存在 train_list.txt 文件中，所有文件均以数据列表的方式存储，数据列表结构如图 13 所示。

1524,1758,1116,414,791,676,541,2213,308,2804,2025,1499,2286,599,1980,1851,2343 0
1524,1758,2243,1389,522,685,1507,2191,417,284 0
1798,1456,1116,2239,992,791,597,541,2213,635,1116,2239,992,961,2107,599,791,1499,2767,2286,395,158
1524,1758,578,1382,2614,1060,541,1796,299,1961 0
1524,1758,2334,86,1758,1237,801,372,1580,2698,540,2600,1483 0
1524,677,1507,782,1932,2422,1853,1458,635,1373,2000,536,624,2191,417,778,2448,1076,108,1668,2349,6
1524,1931,1507,1464,11,380,87,1116,2216,624,2329,998,1599,1569,1213,2370,778,2585,24 0
1798,1787,1524,677,1507,2293,2221,1796,2366,78,2789,2337,2123,1442,1316,1415,1076,245,2585,24,202
1524,677,1507,1033,1721,1050,2693,372,1580,1811,1146,2370,1499,262 0
1524,2684,1507,599,1500,685,1507,2781,2615,877,766,2573,2422,87,816,2572,1951,2693,15,1890,1088,18
1524,1758,1524,1931,1507,96,414,1275,1507,1192,245,1912,1041,1303,673 0

图 13 数据列表结构

3.1.3. 模型参数

实验组用每一类别的每条留言信息的留言主题用来建立数据词典，通过卷积网络训练模型。

为了评价分类的效果通过平均预测损失（cost）和准确率（acc）对分类结果进行衡量，模型训练结果如表 3 所示。

表 3 模型训练结果数据

训练次数	平均预测损失（cost）	准确率（acc）
0	0.80093	0.76891
1	0.59624	0.81270
2	0.52500	0.83050
3	0.49748	0.83456
4	0.47010	0.83929
5	0.45215	0.84222

3.1.4. 实验结果

将训练好的完整分类模型储存到问题一.py 文件中。

随机从附件 3 中抽取 1489 条留言信息储存到预测文本.txt 中对其进行分类预测得到标签结果和结果概率，储存到预测结果.csv 中，形式如图 14 所示。

城乡建设	0.8561473
商贸旅游	0.47375992
交通运输	0.7227191
环境保护	0.4043659
劳动和社会保障	0.99836963

图 14 预测结果统计表

实验结果表明，该模型的预测较为准确，然而仍有对比结果较低的情况，增加实验数据，扩大数据字典和训练数据，分类性能可能进一步增长。若对留言数据进行优化，分类模型的准确率可能有所提高。

3.1.5. 模型优化[1]

可以结合使用类别关键词提取方法，在留言信息文本分类中可以在卷积神经网络分类性能的基础上，减少模型的训练参数和模型内存，同时高效率的去除无意义词对分类预测的影响，提高模型性能。

3.2. 问题二结果分析

根据中文特点，主谓结构，利用结巴库，提取文本数据中关键名词，词云结果如图所示：

图 15 词云结果图

写关键留言的分文件（见附件 txt1），利用正则表达式删去不必要文字

利用结巴库提取关键名词，关键动词如图 16 所示：

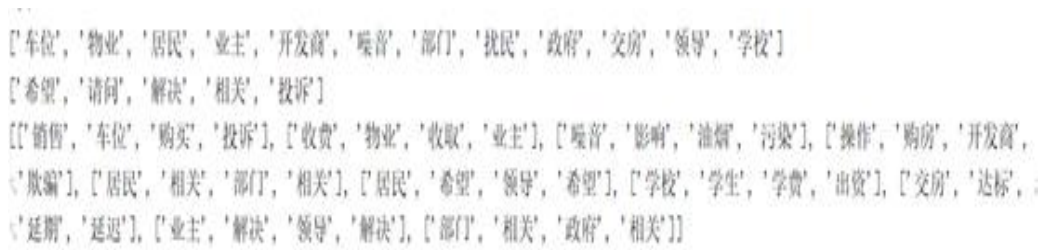


图 16 关键动词结果图

筛选出同时出现关键名词和动词的项，得初步热点留言如图 17 所示。

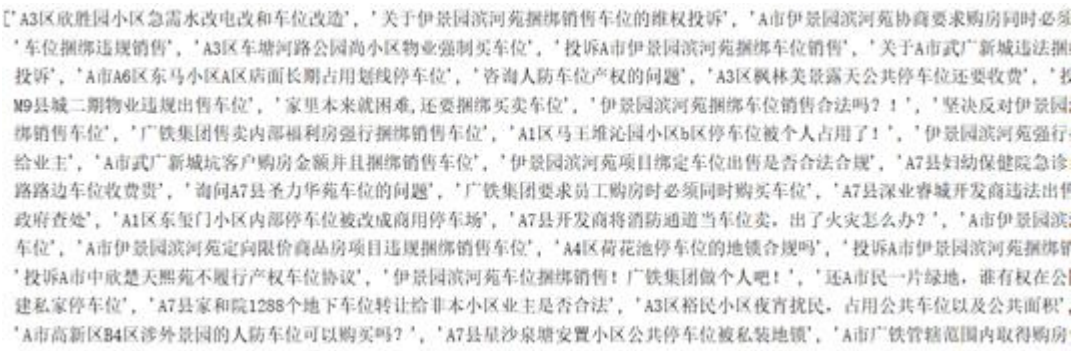


图 17 初步热点留言表

提取热点文本的热点特征文本向量，并对文本进行分析如图 18 所示。

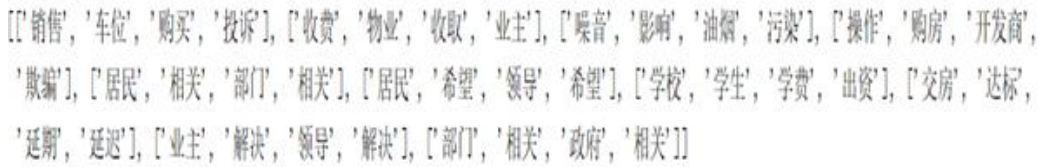


图 18 文本分析结果

计算热点文本权重=关键词权值*文本次数权重值，得热度问题关键向量的热度指标:

热点指数：11.0448 25.1433 9.7656 9.960600000000001 0.5313 7.4217 0.08249999999999999

得到结果，写入热点问题表，热点明细表：程序见附件(热点问题表.py，热点明细表.py)结果见（热点问题表，热点明细表）如图 19。

A	B	C	D	E	F
热度排	问题ID	热度指数	时间范围	地点人群	问题描述
1	1	25.1433	2019/1/15-20	各小区业主，缴纳物业费人	物业乱收费，工作不力、不当问题
2	2	11.0448	2019/1/8-202	主要为有车人民，其中很多	车位捆绑销售，非法占用车位问题
3	3	9.9606	2019/1/4-202	开发商所欺骗的消费者与开	开发商违规操作，欺骗消费者问题反映
4	4	9.7656	2019/1/10-20	居住在不合法开业商铺或者	噪音，污染等影响居民生活问题
5	5	7.4217	2019/1/7-201	各学校学生及家长	学校不当收取学费，或让家长出资行为

图 19 热点问题表，热点明细表

3.3. 问题三结果分析

1. 部分数据的分析如表 4 所示，我们可以看出尽管值很小，但是 Hellinger Distance 的显著性更好。

表 4 部分数据分析内容

留言	回复	Jensen-Shannon Divergence	Hellinger Distance
深圳市缴存住房公积金，是否能在 A 市办理商业住房贷款转公积金贷款？	网友“UU0081604”您好！您的留言已收悉。现将有关情况回如下：目前，A 市住房公积金管理中心不支持持、理解与监督！2018 年 12 月 26 日	0.0217468	0.153709
请问地铁 3 号线烈士公园东站出入口是如何规划的？在晚报大道南侧是否有出入口？如没有能否解释原因？	网友“UU008173”您好！您的留言已收悉。现将有关情况回如下：地铁 3 号线烈士公园站在晚报大道与车站北路交叉口东北向设有 2 个出入口，西北向设有 1 个出入口，晚报大道南侧没有设置出入口。地铁是一个杂的系统工程，其站点和出入口是充分考虑地质条件、建设条件、交通组织、地铁运营、工程可行性的相关要求等综合因素后设置的。	0.0375926	0.202578

	感谢您对我们工作的支持、理解与监督！2018年12月4日		

2. 以第54条答复为例：

表5 第54条答复及结果分析

答复	网友“UU0082042”您好！您的留言已收悉。现将有关情况回如下：据查，spaceplus酒吧位于A市车站中路宇成朝阳广场金朝阳大厦a区t1、t2商业裙房3、4层，所在建筑金朝阳大厦a区地上27层，建筑高度99.55米，地上建筑面积84487.23平方米，设计使用性质第一至五层为商业，t1栋6-25层为酒店式公寓，t2栋6-27层为写字楼。spaceplus酒吧经营范围为第三、四层局部，装修面积2800平方米，使用性质为酒吧，经营单位为A市史斯贝餐饮管理有限公司。经核查，spaceplus酒吧存在消防验收不合格擅自投入使用、公众聚集场所未经消防安全检查擅自营业的违法行为。对此，市消防支队已依法立案查处，并督促指导其整改，目前案件正在办理中。感谢您对我们工作的支持、理解与监督！2018年12月10日

内容回复 60’	回复人名称 4’	落款日期 8’	落款单位 8’	鸣谢回复 10’	正式回复表达 10’	Total	类别
1	1	1	0	1	1	92	**非常完整**

2. 第26条留言如表6所示，进行Token后的结果如表7所示：

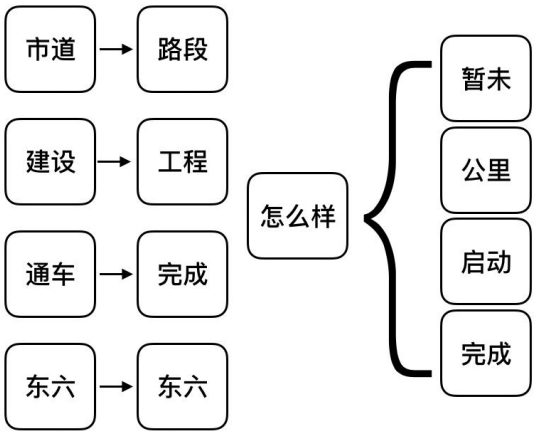
表6 第26条留言内容

留言	A市高铁新城管委会负责建设的A市东六线（劳动东路～机场高粟）段目前建设建展怎么样，何时能够正式通车，现在到A市道，意义重大。
回复	网友“UU008694”您好！您的留言已收悉。现将有关情况回如下：东六线（楚府东路—机场高粟）道路工程全长5.8公里，其中段长度2公里，已完成约800米路基，其余路段因涉及到国有土地（正钢机械厂，星沙机床厂等）征拆问题，暂未启动施工。对我们工作的支持、理解与监督！2018年12月28日

表7 Token结果表

留言	管委会 负责 建设 东六 建设 怎么样 通车 市道 意义
回复	东六 道路 工程 公里 段长 公里 完成 路基 路段 涉及 国有 土地 暂未 启动

可以建立模糊词表，即相同意义特征的词可以互换，从而减少特征流失。从而通过训练模型建立可解释词表称为替代词。而在上部提取出的特征信息，我们可以明显的发现“建设”和“工程”是互补词，“通车”和“完成”是一组互补词，“市道”和“路段”具有相同方向的词性，“东六”相互匹配，“怎么样”可以被“公里”，“完成”，“启动”，“暂未”等词解释。



名称	匹配词个数	权重
市道	1	1/9
建设	2	2/9
通车	1	1/9
东六	1	1/9
怎么样	4	4/9

CRI=0.833

图 20 权重结果示意图

在完成特征词归纳后，应该进行权重分配。对可解释词的匹配数量多的做赋予大权重，可以认为在回复里，花了更多精力解释这个词，也说明这个特征词在留言里的特征程度更高。最后通过计算权重调整后的可解释程度，作为可解释性的标准。

4. 结论

本文从群众的角度出发，结合政府现有信息发布平台的有效发展及延伸，在实现随时随地获取信息和在线留言互动等功能的前提下，逐步完善智慧政务系统。逐步实现群众留言分类的智能化、高效化；进一步推进有关部门适应新形势，及时获取群众关心的热门问题，以提高群众满意度为目标的实践，切实提升为民服务水平，真正让群众感受到信息化、科技化带来的便捷。为群众打造更优质的优质的政务服务，拓宽监督和参与渠道。

结合现下答复意见情况，通过总结揭示了留言答复服务存在的问题，提出以下服务提升的意见和建议。

一、答复内容解读工作加强。详细答复的发布可以使群众及时全面的了解当下政策方针，帮助群众了解自己所反映问题对自己生活带来的影响，以及帮助法规内容更好的实施。

二、有关部门要明确留言答复的工作标准，不允许出了問題“相互<皮”情况的发生。推进国家信息化政务服务建设。依托电子政务内、外网资源，建立信息资源管理中心。各级部门要经常相互交流，吸取经验教训，提高留言答复的应用水平。加强与国际电子政务技术、信息资源、人才培养等领域的交流与合作。

留言服务的发展会给有关部门内部运作和部门重组带来新的要求和内容，同时也带来新的发展契机。“留言政务”是信息时代的产物，它势必要求有关部门进行适应信息时代，不仅是实现运作手段的变化，而且是要重新设计和定位留言服

务职能，再造工作流程，创新并整合服务形式，从而给民众提供综合的、全方位的、个性化的服务。

参考文献

[1] 张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.
[2] Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification,
[3] 岳清清. 深度学习在岩石薄片图像检索中的应用研究[D].西安石油大学,2019.
[4] 周非,李阳,范馨月.图像分类卷积神经网络的反馈损失计算方法改进[J].小型微型计算机系统,2019,40(07):1532-1537.
[5] 孙新胜. 基于多层卷积神经网络的研究与应用[D].杭州电子科技大学,2018.
[6] 黄怡璇,谢健民,秦琴,杨丽颖.影响网络舆情热度评价的主要因素识别研究[J].情报科学,2017,35(10):49-54+62.
[7] 晏榛,杨晓蓉.地市级政府信息公开质量评价实证研究[J].科技资讯,2019,17(22):198-200.
[8] 孙梦皎. 基于用户视角的河北省政府交通出行信息质量评价研究[D].河北大学,2019.
[9] 马一鸣. 政府大数据质量评价体系构建研究[D].吉林大学,2016.
[10] 胡吉明,李雨薇,谭必勇.政务信息发布服务质量评价模型与实证研究[J].现代情报,2019,39(10):78-85.
[11] Christian G. Service Quality Model and Its Marketing Implications ［J］. European Journal of Marketing，1984，18 (4) : 36－40
[12] https://github.com/baidu/Familia/wiki/