第八届“泰迪杯”

数据挖掘挑战赛

作品名称：“智慧政务”中的文本挖掘应用（c 题）

“智慧政务”中的文本挖掘应用

摘要

本文关注“智慧政务”中的文本挖掘问题。在数据预处理中，先对各个附件中的数据进行分词，去停用词，降低后续数据处理的复杂度。

针对问题一，首先使用 LSTM 文本分类模型对留言详情按照分类标签进行分类预测，然后通过 F-Score 来评价一级标签分类模型。分类结果表明，教育文体的分类预测的准确度较高，F1-score 值最高 0.90。

针对问题二，首先识别并提取附件三中留言主题中涉及县、区的词，判定是否可能成为热点问题关键词；然后对留言主题中涉及县、区等词的数据进行相似度计算并聚类出可能是热点问题的数据，结合正则表达式匹配附件三得到热点问题，计算并排序数量前八的热点问题；最后用留言数量、时间范围、点赞数和反对数建立热度评价指标。

针对问题三，对答复意见进行多角度评价，运用计算余弦相似度的方法判断答复意见的相关性；运用正则表达式匹配和人工筛选等方法判断答复意见的完整性，运用正则表达式匹配和系统抽样等方法判断答复意见的可解释性。评价结果表明，答复意见中有 98.7%符合相关性、97.2%符合完整性以及 94.14%符合可解释性。

关键词:LSTM 建模;TF-IDF 模型;正则表达式匹配;余弦相似度

1. 背景与挖掘目标

1.1. 问题背景

近年来，随着微信、微博、市长信箱、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系统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. 挖掘目标

本次建模目的是利用来自互联网的群众问政留言记录和相关部门对部分群众留言的答复意见的数据，解决以下三个问题：

（1）利用文本分词的方法对非结构化的数据进行文本挖掘，根据附件二中留言详情和一级标签的关系，构建 LSTM 分类模型，并用 F-Score 评价模型的好坏。

（2）利用聚类的方法挖掘热点问题，根据附件三中的留言数据将某一时段内反映特定地点或特定人群的问题的留言进行归类。并根据数据的特征定义合理的热度评价指标，给出评价结果。

（3）根据附件四中相关部门对留言的答复意见，通过相关性，完整性，可解释性等角度对答复意见进行评价。

2. 分析方法与过程

本次数据挖掘的总体流程如图 1 所示：

数据读取：通过 Python 软件对附件二、附件三、附件四实现数据的读取。

数据预处理：对题目数据进行数据清洗，其中对附件二的数据进行删除标点

符号、中文分词和去停用词等处理。对附件三、四的数据进行中文分词，去停用词处理，以便降低相似度计算的复杂度。

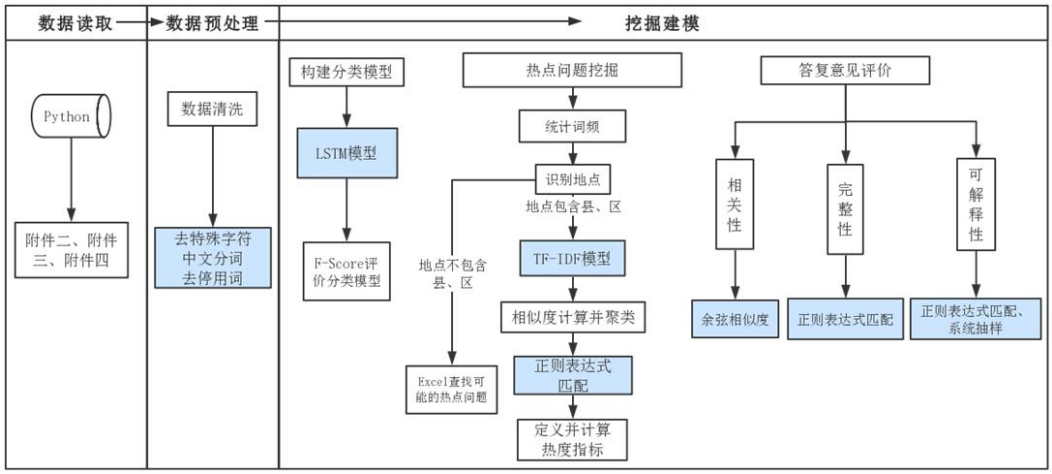

图 1 总体流程图

挖掘建模：对于问题一运用 LSTM 分类模型建立留言标签分类模型，再用 F-score 评价留言标签分类模型。对于问题二进行热点问题挖掘，先通过 doc2bow 构建词袋模型，并得到新语料库，使用 TF-IDF 模型进行文本相似度计算，然后选出可能成为热点问题的留言关键字进行正则表达式匹配附件三，并将数据提取到 Excel，再通过文本特性定义并计算热度评价指标。对于问题三答复意见评价，通过计算余弦相似度、正则表达式匹配和、人工筛选、系统抽样等方法判断答复意见的相关性、完整性以及可解释性。

3. 问题一分析过程与结果

3.1. 思路流程图

解决问题一的流程图如图 2 所示：

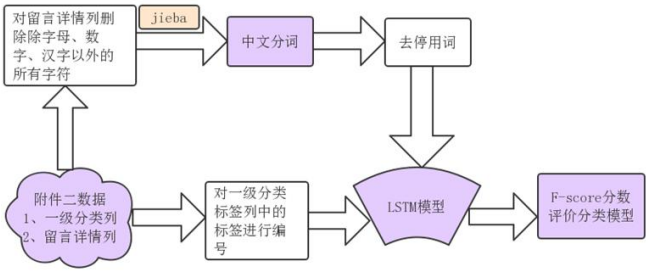

图 2 问题一流程图

3.2. 数据预处理

数据编号：汇总附件二中的一级标签列中包含的标签，再将所含标签与数字（0,1....）建立一一对应关系，对应的数字代表对应的一级标签，以便后面分类模型的建立。

中文分词：在建立分类模型之前，先要把非结构化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，以中文的文本的方式给出了数据。为了便于模型的建立，要先对附件二的留言详情进行中文分词。这里采用 Python 的中文分词包 jieba 对数据进行分词。

去除停用词及其标点符号：附件二中的留言详情包含一些没意义的常用词（stopword）和标点符号，这些词对系统分析文本的内容没有任何帮助，反而会增加计算的复杂度，所以在使用文本数据之前必须将它们清理，其中预处理代码见附件“LSTM.ipynb”。

3.3. 构建 LSTM 分类模型

用 Python 建立基于留言内容的一级标签的 LSTM 分类模型。LSTM 分类模型的优势是能够综合考虑句子中的语义部分，大大的提高了对应的训练精准度，LSTM 分类模型基本流程见图 3。

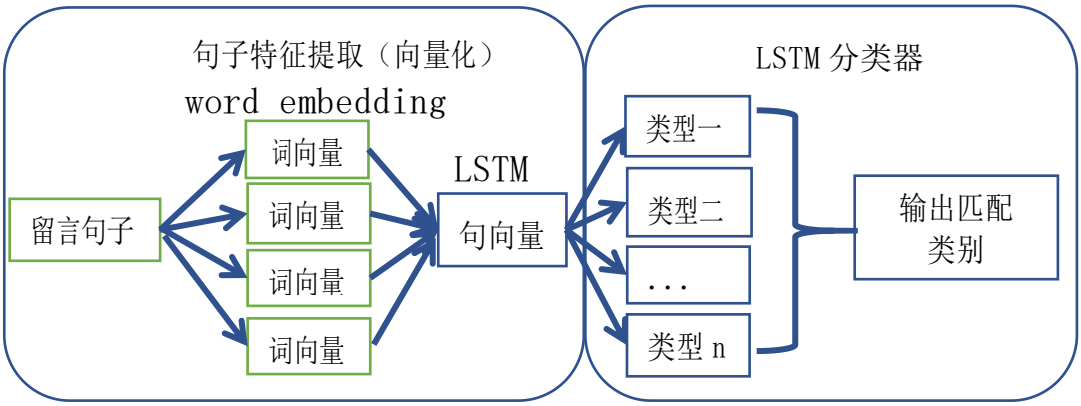

图 3 LSTM 分类模型流程图

基于 LSTM 分类建立的留言分类模型需要经历以下三个步骤：

①将分词后的留言详情信息进行词向量（word embedding）处理。

②设置最频繁使用的词数。

③设置每条留言详情的最大词语数，定义模型，且由于是多分类所以激活函数设置为“softmax”,损失函数为分类交叉熵 categorical_crossentropy。

用分类模型测试数据，其中，程序代码见附件“LSTM.ipynb”。

3.4. F-Score 评价模型

F-Score 是精度和召回率的调和均值，常用于评价分类模型的好坏，对于多分类模型准确率（accuracy）不能反应每一个分类的准确性，因为每类数据不平衡时 accuracy 不能反应出模型的实际预测精度。所以最终建立的分类模型使用 F1 分数进行评价，满足：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。得到各个类的 F1 分数如下：

accuracy 0.8263386396526773

	precision	recall	f1-score	support
城乡建设	0.79	0.81	0.80	305
环境保护	0.86	0.83	0.84	141
交通运输	0.82	0.61	0.70	97
教育文体	0.92	0.88	0.90	251
劳动和社会保障	0.86	0.92	0.89	276
商贸旅游	0.66	0.72	0.69	173
卫生计生	0.85	0.86	0.86	139
accuracy			0.83	1382
macro avg	0.82	0.80	0.81	1382
weighted avg	0.83	0.83	0.83	1382

一级标签分类模型的准确度为 82.63%，其中类别为城乡建设、环境保护、教育文体、劳动和社会保障以及卫生计生的 F_1 分数分别为 0.80、0.84、0.90、0.89 和 0.86，但是由于原始数据中类别为交通运输数据相对较少所以它的 F_1 分数比较低，只有 0.70，而类别为商贸旅游的精度较低只有 0.66，所以 F_1 最低，只有 0.69，具体程序代码见附件“LSTM.ipynb”。

4. 问题二的分析过程与结果

4.1. 思路流程图

解决问题二的流程图如图 4：

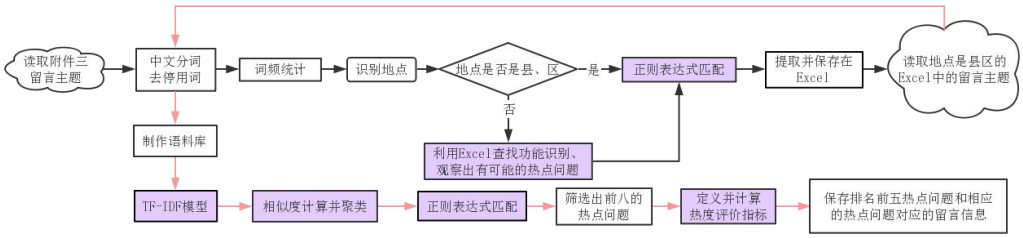

图 4 问题二流程图

4.2. 数据预处理

中文分词：为了方便后面数据的处理，将原始数据附件三和地点是县、区的 Excel 中的留言主题采用 Python 中的中文分词包 jieba 进行中文分词。

去停用词：原始数据附件三和地点是县、区的 Excel 中的留言主题中包含许多没有明确意义的词，为了节省存储空间和提高计算效率，要将这些没有明确意义的词进行删除处理，其中预处理的代码见附件“sim.py”。

4.3. 热点问题预处理

对数据预处理后，统计所有留言主题中词语的词频，根据统计结果判断出现较多次的地点名称，地点名称包括区、县或者具体小区等。再把没有具体区、县的地点识别出来，用 Excel 去查找，这样可以人工的识别出在此地方是否可能有热点问题，然后根据关键字用正则表达式匹配附件三，直接提取热点问题。对识别出来的区、县的地点根据其名称利用正则表达式匹配附件三，并提取到 Excel，其中统计词频识别地点的程序代码见附件“frenc.py”， 结果数据文件汇总见文件夹“地区”。

4.4. 语料库

语料库是一组向量，向量中的元素是一个二元组（编号、频次数），对应分词后的文档中的每一个词。

制作语料库的步骤如下：

（1）将地点是区、县的 Excel 表格中的留言主题通过 dictionary 方法获取词袋（bag-of-words)。

（2）用数字对词袋中所有词进行了编号，并建立编号与词之间的对应关系。

（3）用 doc2bow 制作留言主题的语料库。

4.5. TF-IDF 模型

TF-IDF 模型的主要思想：如果一个词在同类型的一条留言中出现的频率高，并且在其他类型的留言中很少出现，则认为词具有很好的区分能力，适合用来把这类留言和其他留言区分开来。TF-IDF 模型的具体原理如下：

第一步：计算词频（TF），某个词在文本中出现的次数，为了方便不同留言的比较，通常对“词频”标准化，除以文本中词条总数：

$$TF = \frac{\text{某个词在文本中出现的次数}}{\text{文本总词数}}.$$

第二步：计算 IDF 权重，即逆向文件频率，需要运用语料库，用模拟语言的使用环境。IDF 值越大，此特征性在文本中的分布越集中，说明该分词在该文本内容的属性能力越强。

$$IDF = \lg(\frac{\text{语料库中的文本总数}}{\text{包含该词的文本数+1}}).$$

第三步：计算 TF-IDF 值

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

实际分析得到 TF-IDF 值与一个词在留言文本出现的次数成正比，该词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的留言文本的关键词。

4.6. 留言的相似度计算

基于相似度的文本聚类的算法如下：

（1）以第一篇留言主题作为第一个类别，让它与其他所有留言进行相似度计算。

（2）设定一个阈值（阈值由附件三数据的具体计算结果来确定）将结果高于阈值的留言对应的索引值存放在一个列表中。

（3）判断第二条留言主题的索引值在不在之前已经形成的列表中，如果在：跳到第三条留言主题；如果不在：将第二条数据作为第二个类别。

（4）重复以上步骤。如此循环到最后一条留言详情。

该算法的流程图如图 5：

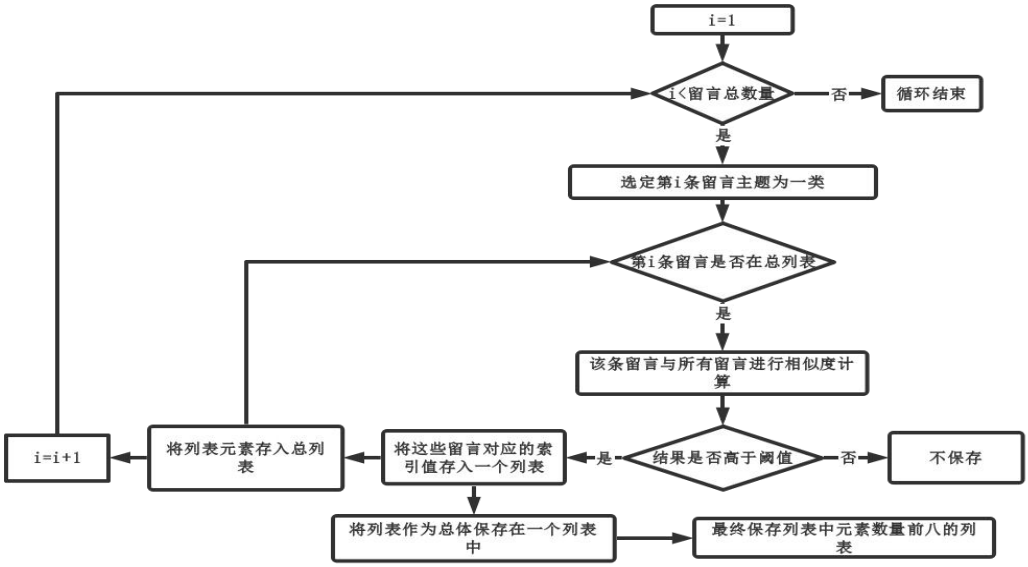

图 5 基于相似度的文本聚类的算法流程图

由前面的算法将得到若干列表，每个列表存放着归为同一类别的留言的索引值，然后计算出列表长度，得到长度前 8 的列表。根据这 8 个列表把对应的留言分别存放到 Excel 表中，其中代码见附件“sim.py”。

4.7. 正则表达式匹配

为了提高聚类精度，用 Excel 查找功能识别留言数据中可能成为热点问题以及对留言数据进行相似度计算得到可能成为热点问题的信息选出关键字，用正则

表达式将关键字与附件三中的留言数据进行匹配，并将提取数据到 Excel，最后筛选出数量前八的热点问题，其中程序代码见附件“rel.py”，数据见文件夹“数量前八的留言”。

4.8. 热度评价指标

根据观察正则表达式匹配后得到的数据中的数据特征，可以认为热度指标与留言的数量和时间范围成正比关系，留言数量越多，时间范围越长，热度越高。考虑到时间范围问题，可以设置一个时间权重。同时点赞数和反对数也是提升议论热度的一个重要因素。所以热度指标可以定义为：

热度指标=权重×（初始热度分+用户交互热度分）；

其中，初始热度分=留言条数（每条 1 分），用户交互热度分=点赞数+反对数（每个 1 分），权重设置如下：

时间范围在 0-3 个月的权重设置为 1；

时间范围在 3-6 个月的权重设置为 1.25；

时间范围在 6-9 个月的权重设置为 1.5；

时间范围在 9 个月以上的权重设置为 1.75。

根据结果分析：排名前五的热点问题中 A 市暮云街道丽发新城小区附近搅拌站污染环境、噪音扰民的热点问题的热度最高，热度指数为 109，部分数据如下表 4-1 ，全部结果见附件“热点问题表.xls”，得到相应热点问题对应的留言信息，部分数据如下表 4-2，全部结果见附件“热点问题留言明细表.xls”。

表 4-1 部分热点问题结果

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	109	2019/11/2至2020/1/26	A市暮云街道丽发新城小区	A2区丽发新城附近修建搅拌厂污染环境，噪音扰民。
2	2	93	2019/7/11至2019/9/1	A市伊景园滨河苑	A市伊景园滨河苑无视职工意愿、职工权益，强行车位捆绑销售行为。
……	……	……	……	……	……

表 4-2 部分热点问题留言明细表

问题ID	问题编号	留言用户	留言主题	留言时间	留言详情	赞成数	反对数
1	267050	A909227	噪音、灰尘污染的A2	2019/11/2 10:18:00	A2区丽发新城附近修建搅拌厂	0	0
……	……	……	……	……	……	……	……
2	212323	A00020702	广铁集团要求员工购	2019/7/11 0:00:00	尊敬的领导：您好！我是一名	0	0
……	……	……	……	……	……	……	……
3	289408	A0012413	在A市人才app上申请	2018/11/15 16:07:12	我叫朱琦梦，是2017年12月3	0	0
……	……	……	……	……	……	……	……

5. 问题三分析过程与结果

5.1. 相关性

数据预处理：读取附件四中的数据，为了降低后续词袋模型等操作的复杂度，采用 Python 中的中文分词包 jieba 对数据进行中文分词，再删除数据中没有明确意义的停用词。

数据预处理后对数据进行测试确定合适的阈值，再对留言详情和答复意见通过计算余弦相似度的方法判断答复意见的相关性。

基于相关性的算法步骤如下：

1、对第一条留言详情和答复意见进行分词、去除停用词。

2、用字典保存留言详情和答复意见中出现的词并编上号。

3、根据词袋模型统计词在留言详情、答复意见中出现的次数，形成向量。

4、计算余弦相似度。

5、将结果保存在列表。

6、重复步骤 1-5，直到所有留言详情和答复意见计算完相似度。

7、根据阈值取出列表中元素对应的索引值，代表数据在原文件中的行数。

8、根据索引值提取文件。

结果分析：经过以上步骤提取出了 143 条相似度低的数据，需要人工对数据进行再次处理，去除对机器未能识别且相关的留言数据。除此之外相似度低的数据中，出现较多形如“……关于您反映的问题，已转……处理”的回复，这些回复表明留言者的问题已经有了相应的处理，所以是相关的，也可以去除。最后经处理剩下 37 条留言被认定为不相关。因此有 98.7%的答复意见符合相关性。部分不相关留言数据如表 5-1，全部不相关数据见附件“3_1.xlsx”，程序代码见附件“three_1.py”。

表 5-1 不相关的答复意见数据

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
6556	UU0081320	打狂犬役苗报销比例是	2018/3/20 15:19:47	又合费用增加了，打狂犬役苗报销比例是多少。	已收悉	2018/3/28 16:05:34
10029	UU0082150	污水直排农田，造成	2016/7/19 20:48:58	污水管道，请问农民的田塘坝深井水怎么解决，	感谢您对我们工作的关心、监督与支持。	2016/8/24 16:02:17
11927	UU0081626	工矿棚户区改造项目	2014/10/2 19:53:03	有关部门查处A市轴承厂工矿棚户区改造项目中	网友：您好！留言已收悉	2014/11/5 16:44:34
25918	UU0081182	云段A1区南路路灯及	2014/4/21 1:40:59	使用，而且整条A1区南路A市段的路灯就在伊莱	“UU0081182”	2014/5/28 15:11:52
30019	UU008151	备全款购买二手房事	2016/11/3 10:00:17	全款购买二手房，房产局资金监管走哪家银行，	已收悉	2016/11/22 12:25:56
37346	UU008937	片区缺少书店，希望能	2019/7/7 11:43:29	提供一个购买书籍的地方和阅读的空间，建议	网友：2019年7月9日	2019/7/10 16:53:16
……	……	……	……	……	……	……

5.2. 完整性

从回复的开头、结尾的格式来分析完整性。答复意见满足如下格式的几点或者全部则可以认定此条答复是完整的。通常答复的内容包含以下几部分：

开头：一般含有称谓、问候语等，如“网友 XXXX 您好”；写清楚发函的缘由，概括交代发函的目的、根据，如“……已收悉”。用过渡语，如“现将有关情况回复如下”等引起下文。

正文：是复函的核心部分，要用简洁得体的语言把要告诉对方的问题、意见讲清楚，使对方接到函后能快速了解来函的意图，准确作出反映，叙写清楚。答复意见要有针对性。

结尾：一般用礼貌性的语言向对方作出答复，例如可以包含以下惯用语：“感谢您对我们工作的支持、理解与监督”、“特此回复”等

落款：由发文机关和成文日期组成。 发文机关写全称或规范化简称。成文日期要用汉字写清楚年、月、日,加盖公章。

基于上面的分析，采用正则表达式匹配关键字的方法提取文件，选出常见的 6 个关键词：“网友：”、“您好”、“你好”、“您反映的问题”“答复”、“回复”进行匹配，并提取文件。出现上述关键词的可认定为完整的。

结果分析：经过筛选有 149 条数据被认定为不完整，为了提高精度，需要进行人工再次筛选，最终结果有 79 条留言数据被认定为不完整。因此有 97.2%的答复意见符合完整性。部分不相关留言数据如表 5-2，全部不完整数据见附件“3_2.xlsx”，程序代码见附件“three_2.py”。

表 5-2 不完整的答复意见数据

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
6556	UU0081320	咨询打狂犬	2018/3/20 15:19:47	请问领导，农合费用增加了，打狂犬役苗报销比例是	已收悉	2018/3/28 16:05:34
10029	UU0082150	反映A市医	2016/7/19 20:48:58	书记您好，我是A6区大泽湖东马社区一名农民，反映	感谢您对我们工作的关心、监督与支持。	2016/8/24 16:02:17
11222	UU0081662	请求为A市	2015/7/29 11:50:15	我于2009年购买比华利山的房子，由于开发商将房屋	网友“UU0081662” 据查，比华利山位于中意二路	2015/8/12 11:47:28
23770	UU0081855	咨询A8县对	2017/3/1 13:08:38	请问政府对双江口镇罗巷集镇新型城镇建设规划集镇	一、罗巷集镇于2013年进行了基础设施建设，公	2017/3/2 15:02:56
……	……	……	……	……	……	……

5.3. 可解释性

可解释性即是答复意见对留言内容的相关解释，一般而言，引用了相关法律法规、政府文件等文件的答复可以判定为满足可解释性，所以对附件 4 中 2816

条数据采用正则表达式匹配并去除认为可解释的答复数据,最终剩下 2050 条数据采用系统抽样的方法抽取样本，然后进行人工识别来分析可解释性。

人工评判标准：如果用户通过阅读答复意见之后没能获取有用的信息、或者用户看了答复意见之后还有疑惑则该条答复意见可以认为不可解释。例如回答不完整，答到一半没有下文；只回了已收悉，没说明任何处理；回答与留言完全不相关；回复交给相关部门处理但没有指定具体的部门等情况。

采用系统抽样从 2050 条数据中抽取样本容量 n=102，抽样距离 k=20，共抽 5 组数据进行人工识别。其中通过系统抽样的 5 组数据人工识别的结果如下：

第一组数据情况：11 条答复意见不可解释，占样本数据的 10.78%；

第二组数据情况：10 条答复意见不可解释，占样本数据的 9.8%；

第三组数据情况：5 条答复意见不可解释，占样本数据的 4.9%；

第四组数据情况：10 条答复意见不可解释，占样本数据的 9.8%；

第五组数据情况：5 条答复意见不可解释，占样本数据的 4.9%。

结果分析：样本数据中平均不可解释的答复意见占 8.04%，所以总体 2050 条数据中不可解释的答复意见数据约有 165 条，即 2816 条原始数据中不可解释的答复意见数据也约有 165 条，因此有 94.14%答复意见符合可解释性。其中部分答复不可解释数据如下表 5-3，全部抽样数据见文件夹“sample”，正则表达式匹配程序代码见附件“three_3.py”，系统抽样程序代码见附件“System.py”。

表 5-3 不可解释的答复意见数据

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
12163	UU0082207	请求701或915路	2014/7/9 13:	701路和91	网友：您好！留言已收悉	2014/8/13 17:43:11
30425	UU0081485	投诉A6区南塘村	2014/2/24 16	本人是A6区	网友：　　您好！您的意见已收悉，我们将迅速向有关	2014/3/10 16:23:31
93271	UU0081810	关于J11市公租房	2014/7/19 11	领导：　　您	网友：您好！留言已收悉	2014/7/24 15:56:02
98755	UU008436	咨询K市升学择校	2018/8/26 10	百年大计教育为	“UU008436”您好！您的帖文已收悉。我办已将您反映的情况转交相关科室	2018/10/15 8:38:11
118451	UU0081779	K6县君泰家园违	2019/12/30 10	K6县君泰家园小	“UU0081779”您好！您所反映的问题已收悉。就您所反映的问题我办已转交	2019/12/31 10:40:33

5.4. 答复意见评价

经过相关性、完整性、可解释性三个角度评价附件四种相关部门给出的答复意见，其中这些答复意见中有 98.7%符合相关性、97.2%符合完整性以及 94.14%符合可解释性，所以相关部门给出的答复意见的质量较好。

6. 总结

本文从自然语言处理和深度学习的角度进行研究，运用中文分词工具、构建LSTM模型、TF-IDF模型和文本相似度计算进行理论分析。首先对附件二、附件三、附件四中的数据进行预处理，通过分析群众的留言信息，对异常值进行了修正与剔除；去除停用词和特殊符号，利用中分分词工具对数据进行分词，得到对我们有用的信息的词条。

针对问题一，使用LSTM文本分类模型对留言详情按照分类标签进行分类预测。最后通过求F1分数来评价一级标签分类模型。针对问题二，先通过doc2bow构建词袋模型，并得到新语料库，使用TF-IDF模型进行文本相似度计算，然后选出可能成为热点问题的留言关键字进行正则表达式匹配附件三，并将数据提取到Excel，再通过数据特性定义并计算热度评价指标。针对问题三，在对留言详情和答复意见进行相似度计算，检验两者之间的相关性；对答复意见探索是否达到某种格式来评价是否完整，和对答复意见是否满足群众的需求判断是否具有可解释性。

基于对自然语言的处理，有利于网络问政平台将群众留言分派至相应的职能部门进行处理，减少了电子政务系统因为依靠人工经验处理留言而存在效率低和差错率高等问题的发生；有利于及时发现热点问题，让相关部门进行有针对地处理提升服务效率，及时解决人民群众的问题；有利于评价答复意见，可以有效地判断相关部门给出的答复意见的质量。

参考文献

[1] 王莹. 基于深度学习的文本分类研究[D].沈阳工业大学,2019.

[2] 甘秋云.基于TF-IDF向量空间模型文本相似度算法的分析[J].池州学院学报,2018,32(03):41-43.

[3] 陈航宇. 正则表达式匹配算法研究[D].燕山大学,2016.

[4] 林于杰. 结合评论生成的可解释性时尚推荐研究[D].山东大学,2019.

[5] 苏东出.基于TF-IDF和余弦相似度的图书馆OPAC系统的研究和实现[J].内蒙古科技与经济,2019(21):67-69.

附录

问题	文件名	备注
问题一	LSTM.ipynb	问题一数据预处理、LSTM 分类模型和 F-Score 的代码
问题二	sim.py	问题二数据预处理和基于相似度的文本聚类的算法的程序代码
	frenc.py	统计词频识别地点的程序代码
	地区	提取出来的地点数据文件
	rel.py	正则表达式匹配的程序代码
	数量前八的留言	筛选的数量前八的留言
	热点问题表.xls	排名前 5 的热点问题数据文件
	热点问题留言明细表.xls	热点问题留言明细的数据文件
问题三	three_1.py	判断答复意见相关性的程序代码
	3_1.xlsx	答复意见不相关的数据文件
	three_2.py	判断答复意见完整性的程序代码
	3_2.xlsx	答复意见不完整的数据文件
	three_3.py	正则表达式匹配的程序代码
	System.py	系统抽样的程序代码
	sample	判断可解释性的样本数据