"智慧政务"中的文本挖掘应用探究

关键词

LSTM DBSCAN Word2Vec 评价模型 智慧政务

摘要

本文通过致力于实现“智慧政务——文本挖掘系统“，一个对”民众“、”工作人员“都友好的政务系统。

第一部分，留言分类系统。为解决数据不平衡问题使用回译进行数据增强；将文本截断为长度为 200 的词序列，用词嵌入模型将词序列转化为由词向量组合成的矩阵；在 Keras 框架下用 LSTM 进行有监督的分类。

第二部分，我们在对留言标题进行预处理后，根据词性和词向量进行特征提取，然后使用 DBSCAN 密度聚类。再根据时间，评论，点赞进行热度排序。

第三部分，我们定义了及时性、完整性、可解释性、相关性四个指标和具体评分算法，并对每一个回复进行打分。

关键词：LSTM DBSCAN Word2Vec 评价模型 智慧政务

Abstract

This paper focuses on the realization of "Intelligent Government - Text Mining System", a government system friendly to the people and staff.

The first part is the message classification system. To solve the problem of data imbalance, data enhancements are performed using backtranslation; the text is truncated into a 200-length word sequence; the word sequence is transformed into a matrix composed of word vectors using a word embedding model; and supervised classification is performed using LSTM within the Keras framework.

In the second part, after preprocessing the message title, we extract feature based on word type and word vector, and then use DBSCAN density clustering. Then sort the heat by time, comment and approval.

In the third part, we define four indicators of timeliness, completeness, explanability, relevance and a specific scoring algorithm, and rate each response. The score approximates the normal distribution.

Keywords: LSTM DBSCAN Word2Vec Smart-government Evaluation model

目录

一、问题重述与分析

1.1 问题重述

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，群众通过手机、电脑等数据产品就能够在网络平台上留言、转发以及评论点赞，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。在传统的数据时代，相关工作人员主要通过抽样调查、内容分析等方式，获取有限的具有代表性的样本信息，并通过统计学方法进行分析。

然而，随着大数据、云计算、人工智能等技术的发展，传统的数据分析已经无法适应当代爆炸性增长的海量舆情信息，样本分析被总体分析所取代。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，快速准确地对热点关注问题进行分析，政府部门能够及时了解民众对某个话题的关注程度和社会迫在眉睫的问题，从而进行正确的方针指导，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 问题分析

1.2.1 问题一：一级标签分类

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。这是一个文本多分类问题，采用基于深度学习的监督式分类算法，并使用 F-score 对分类结果进行评价。

难点分析处理：(1)每条留言内容属于一个长文本，为去除大量无关的信息，

将留言内容分成短文本和提取关键词，使得模型训练更加准确有效；（2）考虑数据不平衡带来的影响，相应进行数据增强。

1.2.2 问题二：热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题。首先进行问题识别，即通过两两文本相似度比较，从众多留言中识别出相似的信息；其次命名实体识别，相同的地点或人群进行归并；最后，定义合理的热度评价指标，实现热点问题的排行，主要按照热点问题的留言数、时间的集中度和点赞反对数的优先级定义。

难点分析处理：（1）命名实体的表达多样化，选择灵活性强的词典，不可忽视近似词语表达；（2）对于海量的留言内容，两两特征比较会引起庞大的计算量和不必要的内存占用，同样作短文本拆分和关键词筛选处理。

1.2.3 问题三：答复意见的评价

本文建立的评价模型参考指标有以下几点：（1）及时性：回复是否及时，政务问题一般都具有很强的时效性，回复的及时性就显得尤为重要；（2）相关性：留言内容与答复意见的相关性，是否答非所问（文本相似度分析），命名实体是否相同（特定的地点或人群保持一致）；（3）完整性：句子本身的完整性，即具备主谓宾通用结构；句子的规范性，有开头结尾、礼貌用语等；（4）可解释性：是否有论点、论据支撑，“根据…规定”等句式。

难点分析处理：（1）及时性、相关性、完整性和可解释性量化描述；（2）构建指标计算和评价答复意见。

二、研究假设与约定

下面对题目原始假设和必要的建模假设做出说明:

2.1 问题一假设

1. 第一题建立关于留言内容的一级标签分类模型，一级分类共有 7 类，假设这样的划分是合理的,即每条留言内容归属的一级分类必然属于 7 类中的一种，在允许个别偏差存在的情况下，不考虑出现的额外一级分类。
2. 文本分类过程中，可能存在文本语义带来的词语交叉问题， 本题忽略词语交叉的影响，事实上，在我们的监督式分类模型训练下，词语交叉的影响确实微乎其微。
3. 数据增强处理:，对留言主题和详情进行回译（中→英→中），再打乱留言详情的句子顺序。词语的表达具有多样化，回译之后留言以新的表达方式呈现出来，加上顺序打乱，可假设回译文本与原始文本是不同的留言内容。

2.2 问题二假设

1. 第二题热点问题挖掘，某一时段内群众集中反映的某一问题可称为热点问题，可能存在同一用户集中反应某个问题的情况，本题假设相同用户的多条留言为不同用户的留言，同一用户多次反映相同问题说明该问题亟需尽快解决，这样不易漏掉群众急切的诉求。
2. 热点问题聚类：由于人群的命名表达过于多样，难以精准定位相同问题。而群众在反映问题时往往会指出问题发生的具体地点，因此本题以特定的地点（具体到 xx 街道 xx 小区）为聚点中心进行热点问题聚类，假设非相同地点

的留言反映的不是同一热点问题，不考虑特定人群的影响，归并得到结果之后，再人工删除明显不符的留言，结果发现按照地点聚类出来的留言基本反映的是同一问题。

2.3 问题三假设

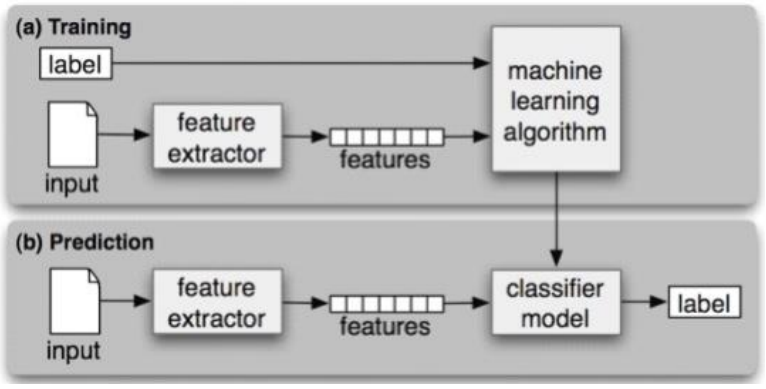
1. 本题从及时性、相关性、完整性和可解释性四个角度进行答复评价，假设四个角度之间彼此没有影响，分别进行打分，最后累积叠加得到总分；
2. 假设答复中出现书名号《》，即可认为答复内容引用了相关文献或法律法规，忽略出现书名号但并非引用文献的情况（如：只是一本课本、文学作品等），事实上，这种情况基本没有出现，说明我们的假设是合理的。

三、模型的建立与求解

3.1 问题一的模型建立与求解

3.1.1 监督式分类模型

第一题是典型的监督式分类问题，下面简单介绍一下监督式分类：



图一：监督式分类流程

在训练过程中，特征提取器用来将每一个输入值转换为特征集。 这些特征集捕捉每个输入中应被用于对其分类的基本信息 。特征集与标签的配对被送入机器学习算法中生成模型。在预测过程中 ， 相同的特征提取器被用来将未见过的输入转换为特征集。之后，这些特征集被送入模型产生预测标签。

3.1.2 数据增强处理

在全部留言中，一级标签各类的留言数之中有部分标签的数据量非常少，造成数据不平衡，这样训练模型时显而易见的问题就是数据量少的标签预测效果极差。所以需要对原始数据进行数据增强处理。我们将留言内容进行回译（中文→英文→中文），自定义翻译函数，调用百度通用翻译 API，然后加入到原始文本数据中；而仅仅回译可能会出现回译文本与原始文本大量重复的情况，于是对回译后的留言详情进行分句，调用 random .shuffle 函数打乱分句顺序。处理情况如下：

一级标签	原始留言数/条	增强处理	处理后留言数/条
城乡建设	2008	回译一次并打乱顺序	4016
环境保护	938	回译一次并打乱顺序	1876
交通运输	613	回译两次并打乱顺序	1839
教育文体	1588	回译一次并打乱顺序	3176
劳动与社会保障	1969	回译一次并打乱顺序	3938
商贸旅游	1215	回译一次并打乱顺序	2430
卫生计生	877	回译一次并打乱顺序	1754

数据增强处理之前，测试集的 F-score 最高可达 0.85；而数据增强处理之后，测试集的 F-score 高达 0.96，说明这样处理还是有很可观的效果。

3.1.3 停用、分词等预处理

我们将一级标签转换成了 ID 数字，每个一级标签用数字代替。由于我们的文本信息内容都是中文，所以要对中文进行一些预处理，包括删除文本中的标点符号、特殊符号，还要删除一些无特殊意义的常用词(stop_words)，因为这些内容对预测文本的内容没有任何帮助，只会增加计算的复杂度并且增加系统开销。然后用 jieba 库对清除停用词之后的文本内容进行分词处理。最后对文本进行截断，本文设置的词语最大长度设置为 200（超过的将会被截去,不足的将会被补 0），于是每个评论详情变成了等长的词语序列。

3.1.4 神经网络模型的架构

下面考虑分类用到人工神经网络模型长短期记忆网络（LSTM）。LSTM 是循环神经网络（RNN）的变种，LSTM 在 RNN 的梯度消失问题上进行了改进。梯度消失就是指后面时间的结点对于前面的结点感知力下降，LSTM 相对于传统 RNN 增加了记忆能力，所以在处理较长文本的问题上更有效果。

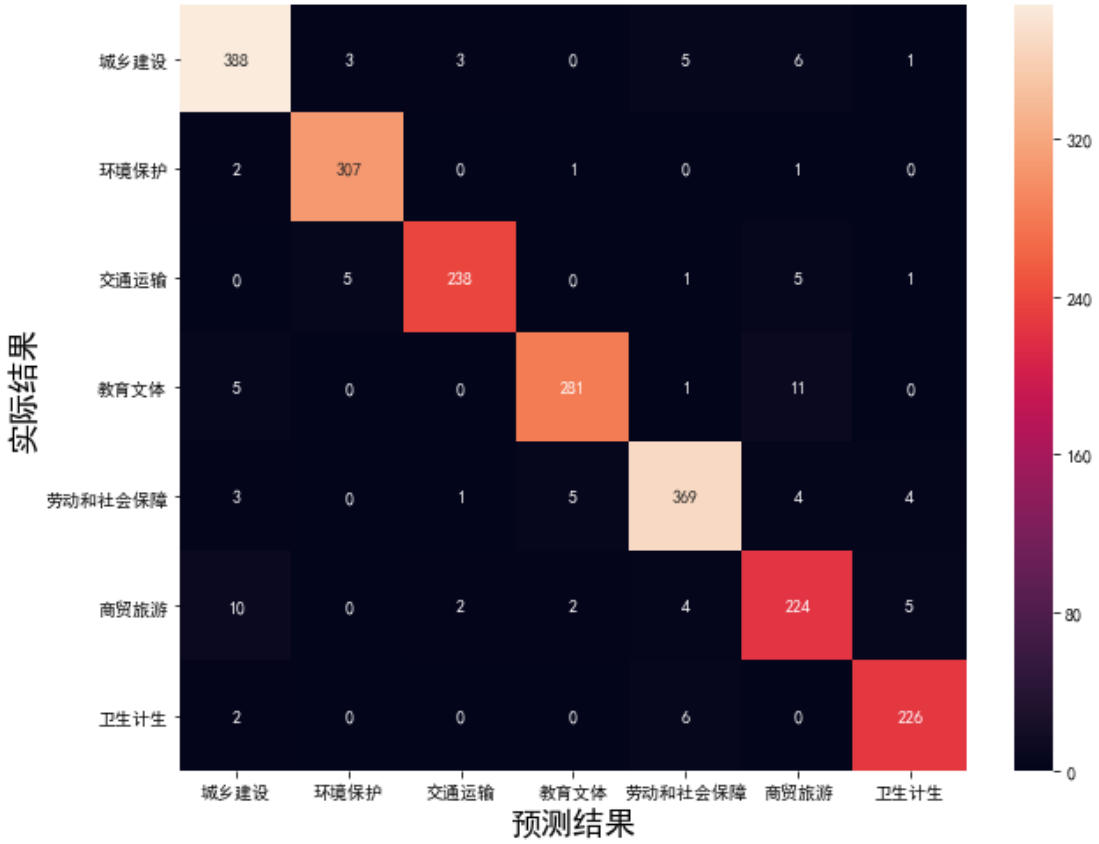
模型的第一次是嵌入层(Embedding)，它使用长度为 100 的向量来表示每一个词语。通过 embedding 层将每个词转换成词向量，维数为 100，于是代表每个留言详情的词语序列也变成了100×200的矩阵。这里也可以通过调用他人预先训练好的 word2vec 模型进行处理。

SpatialDropout1D 层在每次更新时，将输入单元的按比率随机设置为 0，有助于解决过拟合的问题；LSTM 层包含 100 个单元；输出层为 10 个分类的全连接层。由于是多分类模型，所以激活函数设置为'softmax'，损失函数为分类交叉熵‘categorical_crossentropy‘。优化函数经过试验，选择‘nadam’和‘adam’均

可以收敛，但是‘nadam’收敛更快，所以最终选择‘nadam’作为优化函数。

3.1.5 模型的运行与运行结果

首先数据集按照 9：1 分成测试集和训练集，下图是测试集测试结果



图二 各个类别上的预测结果

accuracy 0.9535647279549718

	precision	recall	f1-score	support
城乡建设	0.95	0.96	0.95	406
环境保护	0.97	0.99	0.98	311
交通运输	0.98	0.95	0.96	250
教育文体	0.97	0.94	0.96	298
劳动和社会保障	0.96	0.96	0.96	386
商贸旅游	0.89	0.91	0.90	247
卫生计生	0.95	0.97	0.96	234
avg / total	0.95	0.95	0.95	2132

图三 各个类别上的测试结果评价（包括 F-score）

3.2 问题二的模型建立与求解

第二题需要从大量的群众留言内容中提取出热点问题，因此本题运用文本聚类分析进行热点问题的挖掘。即对大规模文本集进行特征提取，提取出的命名实体和核心谓词；同时研究各个文本之间相似性，以进行事件分类和相似聚类，减少阅览大量反映相似问题的留言所耗费的时间成本和人力物力资源。

3.2.1 热点问题挖掘

本题首先对文本进行预处理。去除空白和停用词后，我们使用 jieba 进行分词，并识别词性。

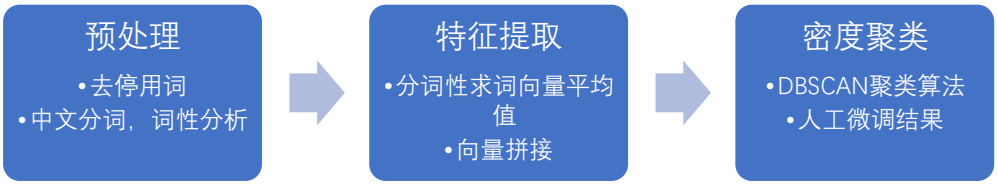
接着，我们对文本进行特征提取。关于特征提取，我们尝试了多种主流办法，包括基于 TF-IDF 的带权词向量均值，直接词向量均值，word2vec 直接文本向量模型。然而，聚类的效果并不好。之后我们考虑到评论的具体内容较长，内容较多，而标题言简意赅，噪声少，我们选择将标题作为聚类的依据，但聚类效果依旧不好。

我们分析了其中的原因：留言和普通文本不同的地方在于：命名实体数量多，种类杂。而对于预训练的 word2vec 模型，缺乏命名实体的相关数据，因此对词向量的嵌入不准确。其次，在这个案例中，统一特定事件的留言很少（与微博热搜存在大量相似转发，每个事件上千上万条微博）不同。最多的一个事件，也不到总量的 1%。最后，词向量平均值的向量维数只有 200 维，句子的信息发生了丢失。因此，我们增加了自定义词典后进行重新分词，自主训练了新的词向量。我们对每个标题的所有名词，动词，地点名词，机构名词，人名分别求词向量平均值，拼合。这样，虽然单个词向量维度从 200 维下降到了 128 维，但是句子向

量的维度从 200 维增加到了 640 维。

关于聚类，我们没有使用 K-means 算法，而是使用了基于密度的 DBSCAN 算法。这是因为 K-means 算法计算得到的每个分类过于庞大，一个分类混杂了多个事件，这与原始数据高度稀疏的分布不符。而若增加分类数，则受初值的影响太大。而 DBSCAN 算法可以将无法聚类的评论都归因于“噪声”，且不受初值选取影响，这就大大解决了数据稀疏性的问题。（使用 scikit-learn 库对数据进行聚类分析）

在聚类结束后，我们除去了聚类方差过大的集合，最后剩下了数目不大的评论。为了避免有效信息损失，我们人工对聚类结果进行了简单修正，着重于将同一类的信息归并、删除反映相同问题但命名实体不一致的个别信息。



图四 文本聚类流程图

其中主要的部分为三个：

1、数据预处理：读取用于聚类的文本信息，进行停用和分词处理；对于时间段的处理较为特殊，最早留言时间记为初始时间，留言时间用具体的留言时间与初始时间的差值来代替（具体为 x 天 x 时 x 分 x 秒）。

2、词语转向量：先使用根据原始数据训练得到的词向量模型，将每个词语转为词向量。然后将标题的词语分为名词，地点名词，人名，机构名词，动词，其他（并丢弃），分别计算各种词性词语的词向量平均值，再将得到的结果拼接为 640 维词向量。

3、聚类算法：使用 DBSCAN 算法进行聚类。具体做法是：随机选择一个数

据点作为起点，检查距离 d 内是否有至少 3 个数据点。若无，标记为噪声点。若有，则将周围的数据点作为一类。

3.2.2 热度评价指标

话题热度，正相关于该话题留言数，点赞数，点赞的时间聚集度，负相关于点踩数。为了描述时间聚集度，我们以天为最小单位，计算某个话题下留言时间的方差。因此，我们定义了这样的指标：

$$有效评论指数 = 留言数 + k \times （点赞数 - 点踩数），这里的k\ 可以取\ 0.5$$

$$时间因子 = \frac{1}{100 + ln(1 + 时间方差)}$$

$$热度 = 有效评论数 \times 时间因子$$

热度排行	热度指数	时间范围	问题描述	相关留言数	点赞数	反对数
1	6.5821	2019/7/11-2019/9/1	A 市伊景园滨河苑强行捆绑车位销售给业主	32	1	18
2	6.0066	2019/11/2-2020/1/9	A2 区暮云街道丽发新城附近违建搅拌站，噪音和粉尘严重影响居民生活	23	2	39
3	0.9868	2019/1/7-2019/4/9	A6 区景城苑小区居民多次反映小区存在着各种亟需解决的问题，却得不到正面回应	3	0	13
4	0.9839	2019/1/16-2019/10/29	A 市人才购房和租房补贴实施办法等相关问题咨询	7	0	4
5	0.7265	2019/9/5-2019/9/25	A3 区梅溪湖看云路一师润芳园小区临街门面烧烤夜宵摊油烟直排扰民	3	0	3

图五 热点问题聚类结果

如图所示，热度排行第一和第二的问题具有明显高于其他问题的热度指

数，“A 市伊景园滨河苑强行捆绑车位销售给业主”共 32 条留言，“A2 区暮云街道丽发新城附近违建搅拌站，噪音和粉尘严重影响居民生活”共 23 条留言，可见关注这两个问题的人很多，且反应时间很集中，在两个月左右。 排行第三的问题，从留言内容上看为同一用户在短时间内多次反映相同的问题，用户从 2018 年开始至今已反映 5 次，而问题拖延许久尚未解决，且反对数较多，反映了群众对此事亟需解决的迫切态度，因此在热点问题第三位是较为符合实际的结果。

3.3 问题三的模型建立与求解

针对附件 4 相关部门对留言的答复意见，我们将从回复的及时性、相关性、完整性及可解释性等角度给出评价:

3.3.1 留言答复的多角度分析

1. 留言答复的及时性: 即回复是否及时, 政务问题一般都具有很强的时效性, 回复的及时性就显得尤为重要;

序号	留言主题	留言详情
1672	三问A6区景城苑小区的问题何时能得到解决	尊敬的胡书记：原帖链接：A6区景城苑小区的问题多年
2730	四问A6区景城苑的问题何时能得到解决	尊敬的胡书记：我们曾就A6区景城苑小区存在的各种亟
3599	再问A6区景城苑小区的问题何时能得到解决	尊敬的胡书记：我们曾就A6区景城苑小区的各种问题于

例如上述关于景城苑小区的问题，同一个业主已经多次反馈，然而小区存在的问题仍迟迟得不到解决，长此以往，会使政府在部分人心里失去公信力，激化群众潜在的矛盾问题。

2. 留言答复相关性：命名实体一致，文本相似度较高

答复中至少要有一致的特定地点或人群, 并保证一致的命名实体下回复的是与留言对应的问题，而不至于出现答非所问的情况 。

例如“A3 区枫林三路向日葵旁边工地一直存在凌晨施工扰民的情况”，“A1 区桐阴里小区一直夜间施工”两个留言均反映了施工噪音扰民现象，但是对应的地点不同；相同的地点，也可能反映多种不同的问题，因此还需要根据留言与答复的文本相似度或关键词锁定来计算答复相关性；

3. 留言答复的完整性：完整的开头、结尾，包括回复中是否有文明用语，一般留言回复都要有“尊敬的网友”，“您好”，“请”等礼貌用语，观察可知这些敬语基本都有，因此完整性主要考虑以下两个方面：

词性完整性： 计算特定名词的词语密度，主要针对答复词语数较少，但是精准回答了留言的答复内容（“言简意赅”）；对言之无物，内容空洞的回复，评分较低。

词语完整性：至少出现一次留言内容中出现的特征词汇，即出现了未必体现出答复的完整；但若不出现，则答复肯定不完整。

4. 留言答复的可解释性：有论点论据支撑，根据...规定，政策等，或者引用文献（根据出现书名号来来判断）；且一般情况下，答复内容越长越详细，可解释性越高。

3.3.2 量化指标

按照打分制进行对答复的每个角度量化描述

1. 留言答复的及时性

答复大多在 15 天左右就会有答复，最快一星期，最长可达三个多月

留言与答复时间间隔	≤7 天	＞7天且≤ 15天	＞15天且≤ 30天	>30 天
分值	+2	+1	0	-1
留言数量	992	740	623	461

占比	35.227%	26.278%	22.123%	16.37%

图六 答复及时性分析

由计算结果可知，83.63%的留言在一个月以内都会得到答复，一个星期之内得到答复的占比最大，可见政府相关部门的答复工作效率较高。

2. 留言答复相关性：

实现思路：将提问提到的特征词汇存入集合 A，答复的特征词存入集合 B，研究 $card(A \cap B)$，即两个集合中重合的特征词数量。

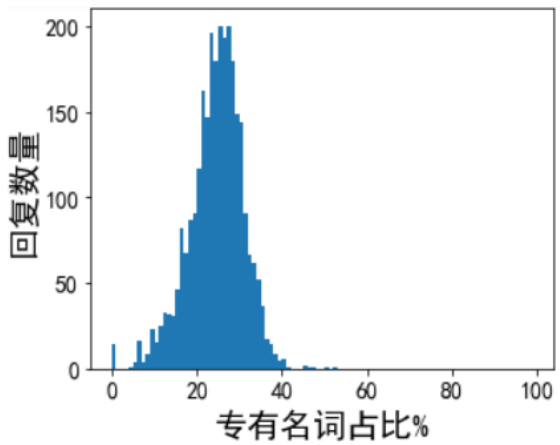
该实现方法的合理之处在于，重合的特征词数量多未必相关性高；但没有重合的特征词，说明答复与留言基本不相关。

Card($A \cap B$)	0	1-3	3 <
分值	0	+1	+2
留言数量	1286	1317	211
占比	45.6677%	46.768%	7.493%

图七 答复相关性分析

3. 留言答复完整性

(1) 词性完整性：



识别专有名词：nr-人名；nt-机构团体名；nz-其他专名

专有名词占比% = 专有名词数/答复词语数*100%

根据专有名词占比比例，我们从 20%开始截断，专有名词占比达 20%及以上得 2

分，0%-20%之间的比例记为 k，按照 10*k 进行给分（例如 15%，得 1.5 分）

比例	$0\% \leq k < 20\%$	$\geq 20\%$
分值	10*k	2
留言数量	581	2235
占比	20.63%	79.368%

图八 答复词性完整性分析

（2）词语完整性

我们选择检测如下词语是否出现：出现一次 1 分，出现二次及以上 2 分。

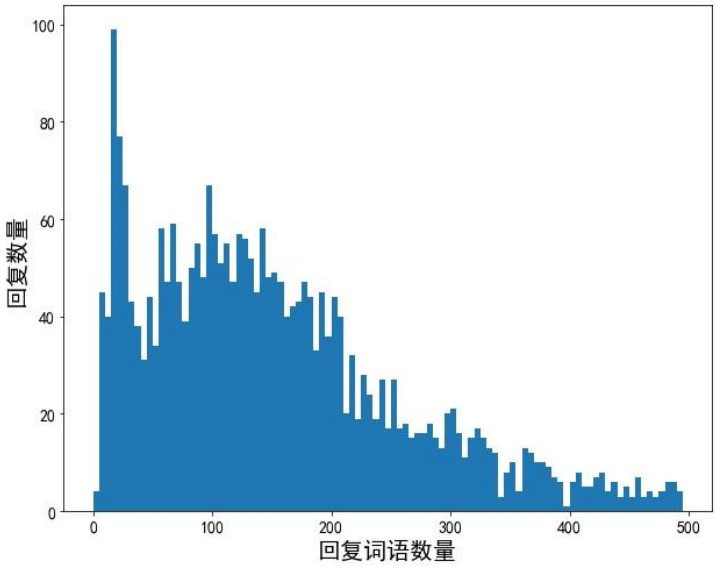
调查,处理,核实,核处,回,咨询,立案,管辖'

特征词数量	0	1	2 ≤
分值	0	+1	+2
留言数量	526	1057	1233
占比	18.679%	37.5355%	43.7855%

图九 答复词语完整性分析

4. 留言答复的可解释性

N 记为每条答复的词语数量，若$N \leq 50$，称该条答复为短回复；若$N > 50$，称该条答复为长回复；



理论支撑：出现“按照、根据、依据、依照”，“相关规定、政策、标准”等特征值；

文献引用：出现书名号《》

下面量化可解释性的指标，评级为三级：

*I*级：答复为短回复且无理论支撑和文献引用

*II*级：(1) 答复仅为长回复；（2）答复为短回复，且有理论支撑或文献引用

*III*级：答复为长回复，且有理论支撑或文献引用

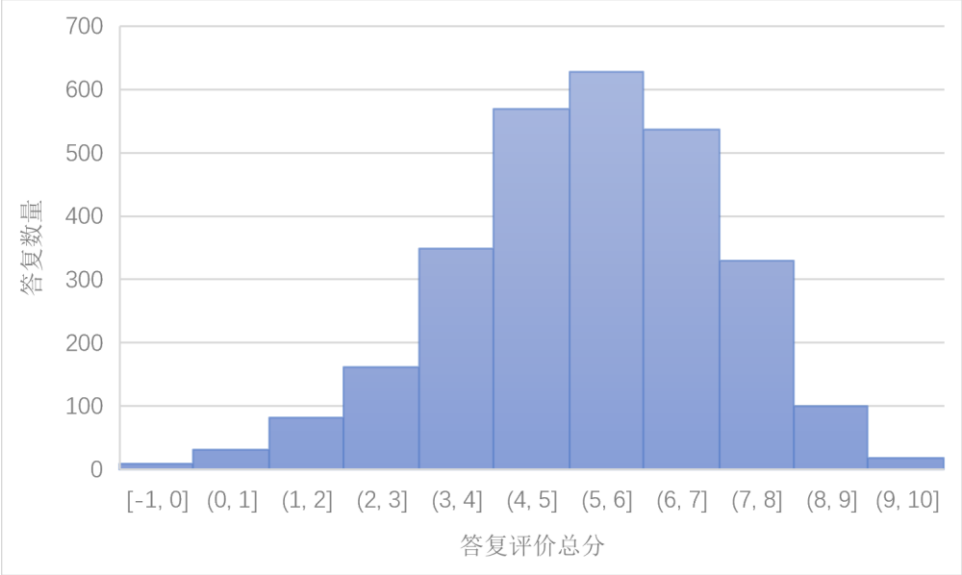
评级	*I*级	*II*级	*III*级
分值	0	+1	+2
留言数量	488	1564	764
占比	17.33%	55.54%	27.13%

图十 答复可解释性分析

3.3.3 评价总分

最后将各个指标的分数加起来，满分为 10 分，按照总分分值划分为五个等级：

评级	E	D	C	B	A
总分 g	$g < 2$	$2 \leq g < 4$	$4 \leq g < 6$	$6 \leq g < 8$	$8 \leq g$
留言数量	75	334	866	1097	413
占比	2.663%	11.86%	30.75%	38.956%	14.667%



图表结果显示，评价总分达到 6 分及 6 分以上的答复占一半以上，分数低于 2 分的仅有 2%左右。说明在我们建立的评价系统中，综合答复及时性、相关性、完整性和可解释性几个方面，政府部门的留言答复工作在一定程度上做的相当到位。下面是选择两个极端结果的评分展示。

零分结果（部分）

感谢您对我们工作的关心、监督与支持。
2016 年 6 月 12 日
已收悉
你好！2019 年 10 月 10 日
0000-00000000　　20163

满分结果（部分）

“UU0081583”您好！　关于您反映的“L5 县两丫坪社区一村民关于请求解决危房补的请求”这一问题，经我镇调查，现将相关问题回复如下。经调查：陈平旺，两丫坪镇两丫坪社区 18 组（原平安村 3 组）居民，系建档立卡贫困户，家庭贫困人口 4 人。1、关于陈平旺建档立卡贫困户的问题。陈平旺建档立卡时间为 2014 年，评定过程完全符合程序，在相关表格上均有其本人签字确认，且在建档立卡之后，结对帮扶责任人也按规定对其进行了结对帮扶走访，并非如他本人所述的毫不知情。2、关于陈平旺没有实施易地搬迁项目的问题。陈平旺已于 2016 年作为易地搬迁计划名单上报，但在搬迁选址阶段，其提供的几处选址都属于“简单位移”，不符合搬迁要求，不予通过。后社区干部也为其提供多条可行的选址建议，陈平旺都不满意不同意，遂其父亲陈异德自愿签订《退出易地搬迁退出书》，并盖有手模印，放弃享受易地搬迁政策。 3、关于陈平旺危房改造项目未获补贴的问题。陈平旺于 2017 年 10 月向两丫坪社区、村镇管理所及国土所申请危房改造建房。经村、镇两级相关部门人员现场勘查、选址放线，拟建于原宅基地前面的斜坡上，房屋出高坎 1.5 米处，位于水田内修建三个立柱，悬空用于作走廊支柱。2017 年 11 月下旬，接群众举报，陈平旺私自将房屋外移占用基本农田建房。经现场勘查：陈平旺房屋现占地 187 平方米（建筑总面积二层共 310 平方米），其中占用基本农田面积为 135 平方米（房屋建筑占地 75 平方米，空坪硬化 60 平方米）。陈平旺占用基本农田建房违反了《中华人民共和土地管理法》第 76 条和《基本农田保护条例》第 17 条的规定，属于非法建筑。 一是根据《L5 县规范农村村民建房管理办法》的通知之第九条规定：选址占用基本农田或占用农用地未办理农用地转用审批的不予批准村民建房申请。二是按照 2018 年危房改造项目政策要求，1 人不能超过 35 平方米，2 人不能超过 45 平方米，3 人不能超过 60 平方米，3 人以上人均不能超过 18 平方米。所以陈平旺不能享受危房改造的资金补助，危房改造计划才未被上报。4、关于陈平旺 100 元车费一事。2017 年 10 月份镇政府分管国土城建工作的领导王银端、国土所所长贺军达、国土所干部石辉、两丫坪社区干部吴菲、黄军长共 5 人应邀去看陈平旺家准备危房改造的宅基地。当一行人准备离去时，陈平旺硬要塞给吴菲 300 元钱说是请大家吃中饭，吴菲坚决不肯接受。两人僵持不下，陈平旺见吴菲态度坚决，只好改口说给司机 100 元车费。5、关于陈平旺请客两丫坪社区支部书记周元齐花费

上千元一事。周元齐于 2016 年 7 月份受邀在陈平旺家吃了一顿中饭，主要咨询危房改造和易地搬迁政策，陪同人有黄军长、李模佑，当时发了一包 25 元的 A1 区王炎，因周元齐不抽烟，离去时并未拿。其后批地基、口头下建房停工通知书周元齐都未直接参与，所以说请客周元齐花费上千元一事并不属实。2018 年 7 月 31 日
“UU0082061”您好，您所反映的情况已转交县住建局进行核处。2018 年 5 月 24 日尊敬的发帖人：您好！非常感谢您对我县建设工作的关心和关注。您反映“呼吁 L3 县政府拉通体育路至龙舟广场便民道路”的问题，我局已收悉，对此高度重视，并安排专人调查核实和认真办理。现将有关情况回复如下： 2018 年 3 月 4 日，中共 L3 县委 L3 县人民政府印发了《关于印发〈L3 县 2018 年度县级领导干部“六联”工作目标考核方案〉的通知》（沅委[2018]11 号）文件，对今年的经济建设及社会发展目标和任务进行了具体细化。其中城建重点项目中包含城北体育场改造项目，目前，体育场改造工程，已完成污水管网改造工程，其他改造工程方案正在进一步完善中（含相关支路建设），我局将积极支持县城体育场改造工程，力争早日复工建设。 根据《L3 县城总体规划（2003-2020）》中城市道路规划，城北片区下河道路规划有体育路、古城南路延长等道路，目前古城南路建设正在实施中，待城区支路陆续建设完成，再辅以健全、高效的城市交通管理方式，将会解决日益拥堵的交通问题。 L3 县发展人人有责。真诚欢迎您和广大网友对我县的城市建设工作积极建言献策。联系电话：0745-4224393。

由此可见，零分回复确实言之无物，内容空洞。满分回复内容详细，确实解决了群众的难题。因此，这个评分标准还是具有科学性的。

四、参考文献

[1] 陈慧, 田大钢, 冯成刚. 多种算法对不同中文文本分类效果比较研究[J]. 上海理工大学 管理学院，2019
[2] 余传明，张小青，陈雷. 基于 LDA 模型的评论热点挖掘：原理与实现[J]. 上海理工大学管理学院，2010
[3] 石凤贵. 基于 TF-IDF 中文文本分类实现[C]. 马鞍山师范高等专科学校软件工程系, 2020：51-54
[4] 孟令达，周喜. 基于区域——频道访问度的民意热点信息挖掘算法[J]. 乌鲁木齐：中国科学院新疆理化技术研究所；北京：中国科学院大学，2013
[5] 金慧峰，程振设. 基于文本挖掘的大学生网络舆情监测和预警模型[J]. 浙江 温州：浙江工贸职业技术学院，2019
[6] 王丽颖，葛丽娜，张翼鹏，王红. 增量式聚类的新闻热点话题发现研究[J]. 广西民族大学信息科学与工程学院，广西科学实验中心，2017
[7]- 派 神 -. 基 于 LSTM 的 中 文 文 本 多 分 类 实 战 [EB/OL].https://blog.csdn.net/weixin_42608414/article/details/89856566,2019-05-06.
[8]最小森林.在 Keras 的 Embedding 层中使用预训练的 word2vec 词向量[EB/OL]. https://blog.csdn.net/u012052268/article/details/90238282,2019-05-15.