“智慧政务”中的文本挖掘应用

摘要

近年来，随着网络和社交媒体的发展，政府了解民意的渠道从传统的问卷调查、电话咨询演变为市长信箱、阳光热线等网络问政的形式，各类相关的文本数据量不断攀升，给传统的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的挑战。同时，随着网络传输性能和硬件计算性能的提高，大数据、云计算、人工智能等技术迅速发展，运用网络文本分析和数据挖掘技术对社情民意信息地研究具有重大的意义。

对于问题 1，通过 re 库去除文本数据中的特数字符防止其干扰文本分类内容，将分类标签数字化，按 8：1：1 的比例切分数据集，采用具有 12 个 Transformer 层的 BERT 模型，通过 fine-tune 操作对 BERT 模型进行迁移学习操作，使其可用于中文分类任务，最后使用训练好的 BERT 模型对群众留言信息进行分类，并利用 f1-score 值进行评价。

对于问题 2，通过 jieba 中文分词工具对群众反映问题信息进行分词，并利用停用词表去除停用词。同时对问题信息进行命名实体识别，识别出地点、人群、组织合并到切词信息中。通过 TF-IDF 算法针对每个群众反应问题信息的切词文本生成一个向量，使用 PCA 将向量压缩为 500 维，利用轮廓系数地方法选择合适的 K 值，采用 K-means 对 TF-IDF 权重向量进行聚类。基于聚类结果进行热度评价，模型以留言数目作为关注度，以用户有效发文数量作为用户影响力，以点赞或反对数作为问题认可度，计算单一时间段下的问题热度，并引入冷却因子，使问题的热度可以随着不同时间段内，留言人数，点赞数，用户影响力变化而动态变化，并通过热度最高和最低的 5 条数据的热度评分，定义模型的热度区分度。

对于问题 3，针对相关性评分问题，通过衡量答复与回复之间的相似性，来表示二者的的相关性，相似性采用句向量的余弦距离来表示，使用基于 BERT 模型的句向量和基于 wor2vec 结合位置编码的句向量来计算余弦距离,并通过对比发现基于位置编码的相关性评分区分度更强。针对完整性评分的问题，通过基于频率和基于 LDA 两种方法提取出答复意见的规范词，并使用基于频率加权和 LDA 加权的评价方法对答复完整性进行评分。针对可解释性评分的问题，通过对答复意见进行因果关系抽取并对原文本和因果关系文本建立词典，利用词典计算因果关系抽取出的文本中互斥内容占原文本的比例，进而对可解释性进行评价。

关键词：文本分类 K-means 算法 热度评价 因果关系抽取 相似度

一. 挖掘目标

本次数据挖掘的目标，是利用自然语言处理与文本挖掘的技术，解决对政务群众留言及评论的分类，分析与评价问题，具体包括：

1、利用 Bert 模型为基础，建立留言内容的一级标签分类模型，对问政平台的群众留言进行分类，并使用 F1-score 为指标，对分类结果进行测评，在公布的数据集上，取得了 93.7%的 F1-score 成绩。

2、利用 K-Means 聚类算法与热度评价模型，对留言进行聚类，在考虑关注度，认可度，用户影响力的基础上，综合基于时间的冷却因子，给出问题热度指数与热度排名。

3. 针对相关性的问题，利用 BERT 与位置编码两种方法形成句向量，通过句向量的余弦相似度，表示留言与回复的相关性。针对完整性评分的问题，通过基于频率和基于 LDA 两种方法提取出答复意见的规范词，并使用基于频率加权和 LDA 加权的评价方法对答复完整性进行评分。针对可解释性评分的问题，通过对答复意见进行因果关系抽取建立词典，利用词典计算因果关系抽取出的文本中，互斥内容占原文本的比例，进而对可解释性进行评价。

二.分析方法与过程

2.1 群众留言分类问题的解决过程

2.1.1 数据预处理

1) 留言信息的内容处理

由于留言中会出现网页链接等会干扰文本分类的内容，所以首先需要对留言文本进行处理，将其中的非文本信息去除。调用 python 的 re 库中的 sub 函数去除掉文本数据中的标点、转义字符和其他特殊字符，将处理后的文本放在 clear1.xlxs 文件中。

2) 分类标签数字化

针对根据用户留言信息将用户留言分类到对应的一级分类标签下的问题，考

虑到一级分类标签以中文文本的形式给出，而各种分类算法需要数字化的标签，故考虑按照等间隔的标准将“城乡建设”等 7 个一级标签转化为[1,2,3……7]的数字化标签形式，为后续分类进行做准备。

3) 训练集，验证集，测试集的划分

在训练的过程中，将留言的主题与留言详情合并后，作为待分类文本。对附件 2 中的所有的数据，按照 8：1：1 比例，划分为训练集，验证集，测试集。数据的格式都是将数据按照留言内容，标签的形式，测试集标签全部相同，且标签数值不属于分类标签集合。

2.1.3 Transformer [1]编码器

在完成文本分类任务时，首先选用了 BERT 模型，BERT 的结构是多层双向 Transformer 编码器。该编码器的输入是 one-hot 编码的词向量；在 BERT 中，输出是 768 维的词向量，通过获得每个词的词向量，进一步完成之后的任务。

Transformer 编码器的关键，是 self-attention 机制。self-attention 机制定义了每一个输入向量 X_i 对应的 query(q_i,)和 key(k_i)-value(v_i)键值对(Xi 是原始词向量的线性变换)，通过三个权重矩阵 Wq, Wk, Wv,来获得。这三个权重矩阵，都是针对上下文所有词向量的权重。即：

$$q_i = W_q X_i, \qquad k_i = W_k X_i, \qquad v_i = W_v X_i$$

之后根据q_i,k_i就可以得 Xj 对 Xi 的权重$\alpha_{i,j}$:

$$\alpha_{i,j} = softmax(\frac{q_i k_j}{\sqrt{d_k}})$$

所以，第 i 个词对应的词向量Y_i就是:

$$Y_i = \sum_j \alpha_{i,j} v_j$$

为了获得更准确的结果，BERT 中的 transformer 编码器实际使用的是 Multi-head self-attention, 即将 q_i , k_i , v_i 分别分解为 $q_i^1 q_i^2 \ldots q_i^s$， $k_i^1 k_i^2 \ldots k_i^s$, $v_i^1 v_i^2 \ldots v_i^s$，按照如上所述相同的方法，得到Y_i^s，再将 s 个输出向量拼接并作线性变换，得到最后的输出结果，即：

---

[1] Ashish Vaswani, Noam Shazeer, NikiParmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and IlliaPolosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

$$Y_i = Concat(Y_i^1, Y_i^2, \dots, Y_i^s)W_o$$

在 BERT 中，s 取 12。

Self-attention 机制最主要的特点，就是在做到并行计算的基础上，计算一个词向量的时候，考虑所有词的影响，这是与 LSTM 区别最大的一点。BERT 正是依靠多层双向 transformer 编码器，获得了较为准确的词向量表示。

2.1.4 BERT 模型

本题中选用的，用于处理中文的 BERT 模型（BERT-Base, Chinese）[2]，具有 12 个 Transformer 层，每层编码器 Multi-head Self-attention 的头的数量是 12，每一层之间的词向量是 768 维的向量。其算法大致流程如下：

1) 数据输入

对于中文分类任务，实际输入 BERT 模型的，是留言中的每一个字根据给出的词典，经过 one-hot 编码方式形成的向量。因为中文的词汇相较于英文词汇的数量过于巨大，若使用词作为 one-hot 编码的基础，会造成输入向量维度过大，不便于计算, 因此对字进行 one-hot 编码。

2) 预训练

BERT 模型的预训练方法有两种：Masked LM 与 Next Sentence Prediction（NSP）。Masked LM 是将句子中的 15%词遮盖，训练模型根据上下文预测被遮盖的词语。NSP 任务是给出两个句子，训练模型判断这两个句子语义是否连贯，能否构成先后句的关系。预训练的过程在下载模型之前已经完成，本题目处理过程中不需要再次训练。

3) fine-tune 操作

经过了预训练的 BERT，用于中文分类任务的模型，需要对模型进行迁移学习 fine-tune 操作。在对 BERT 模型做 fine-tune 操作的时候，不需要改变模型原有结构，用新的数据重新训练最后一层用于分类的层，并对原有模型参数做微调。

2.2 问题二分析方法与过程

2.2.1 数据预处理

2.2.1.1 命名实体识别

由于群众留言信息中的时间和地点信息是非常重要的信息，对聚类和抽取留

[2]https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

言信息的地点及人群的问题有着极大的帮助，故考虑在文本预处理阶段利用命名实体识别的方法将群众留言信息中出现的地点、人群、组织抽取出来，保存在文件中等待后续利用。

命名实体识别的方法大体可以分为：基于规则的方法、基于特征模板的方法和基于神经网络的方法三类，这里使用基于规则的方法和基于神经网络的方法中的基于字的 BiLSTM-CRF 的方法[10][11]进行命名实体识别，具体模型如下：

（1） 基于规则的方法：

使用字符串处理的方式，利用字符的 ascii 码将形如 X 市/区/县（$X \in A{\sim}Z$）、x_i 市/区/县（$X \in A{\sim}Z$，$i \in 1{\sim}9$）识别出来，作为地点的实体识别存储在文件中。

（2） 基于字的 BiLSTM-CRF 的方法：

首先，以句子为单位，将一个含有 n 个字的句子记作一个句子序列：

$$X=(x_1, x_2, \dots, x_n)$$

其中 xi 表示句子的第 i 个字在字典中的 id 对应的 one-hot 向量，维数是字典大小。

模型的第一层为 look-up 层，利用预训练或随机初始化的 embedding 矩阵将 xi 由 one-hot 向量映射为低维稠密的字向量，得到维数为 embedding 维度的 xi，并在下一层输入前，利用 dropout 随机丢弃一些点来缓解过拟合。

模型的第二层为双向 LSTM 层，用来自动提取句子的特征。将一个句子的各个字的 char embedding 序列$(x_1, x_2, \dots, x_n)$作为双向 LSTM 各个时间步的输入，再将正向 LSTM 输出的隐状态序列（$\overrightarrow{h_1}, \overrightarrow{h_2}, \dots, \overrightarrow{h_n}$） 与反向 LSTM 的$(\overleftarrow{h_1}, \overleftarrow{h_2}, \dots, \overleftarrow{h_n})$在各个位置输出的隐状态进行按位置拼接$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \in R^m$，得到完整的隐状态序列$(h_1, h_2, \dots, h_n) \in R^{n \times m}$

经过 dropout 后，接入线性层，将隐状态向量从 m 位映射到标注集的标签数 k 维，从而得到自动提取的句子特征，记作矩阵$P = (p_1, p_2, \dots, p_n) \in R^{n \times k}$。其中$p_i \in R^k$的每一维$p_{ij}$为字$x_i$分类到第 j 个标签的打分值。此时若对 P 进行 softmax，相当于对各个位置独立进行类分类，无法利用已标注过的信息，故需接入一个 CRF 层进行标注。

模型第三层为 CRF 层，用于进行句子级的序列标注。CRF 层的参数为一个

（k+2）×（k+2）的矩阵A，A_{ij}表示从 i 个标签到第 j 个标签的转移得分，进而在为一个位置进行标注的时候可以利用此前已经标注过的标签，其中加 2 因为需在句首尾添加起始和终止状态。若记某一长度为句子长度的标签序列$y = (y_1, y_2, \ldots, y_n)$，则模型对于句子 x 的标签等于 y 的打分为:

$$score(x, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1},y_i}$$

再利 Softmax 得到归一化后的概率为:

$$P(y|x) = \frac{\exp(score(x, y))}{\sum_{y'} \exp(score(x, y'))}$$

根据所得概率即可得到句子 x 的标签为不同 y 时的概率，从而得到 x 最有可能的标签，进而判断 x 是否为所要识别的命名实体。

命名实体识别步骤如下：

（1） 将群众留言信息从附件 3 中逐条依次读入。

（2） 对每条留言信息进行命名实体识别，识别其中包含的地点、人群、组织的信息。

（3） 将针对每条留言信息进行命名实体识别得到的命名实体及对应的原留言信息写入到文件中。

2.2.1.2 TF-IDF 算法

在利用问题一的中文切词方法对群众反应问题信息进行切词后，将每条群众反映问题信息的命名实体识别结果与切词结果进行合并后得到新的切词结果。为使其可以应用于后续的 K-means 聚类，需要将这些词语转化为向量。这里采用 TF-IDF 的算法，将群众反应问题信息转换为权重向量，具体算法如下：

第一步，计算词频，即 TF（TermFrequency）。

TF=某词在文本中出现的次数

为防止 TF 偏向长的文件，将 TF 进行归一化：

$$TF=\frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}}$$

或

$$TF=\frac{\text{某个词在文本中的出现次数}}{\text{该文本出现次数最多的词的出现次数}}$$

第二步，计算逆向文件频率，即（Inverse Document Frequency），由总文件数目除以包含该词语的文件的数目的对数得到。包含某词语的文件越少，IDF 越大，该词语的类别区分能力也越强，公式如下：

$$IDF=\log（\frac{\text{语料库的文本总数}}{\text{出现该词的文本数}+1}）$$

第三步，计算 TF-IDF 值（TermFrequencyDocumentFrequency）。

$$TF\text{-}IDF=TF\times IDF$$

由上述式子求得的 TF-IDF 与一个词在群众反映问题信息中出现的次数成正比，且某个词对分类作用越高，其 TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，对每个文本的生成一个向量，再对向量进行 PCA 降维，得到最终的用于 K-means 的文本。

2.2.1.3 生成词向量

生成基于 TF-IDF 词向量的具体步骤如下：

（1）使用问题一中的中文分词方法对群众反应问题信息进行切词。

（2）考虑到利用 jieba 分词时存在“A 市”等应该作为整体切分的词被切分成“A“和”市“的形式，以及人群、地点和组织在聚类中属于对结果有重要影响作用的词的原因，将在命名实体识别步骤中识别出来的命名实体和对群众留言信息进行 jieba 切分词得到的分词结果进行合并。

（3）在新的分词结果上使用 TF-IDF 的算法，针对每一个群众留言信息生成一个词向量。

（4）由于直接使用 TF-IDF 算法得到的词向量维数过大，若全部使用会占用很大的机器性能和时间，故利用线性降维方法 PCA 对生成的词向量进行降维，将其降至 500 维得到新的词向量。

2.2.2 群众反映问题聚类

生成群众留言信息的 TF-IDF 权重向量后，根据每条留言信息的 TF-IDF 向量，对群众留言反映问题信息进行聚类。这里采用 K-means 的方法将留言反映问题分为 650 类。

2.2.2.1 K-means 聚类的原理

设 X=$\{x_1, x_2, \dots, x_i, \dots, x_n\}$为 n 个 d 维数据点组成的数据集，其中$x_i \in R^d$，K-means 聚类将数据集 X 划分为簇 C=$\{c_k, i = 1,2, \dots, K\}$。其中，每个划分$c_k$代表一个类，每个类有一个类别中心$\mu_i$。相较于欧式距离，余弦相似度在判断文本之间的相似性中有着更好的表现，故选取余弦相似度作为相似性和距离的判断准则，计算该类内各点到聚类中心μ_i的余弦距离和：

$$J(c_k)=\sum_{xi \in c_k} \frac{xi \times \mu_k}{||x_i|| \times ||\mu_k||}$$

聚类目标是使各类的总的距离平方和 $J(C)=\sum_{k=1}^{K} J(c_k)$最小，聚类中心$\mu_k$取$c_k$类各数据点的平均值。

2.2.2.2 利用轮廓系数选择类别数

轮廓系数的定义：

对于某一点 i 来说:

a(i) = average(i 向量到所有它属于的簇中其它点的距离)

b(i) = min (i 向量到与它相邻最近的一簇内的所有点的平均距离)

则 i 向量的轮廓系数为：

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

易知轮廓系数的值介于[-1,1]，越趋近于 1 代表内聚度和分离度都相对较优。故在选择聚类的类别个数时，通过在[0,1000]区间内，以 10 作为取值间隔，选点计算轮廓系数并绘图，选择其中轮廓系数最接近 1 的点作为最终的 k 值。

2.2.2.3 K-means 聚类的具体步骤

（1）在区间[0,1000]内等距离选择 100 个点作为 K-means 的聚类类别数。

（2）对每一个 K 值，从 X 中随机选择 K 个元素作为 K 个簇的初始中心点。

（3）计算剩下元素到 K 个簇中心的余弦相似度，将元素划分到相似度最高的簇中。

（4）根据上一步的划分结果，将每个簇内点取算术平均值得 K 个簇的新中心。

（5）将 X 中元素按照新的中心重新聚类。

（6）重复第四步，直到聚类结果不再变化。

（7）对稳定的聚类结果计算该 K 下的轮廓系数。

（8）选择轮廓系数最接近 1 的 K 值作为聚类的最终 K 值。

（9）对该 K 值重新进行（2）~（6）步后，输出分类结果。

2.2.3 热度评价模型

由于热点问题是指特定时段内，群众集中反映的问题，所以需要限制时段,并考虑特定的问题热度随时间的变化情况，在本模型中，设特定的时段为 t。

1) 问题关注度

问题关注度，类比[7]中提出的，基于 TF-PDF 思想的，用于网页新闻热点挖掘中的“报道关注度”的概念，这里将用户的评论的数量类比媒体的报道数量，定义问题在时间段 t 内的关注度为$f(j,t)$

$$f(j,t)=\frac{D_j*(t)}{\sqrt{\sum_{i=1}^{C} D_i*(t)}}e^{\frac{D_j*(t)}{N(t)}}$$

$D_j*(t)$是话题 j 在 t 时间段内的有效评论数目(一个用户多次评论算一条)，*N(t)* 是在 t 时间段内，有效评论的总数目，C 为总的问题数目（聚类的数目），$\frac{D_j*(t)}{\sqrt{\sum_{i=1}^{C} D_i*(t)}}$

类比[7]中的标准话题频度，这一项设置的目的，是为了避免单个问题的评论数目对问题关注度造成过大的影响。后边使用自然对数 e 的原因，是扩大评论数量不同的问题之间的评分差异，有利于最后的热点问题排序。

2) 留言用户影响力

此外，考虑留言者影响力权重。[3]中提出的用户影响力模型,主要针对微博平台上的发文用户，主要考虑发文数量和用户是否为 VIP 用户，在本问题中，主要考虑其在一段时间内的有效留言数 n(相同类别下，一个用户的多条留言算一条留言)， 若某用户留言频率较高，则他的留言被其他群众关注的可能性越高，所反映的问题，成为热点问题的概率越大。定义用户影响力 h 为：

$$h= \log(n+1)$$

之所以使用对数，是因为用户影响力相对问题关注度是次要因素，通过对数减弱其影响。对于一个包含不同用户的发文主题 i 来说，留言用户影响力因子就是：

$$H(j,t) = \frac{\sum_{k}^{M_j} h_k}{\sum_{i}^{M_i} h}$$

其中，*H(j,t)*是第 j 类评论在 t 时间内的用户影响力，M_i是第 i 类问题中的有效留

言人数，

3) 认同度:

衡量一个问题是否是热点，还要考虑该问题是否能反映群众的集中意见，而赞同数或反对数则可以反映，群众在看到留言的时候，个人对这一问题的看法。同时某些用户看到想要反应的问题已经有人提出，点赞之后就不再留言，赞同数也在一定程度上反映了潜在的留言条数。因此，定义留言的认同度为一条留言的点赞数和反对数的差值，即话题 j 的第 k 条有效留言的认同度为:

$$g_j^k = ag_j^k - \mathrm{dag}_j^k$$

其中，赞成该留言的次数为ag_j^k,反对该留言的次数为dag_j^k，因此一个问题的认同度是：

$$g(j,t) = \frac{\sum_k g_j^k}{\sum_j \sum_k (ag_j^k + \mathrm{dag}_j^k)}$$

4) 综上所述，固定时间内的热点问题模型为：

$$F(j,t) = f(j,t) * H(j,t) + cg(j,t)$$

之所以$f(j,t)$与$H(j,t)$之间是乘法，与$g(j,t)$之间是加法，是因为用户影响力与一类问题下的评论有关，一定会程度上可以作为问题关注度的一个权重因子；而认同度则反映了还有多少潜在的，没有直接写出的留言数目，所以两者之间是加法关系,但由于二者的确有差别，增加一个参数 c，加以区分。

5) 实时热度积累的热点问题模型

由于热点问题具有时效性，一类问题下，若留言之间的时间间隔过长，则不满足“群众集中反映”这一热点问题的特点。因此，在之前模型的基础上，引入冷却因子λ[6],并将时间段 t 切分为$t_1, t_2, \ldots t_i$，考虑时间 t 内的热度累积公式：

$$F(j,t) = F(j,t_i) + \lambda F(j,t_{i-1})$$

在计算过程中，以月为时间单位，进行冷却，只考虑对应问题有留言的月份，若问题 j 在月份 t 内没有评论，则问题的热度不会在月份 t 内冷却，直到下一个有留言的月份才会冷却，在第三部分中，给出了不同冷却因子对应的结果。

6) 模型评价

模型的评价指标，使用[7]中提出的话题区分度指标，考察热度最高的 N 类问题与最低的 N 类问题之间的差距

$$\mathrm{Dis(N)} = \frac{\sum_i^n (\mathrm{F_{hot(i)}} - \mathrm{F_{cold(i)}})^2}{\sum_i^n \mathrm{F_{hot(i)}}^2 + \sum_i^n \mathrm{F_{cold(i)}}^2}$$

Dis(N)越大，说明热度评价模型的区分度越好。

2.3 问题三分析方法与过程

2.3.1 相关性

2.3.1.1 对群众留言信息进行中文分词

在对群众留言信息进行挖掘分析之前，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，群众留言信息以中文文本的方式给出。考虑到群众留言的标题对分类结果也有一定的影响，故先将留言主题合并到留言详情中再进行后续处理，提高分类的准确率。为了便于将中文文本信息转化为计算机能够识别的数字化信息，先对群众回复信息进行中文分词。本文采用 python 的 jieba 库的精确模式进行分词。jieba 库分词采用基于前缀词典的高效词图扫描方法，进而生成句子中汉字的所有可能的成词情况得到有向无环图（DAG），同时利用动态规划的方法找出最大概率路径，进而找出基于词频的最大切分组合，对于未登录词（不在词典中的词）使用 HMM 进行新词发现，得到更好的中文分词效果。

2.3.1.2 去除停用词

上一步 jieba 分词结束后，得到留言主题和留言详情的合并文本的分词信息。但是分析分词结果发现，分词结果中包含大量与分类没有直接关联的分词信息，如：量词、语气词、连词等。针对这个现象，本文使用了权威的停用词表对分词结果进行去除停用词的处理，防止过多与分类无关的词影响分类结果，进而提高分类的准确性。

2.3.1.3word2vector 算法

在对群众留言信息进行分词后，需要将分词后的数据转化为向量，以供数据挖掘使用。这里采用 word2vector 算法[9]，通过神经网络机器学习算法来训练 N-gram 语言模型，并在训练过程中求出 word 所对应的 vector。Word2vector 算法的具体原理如下：

（1） 使用 skip-gram 训练词对

skip-gram 首先设定一个 skip_window 值记录该词选取它的上下文的单词的数量，利用该单词和选中的上下文的词形成词对作为训练数据，作为 quick 的训练标签。

（2）将 word 压缩至 k 维空间，并向量化

将语料中的词转化为向量。首先，使用语料中不重复的单词构建字典，然后对字典中每个单词进行 one-hot 编码，将单词转化为词向量。但是这样的词向量存在两个问题：（1）不能判断任意两词之间的相关性（2）在词汇表非常大时，过高的维度将导致大量计算资源的浪费。

基于上述问题，这里采用 Hinton 提出的映射到 K 维向量的方法，即：用一个简单的多分类的神经网络来训练生成 K 维向量。其中神经网络以 n 个对词进行 one-hot 编码得到的向量作为输入，输出也为词的 one-hot 编码向量，隐含层为 k 个神经元节点，包含 n*k 的权重矩阵，该层可将高维词向量压缩为 k 维向量，再经 softmax 输出 n 维的预测出的概率向量，优化残差函数，训练权重系数。最终训练结束后，这个 n*k 的权重矩阵的每一行即为对应单词的 k 维向量。

2.3.1.4 生成 word2vector 向量

生成 word2vector 的具体步骤如下：

（1）使用 word2vector 算法，每篇文章中的每个单词得到 256 维词向量。

（2）对每篇文章中切词结果中的所有词对应的词向量取均值得到表示该篇文章的向量。

（3）考虑到得到的词向量中各分量可能出现负值，而 LSTM 神经网络的输入值必须为正，对词向量进行归一化处理，将各分量归一化到[0,1]区间内。（是否需要加个式子）

2.3.1.4 相关性评价：

为了衡量答复与留言之间的相关性，通过衡量答复与回复之间的相似性，来表示二者的的相关性，又因为无论是字词或是语句，段落，都可以由词向量表示，而衡量向量之间的相似性，通常选用余弦距离。因此，先将留言与回复表示为句向量，再计算答复与留言的句向量的余弦距离，即：

$$\mathrm{dis}(\boldsymbol{e},\boldsymbol{r}) = 1 - \frac{\boldsymbol{e}\boldsymbol{r}}{\|\boldsymbol{e}\|\|\boldsymbol{r}\|}$$

$\boldsymbol{e},\boldsymbol{r}$分别是留言和回复的句向量，为了便于观察结果，定义相似度为：

$$sim(\boldsymbol{e},\boldsymbol{r}) = 1 - dis(\boldsymbol{e},\boldsymbol{r}) = \frac{\boldsymbol{er}}{\|\boldsymbol{e}\|\|\boldsymbol{r}\|}$$

通过公式可知，该数值越大，说明向量的余弦值越大，即留言与回复之间的相关性越强。

为了得到留言与回复的句向量表示，分别采取基于位置编码句向量和BERT模型的句向量表示。

1) 基于位置编码句（Positional Encoding，PE）向量的相关性

利用 word2vec 得到留言或回复的句子中每个词的向量表示，采用 Sukhbaatar 等提出的位置编码[8]的方式，获得句向量，其中，第 j 个词向量的位置信息向量为l_j，该向量的第 k 维是：

$$l_j^k = \left(1 - \frac{j}{J}\right) - \frac{k}{d}\left(1 - \frac{2j}{J}\right)$$

其中，J 是单词的总数，d 是词向量的维度。

则句向量为：

$$\mathrm{Y} = \sum_j l_j * X_j$$

其中，X_j是第 j 个词的词向量。得到留言和回复的句向量Y_1, Y_2后，计算两个句向量之间的相似度，作为相关性的衡量

2) 基于 BERT 模型的句向量的相关性

将数据中的链接，空格，回车符去除，并将留言的多条语句合并为一条语句，回复同样作此处理，利用 python 第三方库：bert-as-service[3]，计算回复与留言的句向量，计算两个句向量之间的相似度，作为相关性的衡量

2.3.2 完整性

2.3.2.1 完整性定义

一条答复意见的完整性定义为这条答复意见满足答复规范格式的程度。结合政府部门《关于规范信访事项办理回复格式的通知》中的相关要求，答复意见的规范格式必须包含如下五个部分：网友的称呼、网友留言总结、情况说明与答复详情、感谢语以及答复时间。

2.3.2.2 规范词提取

[3]https://github.com/hanxiao/bert-as-service

基于上述答复意见完整性的定义，从附件 4 中选取了 20 条答复意见组成“优秀答复”数据集。在此基础上，我们采用了以下两种方法提取出来答复意见的规范词。

（a）基于频率的规范词：在对“优秀答复”数据集进行数据清洗和切词操作之后，计算每个词在数据集上的频率。根据所求得的概率进行排序，取频率较高的 20 个词作为规范词。第 i 个词的频率值F_i的计算方法如下：

$$F_i = \frac{\text{包含这个词的答复条数}}{\text{数据集中答复总数}}$$

（b）基于 LDA 的规范词：LDA 是一种文档主题生成模型，也是机器学习中特征抽取的重要方法。LDA 模型由 Blei,Ng,Jordan 于 2003 年提出，旨在通过无监督学习的方法从文本中挖掘出隐含主题。LDA 模型假设一篇文章是一些主题构成的概率分布，一个主题是由一些词语构成的概率分布。LDA 模型采用 Gibbs Sampling 算法来学习其概率分布中的参数。其中，Gibbs Sampling 算法运行思路是每次选取概率向量的一个维度，给定其他维度的变量值 Sample 当前维度的值，不断迭代，直到收敛得到 LDA 所要估计的参数。

对于“优秀答复”数据集，采用 LDA 模型进行主题数为 1、主题词为 20 的主题抽取操作，所求得的主题词即为完整性的规范词。同时，LDA 模型还求得了每个主题词的权重L_i

2.3.2.3 完整性评分

在计算得完整性的规范词后，采用如下两种方法对答复完整性进行评分.

（a）基于频率加权的评价方法：对于附件 4 中的每一条答复意见，根据规范词及其频率计算完整性评分的公式如下：

$$V_i = \frac{S_i}{\sum_{j=1}^{n} F_j}$$

其中,F_j表示第 j 个规范词的频率，V_i表示第 i 条答复意见的完整性评分，S_i表示第 i 条答复意见中出现的规范词的F_i总和。

（b）基于 LDA 加权的评价方法：对于附件 4 中的每一条答复意见，根据规范词及其 LDA 权重计算完整性评分的公式如下：

$$V_i = \frac{A_i}{\sum_{j=1}^{n} L_j}$$

其中，L_j表示第 j 个规范词的 LDA 权值，V_i表示第 i 条答复意见的完整性评分，A_i表示第 i 条答复意见中出现的规范词的L_i总和。

2.3.3 可解释性

2.3.3.1 可解释性的定义

答复意见的可解释性，通常指的是答复的意见中某观点的提出是否在该答复中存在相对应的解释，如：答复中提出某个问题暂时无法解决，而有可解释性意味着答复中对该问题为什么无法解决做了解释，例如，是根据某定义或法规致使无法解决。考虑到具有可解释性的句子之间存在较强的因果关系，故通过因果关系抽取来抽取答复意见中具有可解释性的部分，作为可解释性分析的中间数据。

2.3.3.2 因果关系抽取

因果关系抽取实现了从文本中抽取出表示因果的句子（通常采用事件抽取实现），进而在抽取句子上抽取出因果关系的原因句和结果句的目标，其方法大致可分为模式匹配方法、统计学习方法、深度学习方法三大种类。这里采用了模式匹配的方法，分别通过“由果溯因配套式”、“由因到果配套式”、“由因到果居中式明确”、“由因到果居中式精确”、“由因到果前端式模糊”、“由因到果居中式模糊”、“由因到果前端式精确”、“由果溯因居中式模糊”、“由果溯因居端式精确”9 种模式进行模式匹配进行答复意见文本的因果关系抽取，主要步骤如下：

（1） 因果知识库的构建。因果知识库的构建包括因果连词库，结果词库、因果模式库等。

（2） 文本预处理。包括对文本进行噪声移除，非关键信息去除等。

（3） 因果事件抽取。包括基于因果模式库的因果对抽取。

（4） 事件表示。

（5） 事件融合。

（6） 事件存储。将识别出的因果关系保存在文件中。

2.3.3.3 可解释性评分

考虑到一个句子是否具有可解释性，取决于该句内部或该句与相邻句之间是否存在因果关系。故基于因果关系抽取结果，定义某一答复意见的可解释性评分

如下：

$$可解释性=\frac{因果关系抽取得到文本数}{答复意见总的文本数}$$

因为对一个同一答复意见文本进行模式匹配式的因果抽取可能会得到文本部分重叠的两条不同的因果事件，将因果关系抽取到的文本长度简单相加之和与答复意见总的文本长度做比例无法得到正确结果，故考虑使用以下算法：

（1） 对每条答复意见进行切词处理，将切词结果储存到字典中，并将 value 值赋为 0，计算总切词数为 sum1。

（2） 初始化因果关系抽取文本中的互斥的切词个数为 sum2=0

（3） 提取因果关系抽取得到的文本的切词数据，对每一切词，若该切词已存在在字典中，且字典中对应 value 值为 0（表示该词在之前的因果抽取文本中未出现过），令 sum2=sum2+1，并将字典中该词对应的 value 更新为 1。若该切词不在字典中，则在字典中增加该词项，并将 value 值记为 1，同时令 sum1=sum1+1，sum2=sum2+1.

（4） 处理完所有因果关系抽取得到的文本后，计算 sum1/sum2 即得该答复意见的可解释性。

三. 结果分析

3.1 群众留言分类结果分析

使用 BERT 模型，对群众留言进行分类，按照 8：1：1 划分训练集，验证集，测试集，并按字进行分词后，输入 BERT 模型，评测信息如表 3.1，

表 3.1 BERT 在验证集与测试集的分类结果

	验证集	测试集
F1-score	93.9	93.1
精确率	94.2	93.3
召回率	93.6	92.9
准确率	94.5	93.4

从结果可以看出，BERT模型的整体分类结果较好，训练与测试的结果差距不大，说明 BERT 模型通过预训练与 fine-tune 操作，具有良好的学习能力，没有出现过拟合现象。

表 3.2 群众留言具体分类结果

一级标签名称	数据条数	正确分类条数	正确率
劳动和社会保障	175	164	0.937
卫生计生	92	84	0.913
商贸旅游	121	107	0.884
教育文体	168	161	0.958
环境保护	85	78	0.918
交通运输	69	66	0.957
城乡建设	212	197	0.929

具体分类结果如表，从表 3.3 中可以看出，有 6 类的分类正确率都在 90%以上，总体的分类准确率较高，只有商贸旅游类的结果相对不理想，商贸旅游类被错误分类的 14 条数据中，有 6 条都被错误标注为“城乡建设”，从中选取部分数据，如表，从表中可以看出，模型看到了句中“建筑”，“竣工”等词语，将数据错误地分类。除此之外，还有 3 条数据被错误地标为“交通运输”，是受到“客运车”，“高速公路”，“车费”等词语的影响。不过，这类数据的分类相对困难，如果没有专业商业或建筑学背景，按照人类的认知也非常容易出错。

表 3.3 商贸旅游错误分类数据举例

留言内容摘要	错误分类标签
华嘉地产公司汇金城二期项目于……其高层**建筑**按标准按购房合同应该全部为钢化玻璃，结果高层玻璃是普通玻璃。	城乡建设
当时山水名苑二期推迟交房……现在想问下山水名苑二期（**工程竣工**验收备案表）或者(**建筑工程竣工**验收合格)的具体时间是什么时候	城乡建设
……2013 年年底起，怀通**高速通车**后，	

岩垅至黔城班车路线改道，**班车**行至界里坪后，直接上高速连接线到黔城滨江路岩垅班车终点站，路程由原来的 23 公里缩短到 12 公里…….	交通运输
…….县**课**（客，数据中疑为错别字）**运车**说不允许超载以后，客运局收费就开始增加**车费**了…….大一点点的孩子要收费也就算了…….	交通运输

3.2 热点问题挖掘结果分析

3.2.1 聚类 k 值选取

经过数据清洗、TF-IDF 模型生成向量，PCA 将向量降至 500 维后，从 50 到 650 间隔 50 选取 k 值运行 k-means 算法，计算得每次聚类结果的轮廓系数。轮廓系数随 k 值的变化如图 3.1 所示。根据轮廓系数的定义，轮廓系数的取值范围为（-1，1），其值越大代表聚类结果各个类别之间的距离越大，聚类的效果越好。

依据图 3.1，随着 k 值的增大，轮廓系数总体的变化趋势为先上升，在 k=650 处达到最大值，随后略下降。因此，对附件 3 的数据，选取了 k=650 进行 k-means 聚类。

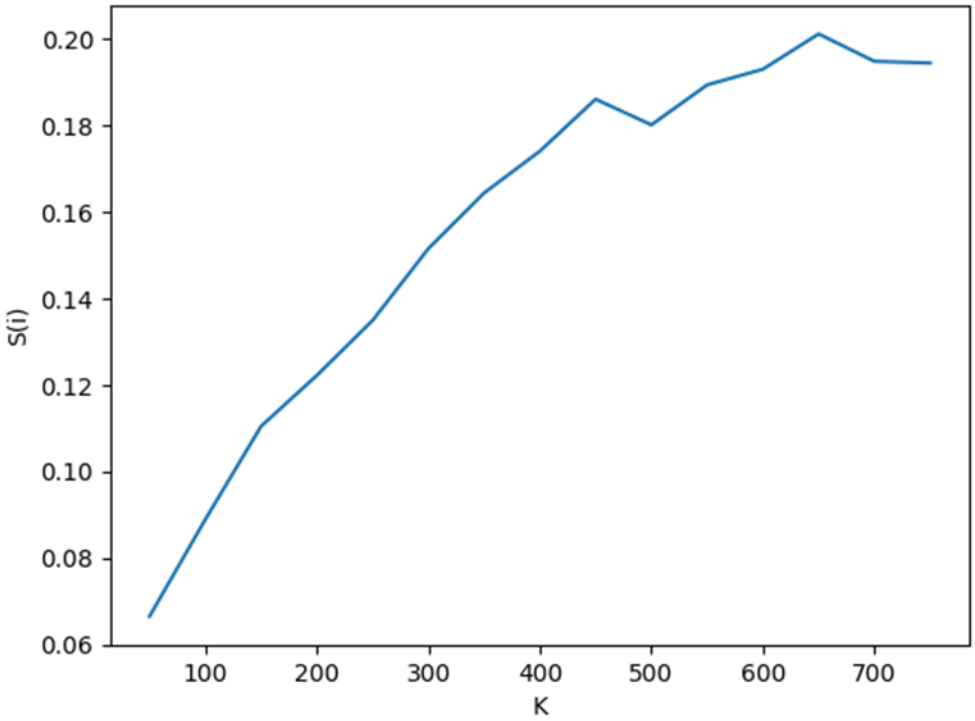

图 3.1 轮廓系数 S(I)随着 k 值的变化图

3.2.2 热度评价指标结果分析

3.2.2.1 冷却因子与区分度

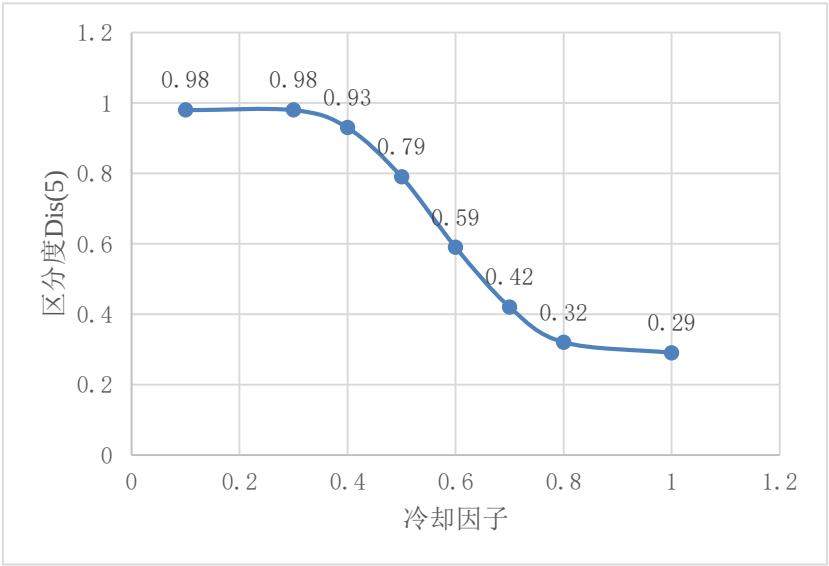

图 3.2 冷去因子与区分度关系曲线图

表 3.4 部分结果展示表

冷却因子	Dis(5)	主题聚类编号	主题热度分数（归一化之后
0.3	0.98	577	1.0
		50	0.47
		541	0.27
		635	0.21
		350	0.19
0.5	0.79	577	1.0
		50	0.53
		527	0.40
		541	0.35
		350	0.33
0.8	0.32	577	1.0
		527	0.91
		347	0.85
		460	0.81
		50	0.75

从热度分析的结果来看，基本符合，关注度与认可度越高，热度越高的特点，如表中所示，虽然冷却因子有变化，如编号 577 对应的问题，始终位于第 1 名，编号 350，541 在冷去因子为 0.8 的时候，为第 6 名，第 7 名，这几个问题对应的关注度（留言数目），认可度（点赞数目）都相对较高。

从冷却因子的效果看，冷却因子减小，冷却效果增强，有助于于提高模型的区分度(从 0.29 增加到了 0.98)，部分时间跨度较长，包含评论较多，但留言分散在较多的的时段内的问题的排名，会随着冷去因子增大而上升，例如第 347 号问题，该问题聚类后共有 18 条留言，其中有三条留言获得了 733，790，821 点赞，该问题的关注度与认可度都很高，但是因为其留言分散在 1-8 月（不含 4 月）共 7 个月份，一月份最多为 8 条，剩余几个月份都只有 2~3 条，因此最早月份（1 月）的热度会被 7 次冷却，在冷却因子较小的时候，最后的热度并不高。而当冷却因子较大，冷却效果较弱的时候，该问题的关注度高，认可度高的优势才能体现。而在几次热度评价时排名都比较高的 577 问题拥有 18 条评论，最高点赞量为 699，时间分布相对集中（分布在 3 个月份），因此其收到冷却效果的影响相对较弱。

考虑到政务问题的复杂性，一个潜在的热点的民生问题被普通民众发现、

留言的周期相对较长，所以选择冷却因子为 0.8 的时候，对应的热度排序，作为提交的热度问题表，以较大的冷却因子，减弱时间分布对热点问题的影响。此外，由于 K-Means 聚类的误差，导致编号 50 问题中混入了其他问题的留言，问题 50 的得分实际较低，因此热点问题表中，第 5 名是实际的热度值的第 6 名，即问题 350。

冷去因子为 0.8 时对应的热点问题词云图如图 3.3

(a)问题 577　　(b)问题 527　　(c)问题 347

图 3.3 热点问题前 3 名词云图

3.3 问题 3 结果分析

3.3.1 相关性分析

3.3.1.1 基于 BERT 模型的句向量的相关性

基于 BERT 模型的相关性评价下，分别选取得分较高与低的 3 条“留言-回复”，如表 3.5，

表 3.5 基于 BERT 的相关性评分

留言详情	答复意见	相关性
别墅买了以后，往往根据自己需要的功能进行装修和改造……因为有些小区业主连搭个棚子都算是违章建筑，这里随便改，随便扩大面积好像都没人管。	关于“想咨询一下……根据《中华人民共和国城乡规划法》第四十条规定……中队将严格依照法律程序进行查处。A7 县城市管理和行政执法局。2015 年 7 月 21 日	0.987
本文投诉涉及的采石场是 2017 开始入驻当地的，采石场是村民的良田，他们以村支书邝米生的关系。承包村里的一个竹子山边上是良田和土地，离人群距离过近……	《K10 县邝胡社区邝家村恶性涉黑采石场投诉》环境举报件，现已办结，现报告如下：……于 2016 年 1 月 18 日和村委会签订了山地承包合同，并有村民签字……	0.986
您好！请问我小孩下半年入小学，外地户口，所有入学证明都有，有购房合同预售，但是没有交房入住请问能	您好！您“关于在 C 市买房的业主小孩小学入学问题” ……就具体入学事宜到您住所或购房房产所在县区	0.985

报名吗？	级教育行政部门咨询。	
劳动广场中央绿化岛已经有几十年的历史，绿化岛也成为了……建议公交车始发站终点站不要设在劳动广场。	网友：您好！留言已收悉	0.721
机动车驾驶教练员职业资格证，是国家职业资格体系的组成部分……	网友：您好！您所反映的问题已知悉，感谢您的建议留言，我们将会对您的建议进行研究斟酌。谢谢您的理解与支持。	0.719
请问，带小孩去打疫苗要带什么证件呢？是所有医院的都可以打吗？	2019 年 1 月 14 日	0.690

从表中信息以及评分可以看出，当回复中含有类似“针对提出的问题,回复如下”的内容时，相关性评分会较高；当回复完全与内容无关，或没有回复内容，仅仅是日期的时候，评分较低。并且，相关性的高低，并没有收到回复意见长短的影响，如表格中倒数第二，倒数第三条数据，虽然回复内容长短不同，但因为相关性较低，所以评分较相似。从结果看，以 BERT 生成的句向量为基础的相关性评分，基本完成了区分留言相关性的任务，但是，该方法的问题是评分的区分度有待提高，相关性不同的数据之间，评分的分差稍小。

3.3.3.2 位置编码的句向量的相关性

基于位置编码(PE)的句向量相关性评分结果，如表 3.6 所示：

表 3.6 基于 PE 的相关性评分

留言详情摘要	答复意见摘要	相关性
经调查，反映我镇镇长张登武不作为的信访人为黄尊富……多次到市、县相关部门进行缠访闹访，对县、镇主要领导进行电话骚扰，在群众中造成不良影响……。	……经镇综治办调查核实，信访人黄尊富……四是信访人反映诉求必须依法依规进行，不得以发短信威胁或以打电话骚扰等非正常方式干扰、影响国家机关或国家工作人员的正常工作、生活……	0.997

首先，感谢 A 市委办信息处对网络问政的重视……业主的正常权益有人出头保障，那租户的呢？是不是就无人问津了？……小区地面公共停车位为什么要单独给部分车主享用？请明晰“公共停车位”的具体含义。	……物管办联合和馨园社区工作人员积极了解情况……业主并不是将所有权益都让渡给租户，比如说租户并没有权利参与社区的业主大会……公共停车位是和馨园集体经济组织所有。	0.989
融圣国际的几点问题反映……针对小区的问题我们小区全体业主反映如下：一、开发商存在的问题：……没有按照要求对外进行招标公告，存在利益输送嫌疑。以上问题请领导调查处理，还小区一片蓝天。	……经区城乡建设局调查了解，现将有关情况回复如下：一、开发商存在的问题……小区二十多台电梯的维护保养没有按照要求对外进行招标公告，存在利益输送嫌疑问题……	0.981
书记您好，我是 A6 区大泽湖东马社区一名农民，反映 A 市医学院污水直排农田……	感谢您对我们工作的关心、监督与支持。	0.386
我是华都小学一名学生的家长，现华都小学已经正式开学，现有师生 400 多人，大多数为一年级学生……	L1 区网宣办 2015 年 9 月 18 日	0.378
请问，带小孩去打疫苗要带什么证件呢？是所有医院的都可以打吗？	2019 年 1 月 14 日	0.292

从 word2vec 结合位置编码的相关性评分来看,由于引入了句子中的词向量的位置信息，相比于 BERT 的句向量，导致相关性评分的区分度增强,从表中数据可以看出，相关性较强的数据与较弱的数据分差达到 0.705，而在 BERT 中，这一数值为 0.297。但也是因为引入了位置信息，句子长度会对结果造成一定的影响。

3.3.2 完整性分析

基于答复意见的完整性规范定义，从附件 4 中选择了 20 条答复构成了“优秀答复”数据集，通过 LDA 模型和频率计算两种方法提取的规范词如表 3.7 所示。其中年、月、日等词代表了答复规范中答复时间部分，感谢、理解等词代表了答复规范中感谢语部分，网友、收悉等词表示了答复规范网友称呼及网友留言

总结部分。

表 3.7：基于 LDA 和频率提取的规范词及权值表

规范词	LDA 权重	规范词	频率
年	0.012	工作	1.00
月	0.010	年	1.00
市	0.009	您好	1.00
区	0.007	日	1.00
工作	0.007	月	1.00
日	0.007	支持	1.00
情况	0.006	感谢您	0.95
支持	0.005	回复	0.95
您好	0.005	收悉	0.95
回复	0.005	情况	0.90
监督	0.005	网友	0.90
感谢您	0.005	现将	0.90
网友	0.004	监督	0.85
现将	0.004	留言	0.85
收悉	0.004	理解	0.80

针对附件 4 部分答复意见的完整性评分如表 3.8 所示。其中，前三条答复意见基于频率的得分为 1，基于 LDA 的得分为 1 和 0.9，这三条答复意见完整的包含了网友的称呼、网友留言总结、情况说明与答复详情、感谢语以及答复时间五个部分，较好地满足了完整性规范。后三条评论都仅仅出现了答复详情这部分，其两种得分都是 0。综上所述，完整性的评分可以全面地反应答复意见的满足完整性规范的程度。

表 3.8：基于规范词的完整性评分

答复意见	评分 1	评分 2

网友“UU00812”您好！您的留言已收悉。现将有关情况回复如下：……建设一所社区卫生中心，目前正在拆迁腾地。感谢您对我们工作的支持、理解与监督！2019 年 1 月 15 日	1	1
网友“UU008792”您好！您的留言已收悉。现将有关情况回复如下：……在未整改到位之前，楚江新区建设工程质量安全督查站将不同意竣工验收。感谢您对我们工作的支持、理解与监督！2019 年 1 月 11 日	1	0.90
网友“UU008480”您好！您的留言已收悉。现将有关情况回复如下：……目前该地点进入油漆施工阶段，计划于元月上旬竣工，对市民造成的不便深感歉意，也感谢您对城市建设工作的关心和理解。感谢您对我们工作的关心、监督与支持。2016 年 12 月 27 日	1	1
你好。残疾人证的办理需要向你户籍所在地的县级残联申请，由定点医疗鉴定机构鉴定类别和级别。	0	0
请咨询 K 市人社部门。	0	0
国网西地省 J9 县供电公司	0	0

3.3.3 可解释性分析

基于答复意见可解释性的定义，首先对附件四中的答复意见进行因果关系抽取，部分答复意见的抽取结果如表 3.9 所示。

第一条答复意见为对解决方案的单纯阐述，并未对其中的任何概念进行解释，故因果关系抽取结果为空。

第二条答复意见的因果关系抽取抽取出“因该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大，周期较长”，“因近期持续雨天，为保证道路施工质量，需在晴好天气才能正常施工”两个具有因果关系的事件，分别对“施工周期较长”、“需要好天气才能正常施工”两个概念进行解释，进一步解答网友对“湖段怎么还没修好”的疑问。

第三条答复意见的因果关系抽取抽取出“按照《A 市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在 A 市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄 35 周岁以下（含），首次购房后，可分别申请 6 万元、3 万元的购房补贴”，通过引用规定对购房贷款问题进行解释，

体现了可解释性。

第四条答复意见的因果关系抽取将“该工棚占用人行道的情况属实，确实对市民出行造成了一定影响”重复抽取，导致抽取结果不可简单相加作为可解释性评分的依据，故考虑需要使用前文提出的可解释性评价算法。第五条回复没有概念的提出，因而没有可解释性。

表 3.9：因果关系抽取结果

答复意见	因果关系抽取结果
网友“UU00877”您好!您的留言已收悉。现将有关情况回复如下：……尽快启动万国城小区人行天桥建设并投入使用，方便出行。感谢您对我们工作的支持、理解与监督！2019 年 1 月 8 日	[]
网友“A00023583”：您好！……因该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大，周期较长。加之坪塘集镇原有管线、排水渠道较多，需先处理管线和渠道才能进行道路施工，且因近期持续雨天，为保证道路施工质量，需在晴好天气才能正常施工。……	[{'tag': '因', 'cause': '该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大', 'effect': '周期较长'}, {'tag': '因', 'cause': '近期持续雨天，为保证道路施工质量', 'effect': '需在晴好天气才能正常施工'}]
网友“A000110735”：……按照《A 市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在 A 市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄 35 周岁以下（含），首次购房后，可分别申请 6 万元、3 万元的购房补贴。“首次购房”是指在 A 市限购区域内首次购买商品住房	[{'tag': '按照', 'cause': '《A 市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在 A 市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄 35 周岁以下（含），首次购房后', 'effect': '可分别申请 6 万元、3 万元的购房补贴'}]

（含住宅类公寓）。……	
网友“UU008599”您好！……该工棚占用人行道的情况属实，确实对市民出行造成了一定影响。……感谢您对我们的工作的理解与支持！特此回复。2018 年 5 月 23 日	[{'tag': '造成', 'cause': '确实对市民出行', 'effect': '了一定影响'}, {'tag': '造成', 'cause': '该工棚占用人行道的情况属实，确实对市民出行', 'effect': '了一定影响'}]
网友：您好！留言已收悉	[]

利用算法在因果关系抽取结果上对答复意见的可解释性评分进行计算，选择评分较高的两条数据和评分较低的两条数据以及评分为 0 的两条数据展示在表 2 中。为方便读者查看，将因果抽取结果的格式化数据转化为英国关系抽取中出现的答复意见中的语句展示在表 3.10 中。根据结果分析可以看出，前两条答复意见数据中存在因果关系的文本占比较大，可解释性评分较高，3、4 条文本中存在因果关系的文本占比较少，可解释性评分较低，5、6 条文本中没有存在因果关系的文本，故可解释性评分为零。

表 3.10：基于因果关系抽取的可解释性评分

答复意见	因果关系抽取结果	可解释性评分
“UU0081841”……您好！我市于 2017 年出台《关于进一步推进人才优先发展的 30 条措施》（简称“人才新政 30 条”），对新引进的高层次人才进行人才分类，并享受安家补贴、子女就学、配……关于市级的人才补贴有关政策建议您向市人社局人力资源管理科咨询，联系电话：0000-00000000。2019 年 7 月 17 日	[{‘我市于 2017 年出台《关于进一步推进人才优先发展的 30 条措施》（简称“人才新政 30 条”），…… 关于市级的人才补贴有关政策建议您向市人社局人力资源管理科咨询，联系电话：0000-00000000。’}]	0.6931
根据《E9 县 2019 年城乡居民医疗保	[{‘根据《E9 县 2019 年城乡居民医疗	0.6484

险参保缴费工作方案》的通知（[政府发文]55 号）的文件精神，……6 月份对全县参保城乡居民进行家庭帐户注资。2019 年 4 月 24 日	保险参保缴费工作方案》的通知（[政府发文]55 号）的文件精神，……6 月份对全县参保城乡居民进行家庭帐户注资。'}]	
网友“UU008404”您好！您的留言已收悉。现将有关情况回复如下：……园方按公告要求，全程透明公开公正组织招生各环节工作。……感谢您对我们工作的支持、理解与监督！2018 年 7 月 10 日	[{‘园方按公告要求，全程透明公开公正组织招生各环节工作。’}]	0.0039
“UU0082197”您好！我局自 5 月份至今收到您多次反映关于 K4 县九龙时代广场小区垃圾处理房问题的投诉……由于多方原因，始终没有达成共识。……新的进展情况我局将及时通报小区业委会及业主代表。2019 年 7 月 8 日	[{‘由于多方原因，始终没有达成共识。’}]	0.0078
“UU008194”您的留言已收悉。关于您反映的问题，已转市交通运输局调查处理	[]	0
您的留言已收悉。关于您反映的问题，已转 I1 区委、区人民政府调查处理。	[]	0

从上述结果分析中可以看出，基于因果关系抽取出的可解释性评价，能够较好地反映在答复意见中某一新观点的提出是否同样存在对该观点的阐释，以及句与句之间是否有可解释的逻辑关系。可解释性作为答复意见的质量的一个评价指标，能够很好地评价该答复意见对提问者来说是否清晰易读、有理有据、有说服力，这是优秀的答复意见应该满足的标准。

四、结论

对政务群众留言信息进行分析研究，对了解群众关心的问题和需求，及时发现群众面临的问题，对不合理的政策法规进行调整有着重大意义，同时也是文本分析的一个重要课题和难题。传统的人工留言划分和热点整理逐渐无法满足庞大的网络留言数据信息。本文采用在 NLP 领域有良好表现的 BERT 模型对群众留言信息进行分类，采用结合命名实体识别的 K-means 算法对群众反应热点问题进行聚类，并使用 NLP 领域的相关评价参数对答复意见的相关性、完整性、可解释性进行评价，得到了量化指标。

由分类结果可以看出，通过数据挖掘实现的政务留言信息类别分类正确率可达 93%，能够一定程度上代替人工的分类结果。同时采用 K-means 等聚类方法，能在短时间内迅速掌握群众集中反应的某一问题，从而给以快速的响应，提高政务系统的效率。针对答复意见的评分，能够了解答复意见是否完整规范、与用户反映问题相关、易于用户理解，实时给以回复评分，可根据此评分对不合格的评价进行整改，从而保证用户的反应问题都能够及时解决。

本文提出的分类模型、聚类模型及答复意见的评价标准对政务留言系统的各项业务均有较大帮助，并且在其他的文本分类、聚类及 QA 评价问题上也有较好的泛化能力。

参考文献：

[1] AshishVaswani, Noam Shazeer, NikiParmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, LukaszKaiser, and IlliaPolosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[2] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton &Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[3] 吴靓婵媛. 基于社区发现的网络舆情热点主题识别研究. 硕士论文, 南京理工大学, 7 2017.

[4] 尚鸿运. 中文微博的热点话题检测及趋势预测算法研究, 天津大学.

[5] 张杨子. 面向对话系统回复质量的自动评价研究. 哈尔滨工业大学.

[6] 潘立祥. 网络热点话题发现系统研究. 硕士论文, 华中科技大学, 5 2015.

[7] 罗亚平, 王枞, 周延泉. 基于关注度的热点话题发现模型. 中文信息处理国际会议, 2007.

[8] Sainbayar Sukhbaatar ，Arthur Szlam ，JasonWeston ，Rob Fergus，End-To-End Memory Networks

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111-3119

[10] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[11] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[C]//Proceedings of NAACL-HLT. 2016: 260-270.