“智慧政务”中的文本挖掘应用

摘要

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。因此，运用自然语言处理、文本分析和数据挖掘技术对各类社情民意数据的研究，构建智慧政务系统具有重要意义。

对于问题一，先通过正则表达式过滤去除文本中一些无关的符号，得到纯中文文本信息，并利用 jieba 分词工具对中文文本进行分词，去除掉对文本信息分析无关的停用词。利用 TF-IDF 算法将词语转换成词向量，使用 Word2Vec 工具中 Skip-Gram 模型将数据训练成词向量。最后，建立一个两层的双向 LSTM 分类模型对词向量进行训练，构建出文本分类模型。

对于问题二，先采用与问题一中相类似的数据预处理方法对文本进行数据预处理。再采用 LSA 算法对向量空间进行降维处理。并使用 SVD 分解，消除同义词、多义词的影响，提高后续处理的精度。采用余弦相似度来计算文本的相似度。采用 DBSCAN 聚类算法对文本进行聚类。通过聚类，可以得出留言信息中相似问题的数量。最后，综合考虑“同类问题的点赞数”、“同类问题的数目”和“同类问题发布的时间差”对热度进行评价，得出某段时间内热点问题。

对于问题三，本文通过“相关性”、“完整性”和“可解释性”三个指标对答复的质量进行评价。相关性评价问题答复与问题题目之间的关联程度。完整性主要指问题答复的规范程度，是否涵盖标准答复中的各个要素。可解释性主要指问题答复是否有依据，是否有政策、法律法规、数据等相关因素的支撑。通过对相关性、完整性、可解释性这三个指标的综合分析，对留言问题的答复质量进行评价。

关键词：中文分词、TF-IDF 算法、LSTM、LSA 算法、SVD 分解、DBSCAN 聚类

Abstract

In recent years, with the online questioning platforms such as WeChat, Weibo, mayor's mailbox, sunshine hotline and so on, it has gradually become an important channel for the government to understand public opinion, gather people's wisdom, and gather public opinion. The work of the relevant departments, which mainly relied on manual work to divide the message and organize hotspots, brought great challenges. Therefore, the use of natural language processing, text analysis, and data mining techniques for the research of various social conditions and public opinion data to build a smart government system is of great significance.

For question one, first remove some irrelevant symbols in the text through regular expression filtering to obtain pure Chinese text information, and use the jieba word segmentation tool to segment the Chinese text to remove stop words that are irrelevant to the analysis of text information. Use the TF-IDF algorithm to convert words into word vectors, and use the Skip-Gram model in the Word2Vec tool to train the data into word vectors. Finally, a two-layer two-way LSTM classification model is established to train word vectors, and a text classification model is constructed.

For problem two, first use a data preprocessing method similar to that in problem one to preprocess the text. Then the LSA algorithm is used to reduce the dimension of the vector space. And use SVD decomposition to eliminate the influence of synonyms and polysemy, and improve the accuracy of subsequent processing. The cosine similarity is used to calculate the similarity of the text. Use DBSCAN clustering algorithm to cluster text. Through clustering, the number of similar questions in the message can be obtained. Finally, comprehensively consider the "like points of similar problems", "the number of similar problems" and "the time difference of similar problems" to evaluate the heat, and get the hot issues within a certain period of time.

For question three, this article evaluates the quality of the response through three indicators: "relevance", "completeness" and "interpretability". The degree of relevance between the relevance evaluation question answer and the question title.

Completeness mainly refers to the degree of standardization of the question answer, whether it covers all elements in the standard answer. Interpretability mainly refers to whether there is a basis for answering questions and whether they are supported by relevant factors such as policies, laws, regulations, and data. Through the comprehensive analysis of the three indicators of relevance, completeness, and interpretability, the quality of the answers to the message questions is evaluated.

Keywords: Chinese word segmentation, TF-IDF algorithm, LSTM, LSA algorithm, SVD decomposition, DBSCAN clustering.

目录

1. 问题背景

当今社会正处于信息大爆炸的时代，大数据技术的作用也越来越重要。政府是信息的管理者，甚至一定程度上是数据的最先获得者。在这个大数据时代，政府部门应用大数据思维，通过大数据技术来实现数据信息的共享，将数据的功效最大程度地发挥出来是当今社会的主要趋势。

随着社会的进步与发展，在微信、微博、市长信箱、阳光热线等网络问政平台上进行了解民意、汇聚民智、凝聚民气成为了主流，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智能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，随着大数据时代的发展，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，这将对大大地提升政府的管理水平和施政效率。

因此本文要运用收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，利用自然语言处理和文本挖掘，建立有效的数学模型进行评价。

2. 挖掘目标

本次建模的目标是通过网络问政平台提供的群众问政留言记录及相关部门对部分群众留言的答复意见，利用 Jieba 中文分词工具对留言进行分词、双向 LSTM 分类模型以及 DBSCAN 聚类的方法，达到以下三个目标：

（1）利用文本分词和文本分类的半监督机器学习方法对群众留言进行文本挖掘，经过文本数据不断地训练模型，建立关于留言内容的一级标签分类模型。结合 F-score 对分类模型进行评价。

（2）利用 DBSCAN 聚类的方法将群众问题的留言进行归类，并定义合理的热度指标、计算热度值，挖掘群众反映的热点问题。分析排名前五的热点问题，了解群众反映的最大问题，有助于相关部门能有效处理，提高服务效率，群众的问题也能及时解决。

（3）建立一套完整的评价方案，将群众答复意见的质量从相关性、完整性、可解释性进行评价，通过方案筛选有用的答复意见，可以减轻政府部门的工作压力，及时认识到群众的问题是否解决。

3. 分析方法与过程

总体流程图

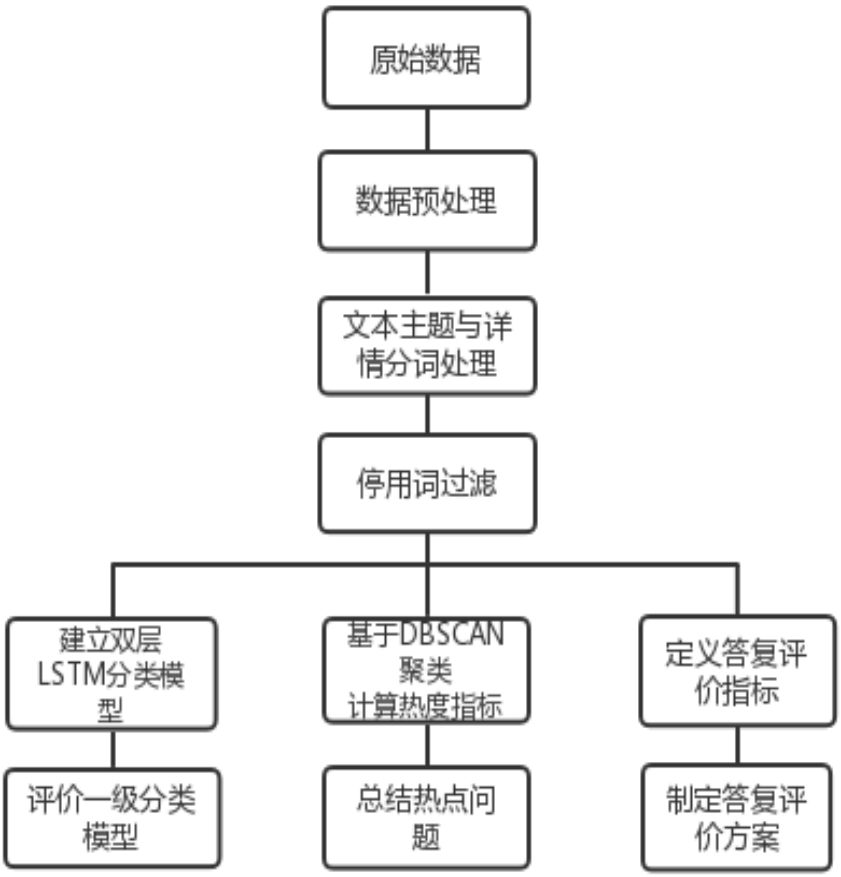

图 3-1 总体流程图

本论文的分析流程可大致分为以下三步：

第一步：对提供的原始数据进行预处理，包括数据清洗、中文分词、停用词过滤等操作。

第二步：文本留言经过处理后，建立相关的分类模型或运用多种方法对于留言数据进行分类和整理。

第三步：从对应的结果中选择合理的分类模型、总结排名前五的热点问题和评价留言答复。

3.1. 问题 1 分析方法与过程

3.1.1 数据预处理

3.1.1.1 数据清洗

为了提高文本分类的准确率，需要对文本中一些不重要的字符进行清洗过

滤。在题目给出的数据中，出现了很多不影响分类的无用字符，干扰问题的分析，本文使用设计正则表达式过滤掉中英文标点符号、空格和数字等无效字符。本文所给的数据如图 3-2 所示：

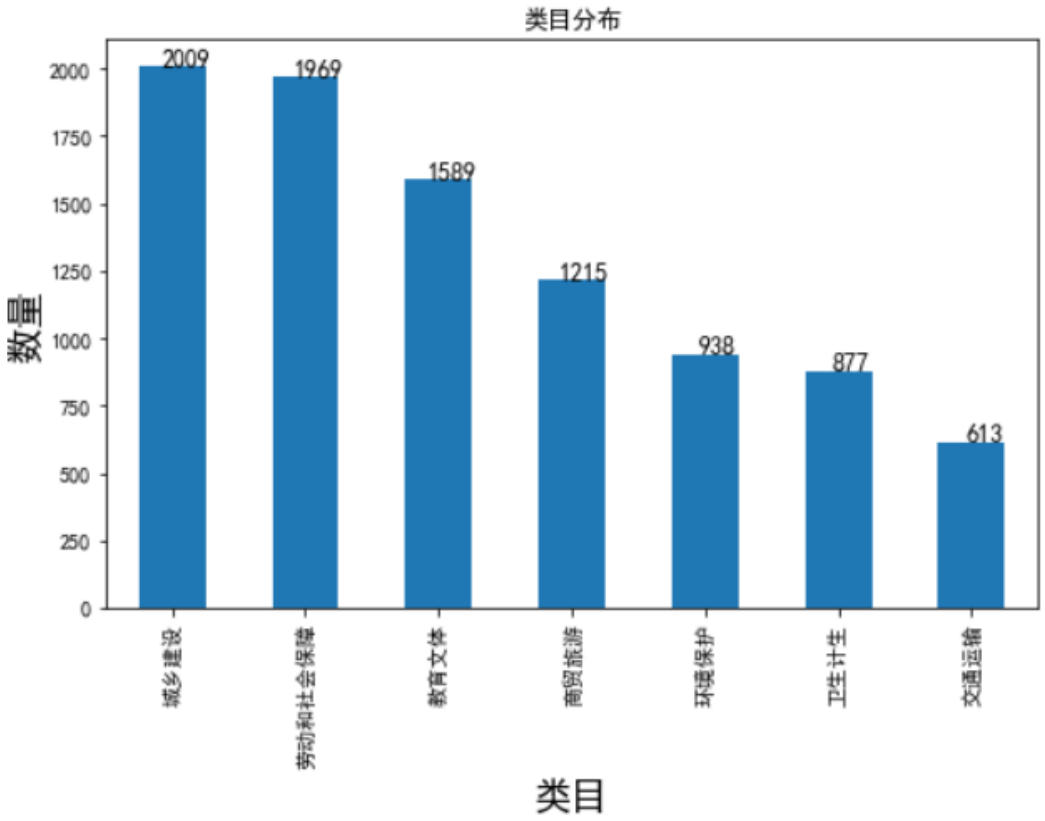

图 3-2 类目分布图

3.1.1.2 文本分词

在对留言信息进行挖掘分析之前，需要先将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。在数据清洗完成后，留言主题与留言详细都是以中文文本的形式给出数据，为了便于转换，要先对这些描述性文字进行文本分词。文本分词就是将中文字符组成的句子分割成由中文词组组成的句子，在本文中我们采用 python 中的 Jieba 分词包进行分词，并去掉一些影响结果的停用词。Jieba 分词分词速度快，分词效果好，并且可以词性标注。Jieba 分词的思想主要是，首先通过对照字典生成句子的有向无环图，在根据选择的模式不同，依照词典寻找最短路径，然后对句子进行截取或直接截取。对于不在词典中的词，将使用 HMM 进行新词的发现，这些能使得有更好的分词结果。

我们对文本进行分词处理后，分词结果如表 3-1：

序号	留言详情	留言主题	一级分类
0	大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑 集团……	西湖 建筑 集团 占道 施工 安全隐患	城乡建设
1	在水一方 人为 烂尾 多年 安全隐患	书院 路 主干道 在水一方 大厦 一楼……	城乡建设
2	市政府 交警支队 安监局 环保局 区政府 市区 杜鹃 文苑 小区 业主 涉及网上 写信……	杜鹃 文苑 小区 外 非法 汽车 检测站 开业	城乡建设

表 3-1 **分词结果表**

3.1.1.3 TF-IDF 算法

在对留言主题与留言详情分词后，需要对这些词语转换为计算机能识别的向量，进一步以便于挖掘分析的使用。本文使用 python 中的 sklearn 库自动转换为词向量。从文本中提取的特征对文本分类的重要程度不一样，对于重要的特征需要较大的权重，而不太重要的特征则需要较小的权重。本文采用 Jieba 自带的语义库——TF-IDF 算法转换为权重向量。该算法的主要思想是一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能代表该文章，即适合分类。

TF-IDF 算法的具体原理如下[1]：

第一步，计算词频，即某一给定的词语在该本文中出现的频率，即 TF 权重。

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

或

$$TF_w = \frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中出现次数最多的次的出现次数}}$$

第二步，计算逆文件频率，表示包含某个词语对文章分类的重要性度量指标，即 IDF。

$$IDF = \log\left(\frac{\text{词料库的文档总数}}{\text{包含词条}w\text{的文档数}+1}\right)$$

第三步，计算 TF-IDF 值。

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

3.1.1.4 训练词向量

训练词向量就是将词语用向量表示。将词义相近的词转化为向量后，他们之间的欧氏距离也会很小。本文使用 Word2Vec 工具训练词向量，实现词语的向量化表示。在 Word2Vec 工具包中，我们采用 Skip-Gram 模型将数据训练成词向量。

将训练好的文本数据份成训练集和测试集。将训练及中的词向量再用于深度学习模型的训练，保存训练好的词向量，以便于下一步的文本挖掘分析。训练后的几个具体的词向量如表 3-2 所示。

部分词	向量
环境保护	[0.11437468，-0.02319438，-0.17052916，-0.37452495,……]
业主	[-0.09696922，-0.11551602，0.00216453，-0.31718507，……]
小区	[-0.21214126，-0.01224018，-0.2939689，-0.4294774，……]

表 3-2 词向量表

从图中可以看出，环境保护、业主、小区等被表示成 300 维的向量，本文共生成向量 2375 个。

3.1.2 文本分类

3.1.2.1 基于 LSTM 的分类模型

本文建立一个两层的双向 LSTM 分类模型对文本进行分类，LSTM 更有利于获取有上下文关系的特征向量。有些时候预测可能需要由前面若干输入和后面若干输入共同决定，双向的 LSTM 分类模型会更加准确[2]。

建立的 LSTM 模型的结构如图 3-3，包括输入层、两层双向 LSTM 层、全连接神经网络层、输出层。

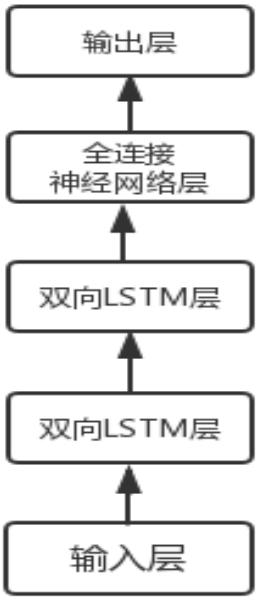

图 3-3 LSTM 模型结构图

第一层为输入层，输入层主要是将训练集词向量传入到 LSTM 模型当中去。

第二层为双向 LSTM 层，主要用来获取句子向量的上下文信息。为了防止过拟合，本文使用了深度学习中的 dropout 进行正则化，dropout 的原理是临时删除一部分神经元，梯度下降更新其他神经元的权值，然后下次临时删除其他神经元同样进行这样的操作。

第三层为双向 LSTM 层，由于一层双向 LSTM 层的效果不是很好，所以增加一层双向 LSTM 层使得结果更准确。

第四层为全连接神经网络层，由于 LSTM 输出的向量的维度是指定的，为了能计算损失，全连接层用于将输出的向量转换为标签向量的维度，达到标签的向量维度与输出向量维度一致。

第五层为输出层，输出分类向量。

3.1.2.2 具体训练过程

由于 LSTM 分类结果为向量，为了实现标签与分类向量的比较，需要对标签向量化。但观察到留言文本内容的长度与标签的不同，所以要对标签采用补零法，达到两者的维度相同。本文采用半监督学习的基本思想将为未标签样本进行标签，将训练集和测试集都放入模型，利用少量标注样本和大量未标注样本进行机器学习。

本文基于 LSTM 分类模型训练的具体流程如图 3-4。

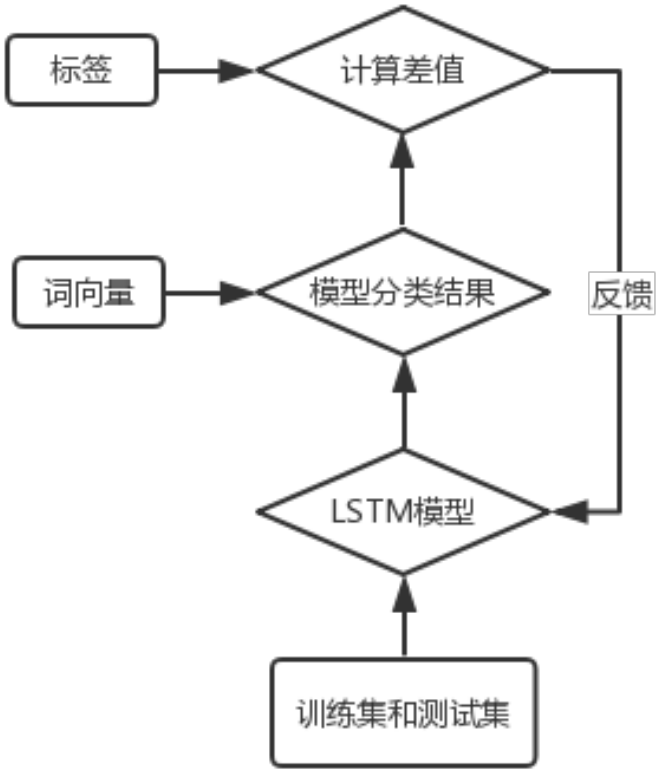

图 3-4 LSTM 分类模型训练流程图

3.1.3 模型的评估

本文首先会采用普通的 F-score 对分类方法进行评估：

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

同时本文也利用基于混淆矩阵的 Kappa 系数对分类模型进行评估。

混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用 n 行 n 列的矩阵形式来表示。具体评价指标有总体精度、制图精度、用户精度等，这些精度指标从不同的侧面反映了图像分类的精度[3]。

kappa 系数是一种衡量分类精度的指标。它是通过把所有地表真实分类中的像元总数（N）乘以混淆矩阵对角线（Xkk）的和，再减去某一类地表真实像元总数与该类中被分类像元总数之积对所有类别求和的结果，再除以总像元数的平方减去某一类地表真实像元总数与该类中被分类像元总数之积对所有类别求和的结果所得到的。[4]kappa 计算结果为-1~1，但通常 kappa 是落在 0~1 间，可分为

五组来表示不同级别的一致性：0.0~0.20 极低的一致性、0.21~0.40 一般的一致性、0.41~0.60 中等的一致性、0.61~0.80 高度的一致性和 0.81~1 几乎完全一致。

3.2. 问题 2 分析方法与过程

3.2.1 文本预处理

先对留言文本进行与上题一样的完成文本预处理，除此之外，在聚类之前还需要对文本进行一些特殊的处理。在大数据的环境下，考虑到数据一般会比较稀疏，利用稀疏向量从训练集中抽取特征。

在文本挖掘中，由于向量空间维度太高，本文要用到 LSA 算法进行降维的处理。通过对大量的文本集进行统计分析，从中提取出词语的上下使用含义。技术上使用 SVD 分解等处理，消除同义词、多义词的影响，提高后续处理的精度。

根据降维后的特征，计算每一条留言中每一个词的权重，采用与上题相同的 TF-IDF 方式。降维后的矩阵还要进行规范化。完成上述的处理之后，可用于后面计算留言之间的相异度和进行文本聚类[5]。

3.2.2 文本相似度

留言 X 与留言 Y 之间的相似性 sim(X,Y)，采用余弦相似度来度量。余弦相似度用向量空间中两个向量对的夹角的余弦值来衡量两个文本之间的相似度，余弦相似度更注重两个向量在方向上的差异，适合文本之间相似度的计算。计算公式如下：

$$sim(X,Y)=\frac{\sum_{i=1}^{n}(X_i\times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\times\sqrt{\sum_{i=1}^{n}(Y_i)^2}}$$

sim(X,Y)的取值范围为[0，1]，相似性越接近于 1，则说明留言 X 与留言 Y 的内容越相近；反之，则说明可能性越小。

3.2.3 DBSCAN 聚类

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为

密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。[6]

尝试了几种文本聚类的算法，考虑到我们不知道要划分的聚类个数，本文采用 DBSCAN 聚类算法进行聚类。

考虑到数据集是稠密的，并且数据集不是凸的，DBSCAN 会比 K-Means 聚类效果好很多。DBSCAN 聚类算法是基于密度的聚类算法，目前数据较少，采用 DBSCAN 算法。DBSCAN 算法把密度相连的点的最大集合定义为类簇，其思想是找出核心点对应的密度可达的点构成类簇。DBSCAN 算法聚类速度迅速，并且可以处理噪音数据。

根据聚类结果，一步步调试 epsilon 和 minPts 参数后，确定 epsilon 与 minPts 分别为 0.62、2。DBSCAN 算法的一般步骤为[7]：

步骤一，任意选择一个点（既没有指定到一个类也没有特定为外围点），计算它的 NBHD(p,epsilon)判断是否为核点。如果是，在该点周围建立一个类，否则，设定为外围点。

步骤二，遍历其他点，直到建立一个类。把 directly-reachable 的点加入到类中，接着把 density-reachable 的点也加进来。如果标记为外围的点被加进来，修改状态为边缘点。

步骤三，重复步骤 1 和 2，直到所有的点满足在类中（核点或边缘点）或者为外围。

3.2.4 热度评价指标

3.2.4.1 定义热度评价指标

针对群众集中反映的热点问题来说，首先一个问题反映的人数越多，则说明该问题在某一时间段内是受群众关注的，因此，将同一问题的留言条数作为热点问题热度的一个影响指标；其次，群众的点赞数同样也影响问题反映的程度，在某段时间内，某留言问题的点赞数越多，说明该留言的问题受关注较高；最后，留言问题提出的时间范围差也是影响热点问题的指标之一，假使留言时间差越大，而点赞数却越少则说明该留言的问题受关注较低。

通过上述分析可知，从三个角度对热点留言问题的影响指标进行分析：

（1）同类问题的点赞数

（2）同类问题的数目

（3）同类问题发布的时间差”

3.2.4.2 热度评价方法

文本综合考虑三个角度对热点问题的影响，以天为时间单位，对留言热度进行计算，公式如式　所示[8]。

$$Hot_T(t)=\sum_i\left(\alpha\frac{S_n}{S_N}+\beta\log p_n+\gamma w_t\right)$$

其中，S_N为某段时间内所有留言条数，S_n为问题 T 相关的留言条数，p_n为问题 T 的点赞数，α、β、γ为调整因子，本文选取$\alpha=0.8$，$\beta=0.1$，$\gamma=0.1$。

3.3. 问题 3 分析方法与过程

3.3.1 评价指标定义

（1）相关性：在进行问题答复质量相关性的评价过程中，答复的相关性主要指问题答复与问题题目的关联程度。

（2）完整性：在进行问题答复质量完整性的评价过程中，答复的完整性主要指问题答复的规范程度。

（3）可解释性：在进行问题答复质量相关的评价过程中，答复的可解释性主要指问题答复是否有依据，是否有政策、法律法规、数据等相关因素的支撑。

3.3.2 相关性评价分析方案

在评价问题答复与问题的关联程度时，我们可以基于问题一中的方法，对附件 4 中的答复进行主题信息提取并分类。通过判断问题所属的类别与答复所属类别是否一致，进而来判断问题的答复与问题是否相关，从而对答复的相关性进行评价。

3.3.3 相关性评价过程分析

（1）假设问题信息和答复信息可以划分为若干个主题类，所有主题类别可

定义为 C ，则有：

$$C = \{C_1, C_2, C_3, \ldots\ldots, C_{n-2}, C_{n-1}, C_n\}$$

其中，$C_1, C_2, C_3, \ldots\ldots, C_{n-2}, C_{n-1}, C_n$ 为各主题类别的标签值。

（2）根据问题一中的模型，定义问题为 Q ，提取出问题的所属主题类：

$$C_Q = F(Q)$$

其中，$F(x)$ 为问题分类函数模型。

（3）基于问题一中的模型，对附件 4 中问题的答复信息进行训练，定义答复为 A ，提取出答复的所属主题类：

$$C_A = f(A)$$

其中，$f(x)$ 为答案分类函数模型。

（4）假设根据问题 Q 的主题类别标签提取结果 $C_Q = C_i$ ，假设根据答复 A 的主题类别标签提取结果 $C_A = C_j$ ，若 $i = j$ ，则称答复 A 与 问题 Q 是完全相关的，若 $i \neq j$ ，则称答复 A 与 问题 Q 是不完全相关的。

（5）定义相关程度为 ϑ ，当完全相关时，$\vartheta = 1$ ；当不完全相关时，$\vartheta = g(C_A, C_Q)$，（其中，$g(x, y)$用于计算两个变量的相关程度。

3.3.4 完整性评价分析方案

答复的完整性主要指问题答复的规范程度，一个完整规范的答复需要包含全部的要求信息。通过判断问题是否包含全部的要求信息，进而判断答复的完整程度，从而对答复的完整性进行评价。

3.3.5 完整性评价过程分析

（1）假设一个完整的答复需要包含答复时间、问题分析描述、造成问题的原因、问题的处理结果、采取措施的解释信息这五大要素。

（2）定义：完整的答复信息为 A ，答复时间为 T ，问题分析描述为 P ，造成该问题的原因为 R ，问题的处理结果为 S ，对应的解释信息为 I ，问题完整程度为 η 。

（3）因此，完整的答复信息 A 可以表示为：

$$A = \{T, P, R, S, I\}$$

（4）那么对于 $\forall a \subseteq A$，都可称为对问题的答复。

（5）当$a = A$ 的时候，a 即为一个规范的完整性答复；当$a \subset A$ 的时候，a 即为一个非规范非完整的答复。

（6）因此，问题的完整程度 η 即可表示为：

$$\eta = \frac{|a|}{|A|}$$

（7）根据假设推广，若答复信息由 m 个要素构成，则有

$$A = \{E_1, E_2, E_3, \dots\dots, E_{m-2}, E_{m-1}, E_m\}$$

$$|A| = m$$

因此，当答复信息由 m 个要素构成时，则问题的完整程度 η 可以表示为：

$$\eta = \lim_{i \to m} \frac{|a|}{|A|}$$

3.3.6 可解释性评价分析方案

在评价问题答复的可解释性时，先对模型所面向问题的政策、法律法规等数据进行预处理，通过分词提取其中的关键词，先构建一个“政策法规语料库”。然后对问题答复中的关键词进行提取，对关键词进行判断。

3.3.7 可解释性评价过程分析

（1）定义根据模型所面向问题的政策、法律法规等数据构建的语料库为 D，根据指定答复构建的关键词库为 K。

$$D = \{d_1, d_2, d_3, \dots\dots, d_n\}$$

$$K = \{k_1, k_2, \dots\dots, k_m\}$$

（2）定义问题答复中关键词的总数为 N，即 $N = m$，问题中关键词在预料库中的数量为 N_i，则有：

$$N_i = |\, D \cap K \,|$$

（3）定义可解释程度定义为 ξ，则有：

$$\xi = \frac{N_i}{N}$$

3.3.8 综合分析

根据上述分析，我们定义了 ϑ，η，ξ，分别对相关性、完整性、可解释性

进行了分析，现对这三个指标进行综合分析。

定义答复质量综合指数为 e ，则有：

$$e = x \cdot w$$

其中 $x = [\vartheta \quad \eta \quad \xi]$，$w$为权重系数，$w = \begin{matrix} w1 \\ [w2] \\ w3 \end{matrix}$

4. 结果分析

4.1. 问题 1 结果分析

对于分类方法进行评价，本文采用 F-score。将训练的每论结果进行可视化展示，见图 4-1。

图 4-1 可视化结果

从图可以看出，模型训练集的精确度大致达到了 90%以上，损失函数均小于 10%，模型展示出较好的分类准确度。

为了能直观地比较出两层的双向 LSTM 分类模型的优劣性。本文将深度学习两层的双向 LSTM 分类模型与 Xgboost[9]分类模型进行比较。Xgboost 模型和 LSTM 分类模型的精确度结果如图 4-2。

```
logloss: 0.298
              precision    recall  f1-score   support

        城乡建设       0.85      0.84      0.85        62
        环境保护       0.97      0.91      0.94       210
        交通运输       0.88      0.93      0.90        83
        教育文体       0.89      0.90      0.90       120
      劳动和社会保障       0.88      0.89      0.88       197
        商贸旅游       0.91      0.94      0.93       154
        卫生计生       0.93      0.92      0.92        95

    accuracy                           0.91       921
   macro avg       0.90      0.90      0.90       921
weighted avg       0.91      0.91      0.91       921


Epoch 57/60
8289/8289 [==============================] - 4s 429us/step - loss: 0.0666 - f1: 0.9211 - val_loss: 0.3679 - val_f1: 0.7095
Epoch 58/60
8289/8289 [==============================] - 3s 419us/step - loss: 0.0645 - f1: 0.9197 - val_loss: 0.3731 - val_f1: 0.7075
Epoch 59/60
8289/8289 [==============================] - 4s 428us/step - loss: 0.0666 - f1: 0.9206 - val_loss: 0.3731 - val_f1: 0.7091
Epoch 60/60
8289/8289 [==============================] - 4s 511us/step - loss: 0.0636 - f1: 0.9206 - val_loss: 0.3793 - val_f1: 0.7065
```

图 4-2 Xgboost 模型的精确度结果

由于 LSTM 模型训练次数过多，文本展示最后几轮结果，计算每轮结果的平均值，得到损失函数的均值为=0.0636，F1-score 的均值为 0.9206。而 Xgboost 模型的 logloss 值与 f1-score 均值分别为 0.298、0.91。以上对比发现，LSTM 模型对此题的文本分类有较高的准确度[10]。

本题的混淆矩阵如图 4-3 ：

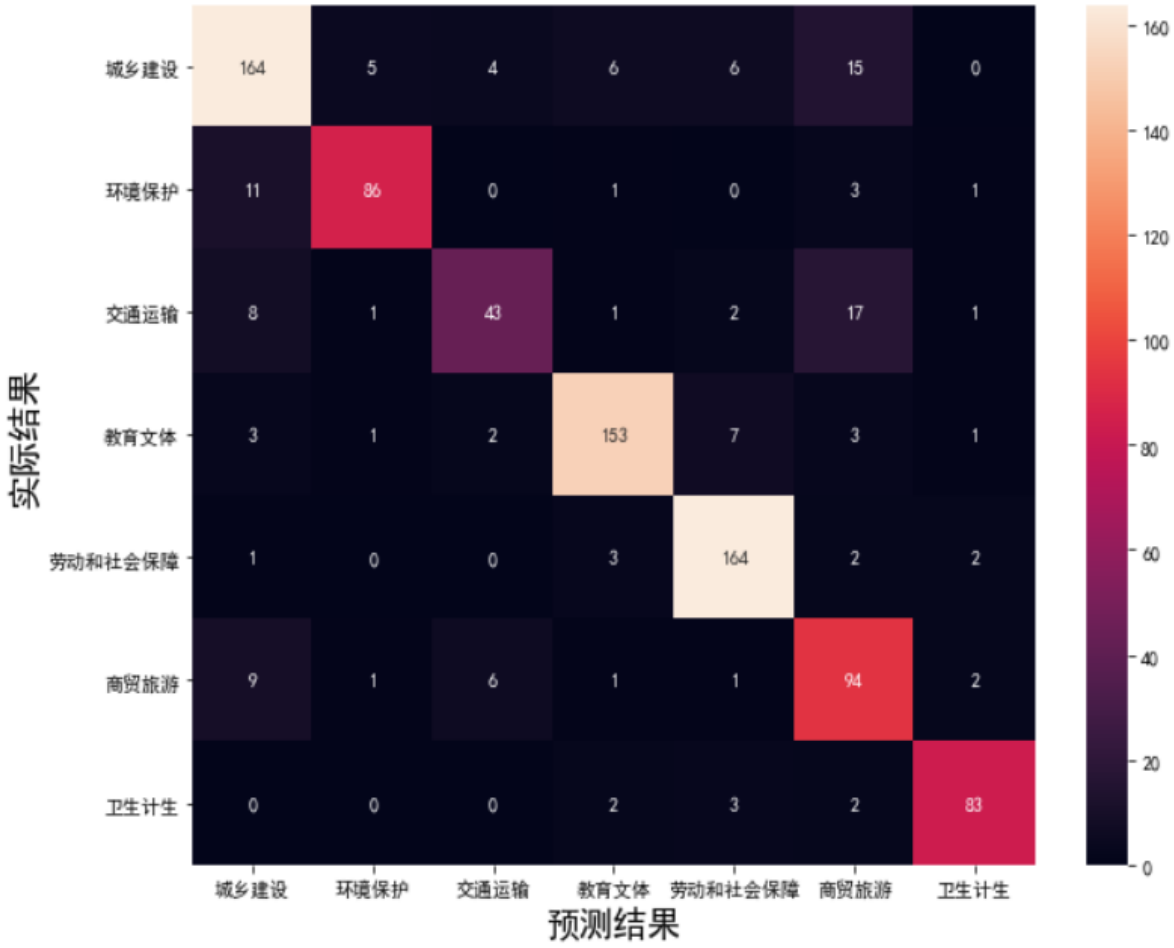

图 4-3 混淆矩阵

本文还采用 Kappa 系数用于衡量分类精度。基于混淆矩阵图的计算得到模型评价参数为 0.8267，说明预测结果与实际结果几乎完全一致。这表明了此题所用两层的双向 LSTM 分类模型预测结果准确度高，结果合理。

4.2. 问题 2 结果分析

将 DBSCAN 聚类的结果可视化，DBSCAN 聚类图如图 3-3 所示。从图 4-4 可以看出，经过 DBSCAN 聚类后，文本大致分为 624 个聚类数，并将同为一类的留言打上同一标签。

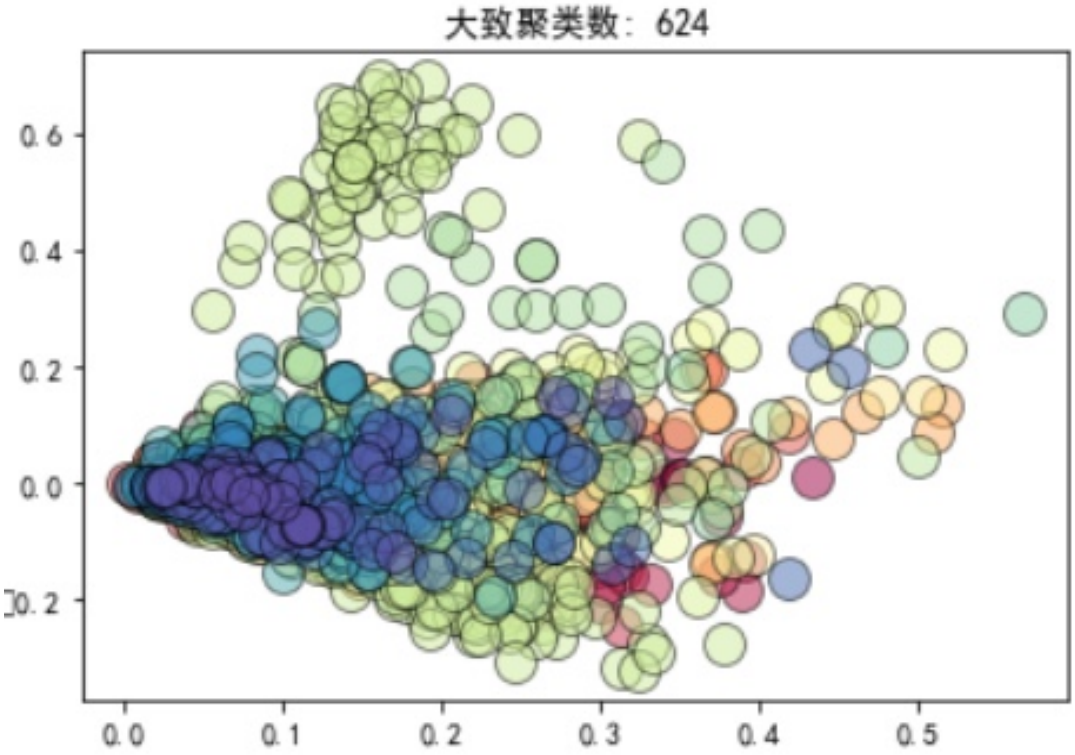

图 4-4 聚类结果图

根据族的聚类结果，输出不同族聚类结果，下面给出族群序号为 0 的聚类实际效果。

族群序号	留言 ID	留言用户	留言主题	点赞数	反对数
0	0	A0182491	A 市经济学院体育学院变相强制实习	9	0
0	12	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	0	0
0	24	A220235	A 市经济学院强制学生实习	0	0
0	26	A3352352	A 市经济学院强制学生外出实习	3	0
0	27	A1204455	A 市经济学院组织学生外出打工合理吗？	1	0

表 4-1 不同族聚类结果

从表 4-1 可以看出，通过聚类后的文本分类将强制实习这一问题能较准确地归为一类，并打上了同一标签，族群序号为 0。

根据本文定义的热度指标对留言进行热度计算，计算的热度指标如表 3-2：

热度排名	热度指标	时间范围	地点/人群	问题描述
1	0.791413	2019/01/11 至 2019/10/27	A 市西地省	调查西地省多次发生 P2P 平台诈骗案破案进程
2	0.76875	2019/4/7 至 2019/04/11	A 市梅溪湖周围小区	附近小区教育资源分配不均小学入学问题没有解决
3	0.660202	2019/4/18 至 2019/09/06	A 市渝长厦高铁	高铁距离绿地海外滩小区太近严重扰民
4	0.539207	2019/01/07 至 2019/12/20	A 市普惠性幼儿园	普惠性幼儿园收费过高且分配名额有限孩子无法入学
5	0.525326	2019/02/21 至 2019/04/12	A6 区月亮岛路	月亮岛路架设高压电线环评造假有害严重影响附近居民生活

表 4-2 **热度计算结果**

从表 4-2 可以看出，A 市西地省的 P2P 平台诈骗案破案进程热度最高，为 0.791413，热度排名第一，A 市梅溪湖周围校区教育资源分配不均小学入学没有解决、A 市渝长厦高铁距离绿地海外滩小区太近严重扰民、A 市普惠性幼儿园收费过高且分配名额有限孩子无法入学、A6 区月亮岛路架设高压电线环评造假有害严重影响附近居民生活分别排名 2、3、4、5。

4.3. 问题三结果分析

根据本文提出的评价方案，对问题答复相关性、可解释性、完整性三个指标进行分析。通过分析问题与答复主题的相似度，分析答复中引用相关法律法规的比例，分析答复中是否包含标准答复的各个要素进行综合评价。问题三的结果如表 4-3 所示。

留言编号	留言用户	留言主题	答复意见	是否
2554	A00023583	A3 区潇楚南路洋湖段怎么还没修好	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3 区洋...	0
4233	UU0081297	咨询A市公积金新政的相关问题	网友“UU0081297”您好！您的留言已收悉。现将有关情况回复如下：根据《关于调整住房公积金贷款政策有关...	1
4892	UU0081464	咨询外省人落户A市的相关问题	网友“UU0081464”您好！您的留言已收悉。现将有关情况回复如下：根据《A市公安局关于进一步规范市外迁入...	1
5559	UU0081326	咨询A市医保异地报销政策	网友“UU0081326”您好！您的留言已收悉。现将有关情况回复如下：若是由于...	0
7518	UU0082142	A7 县星沙人民呼吁加强交通建设	网友“UU0082142” 您好！您的留言已收悉。现将有关情况回复如下： 你好，目前我局正在编制《A市轨道交通..	0

表 4-3 答复意见表

从表4-3中可以看出，一个留言编号为“4233”、“4892”的留言的答复在完整性、可解释性、相关性三个方面都表现的较好，不仅答复规范，答复与问题关联度大，而且还引用了相关的法律法规，总体较好，而“2554”、“5559”、“7518”表现的不好，至少缺少了三个要素中的一项。

5. 结论

对网络问政平台的留言信息进行分析研究，建立基于自然语言处理技术的智

慧政务系统，对政府提高管理水平和施政效率具有重大意义。传统的文本及解读已经不能满足数据量庞大的社情民意文本数据。本文采用根据双向 LSTM 分类模型和 DBSCAN 聚类方法，建立合理的关于留言内容的一级标签分类模型和将某一时段内反映特定地点和特定人群问题的留言分类，并定义合理的热度评价指标，整理出热点问题。针对群众对留言的答复意见，定制了一套从相关性、完整性和可解释性出发的评价方案。

由分析结果可以看出，本文基于双向的 LSTM 的分类模型分类结果较准确，留言内容根据一级标签分类的准确率高达 90%以上。将本文进行 DBSCAN 聚类得到相似留言内容的归类，定义合理的热度指标发现 A 市西地省的 P2P 平台诈骗案破案进程、，A 市梅溪湖周围校区教育资源分配不均小学入学没有解决、A 市渝长厦高铁距离绿地海外滩小区太近严重扰民、A 市普惠性幼儿园收费过高且分配名额有限孩子无法入学、A6 区月亮岛路架设高压电线环评造假有害严重影响附近居民生活，政府应该着力尽快解决这些问题。

6.参考文献:

[1] 张闯. 基于深度学习的知乎标题的多标签文本分类[D].北京交通大学,2018.

[2] 邬明强,邬佳明,辛伟彬.Word2Vec+LSTM 多类别情感分类算法优化[J].计算机系统应用,2020,29(01):130-136.

[3] 张安定. 遥感原理与应用题解：科学出版社，2016

[4] 唐万，胡俊，张晖,等. Kappa 系数:一种衡量评估者间一致性的常用方法(英文)[J].上海精神医学(Shanghai Archives of Psychiatry)，2015(1):62-67.

[5] 凤丽洲. 文本分类关键技术及应用研究[D].吉林大学,2015.侯泽民， 巨筱. 一种改进的基于潜在语义索引的文本聚类算法[J]. 计算机与现代化，2014，7：24-27.

[6] https://baike.baidu.com/item/DBSCAN/4864716?fr=aladdin#ref_[1]_3063170 A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise ．百度学术[引用日期 2018-03-29]

[7] 陈毅森,李清亮,曹德强.基于常驻小区的投诉热点聚类方法研究及应用[J].广东通信技术,2019,39(12):71-76+80.

[8] 吴刚勇,张千斌,吴恒超,顾冰.基于自然语言处理技术的电力客户投诉工单文本挖掘分析[J].电力大数据,2018,21(10):68-73.

[9] 段立,徐鸿宇,王懿,赵莉,刘冲,郭娇.基于 word2vec 和 XGBoost 相结合的国网 95598 客服投诉工单分类[J].电力大数据,2019,22(12):50-57.厉建宾，朱雅魁，付立衡 ．基于大数据技术的客户诉求分析

[10] 谢娟英,王春霞,蒋帅,张琰.基于改进的 F-score 与支持向量机的特征选择方法[J].计算机应用,2010,30(04):993-996