基于机器学习和 NLP 的“智慧政务”体系优化研究

摘要

网络问政平台正成为人民问政咨询、政府了解民意的主要渠道，大量的文本数据量，使得人工划分工作空前冗杂。为解决这一难题，本文建立基于 NLP 问题的 SVM 机器学习模型、基于卷积神经网络的 TextCNN 模型和 LDA 主题模型，实现了对留言的标签分类、热点问题的提取以及政府答复意见的质量评估。

针对问题一：为较为准确地对留言进行一级标签分类，本文建立了**基于 NLP 问题的 SVM 机器学习模型**。首先解析已知留言的一级标签，然后对留言详情进行分词等文本处理，之后计算文本的特征值，将样本向量化表示，再将其归一化，得到训练集和数据集，接着用 SVM 模型进行机器学习，最后采用 F-Score 指标对分类结果进行评价，评价分数为 0.54，说明 SVM 分类模型较为准确。

针对问题二：对热点问题进行挖掘，定义合理的热度评价指标，然后对热点问题进行排序和总结。本文首先用 K-means 聚类算法对留言进行初步聚类，根据初步聚类结果，本文引入**时间衰减机制**，将五个热度评价指标定义为**平均时间衰减值、留言个数、平均留言文本长度、反对数、点赞数，**然后建立**基于卷积神经网络的 TextCNN 模型**对留言进一步归类，接着通过**层次分析法**计算五个热度评价指标的权重，将五个热度指标加权求和得到热度值，热度值越高，表明该问题越应该受到重视，最后根据热度值的大小得出热点问题。

针对问题三：为对各部门答复意见的质量进行较为准确的评价，本文建立 **LDA 主题模型**进行定量分析。首先对答复意见进行文本预处理，并构建新的语料库和词典，然后根据模型对文本进行模型训练，得到答复意见的相关性、完整性和及时性三个指标的得分，接着对三个指标**加权求和，**最后生成各答复意见的评分情况。评分越高，说明该答复意见越好。下面给出所有答复意见的平均打分情况：

相关性	完整性	及时性	答复评分
0.62	0.63	0.56	63.31

关键词：自然语言处理、SVM、K-Means 聚类算法、TextCNN、LDA 主题模型

Abstract

The network political platform is becoming the main channel for the people to consult and the government to understand the public opinion. In order to solve this problem, the SVM machine learning model based on NLP problem, the TextCNN model based on convolutional neural network and the LDA theme model were established in this paper, which realized the label classification of the comments, the extraction of hot issues and the quality assessment of the government's replies.

Aiming at problem 1: in order to accurately classify the first-class tag of the message, this paper establishes a **SVM machine learning model based on NLP problem**. Parsing message level labels known first, and then to participle text processing, such as message details after computed text characteristic value, the sample vectorization, said it normalized, get training set and data set, with machine learning SVM model, then finally the F - Score index to evaluate the result of the classification and evaluation Score of **0.54**, shows that the SVM classification model is more accurate.

Aiming at problem 2: mining hot issues, defining reasonable heat evaluation index, and then sorting and summarizing hot issues. This paper using the K - means clustering algorithm to initial clustering of messages, according to the initial clustering results, this paper introduced **time attenuation mechanism**, five heat evaluation index is defined as the **average time attenuation values, message number, average length of text messages, the number of people in favor and the number of people against** and then **TextCNN based on convolution neural network model** is established for the message further classification, and then through the analytic hierarchy process (AHP) to calculate heat five evaluation index weights, the five heat index weighted summation of heat value, the higher heat value, indicates that the problem should be taken seriously, finally according to the size of the heat value of hot issue.

Aiming at question 3: in order to accurately evaluate the quality of the replies from various departments, this paper establishes **LDA theme model** for quantitative analysis. Firstly, the text of the replies is preprocessed, and a new corpus and dictionary are constructed. Then, the text is trained according to the model to get the scores of the three indicators of relevance, completeness and timeliness of the replies. The **weighted sum** of the three indicators is used to generate the score of each response. The higher the score, the better the response. Below is an average score for all responses:

Relevance	Integrity	Timeliness	Response score
0.62	0.63	0.56	63.31

Keywords: natural language processing, SVM, k-means clustering algorithm, TextCNN, LDA theme model

目录

一、问题重述

1.1 相关背景

随着网络的日益普及，互联网在中国民众的政治、经济和社会生活中扮演着日益重要的角色，成为公民行使知情权、参与权、表达权和监督权的重要渠道。政府官员通过各种形式在网上与百姓沟通，了解民情、汇聚民智，以达到取之于民，用之于民，使得政府的信息更加透明畅通，从而实现科学决策、民主决策，真正做到全心全意为人民服务。

近年来，各类社情民意相关的文本数据量不断攀升，使得人工划分留言和整理热点等工作极为冗杂。因此，借用大数据对政务系统进行智慧管理对于提升政府的管理水平和施政效率尤为重要。

1.2 题设数据

题目提供了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。

（1）附件 1 是网络问政平台下，根据留言内容所得的内容分类三级标签体系，以分支的形式展现了由社会各层面到具体问题的明细分类。

（2）附件 2 和附件 3 则是包含了留言用户、留言时间及具体留言内容，前者进行了一级分类，后者涉及到了留言所获其他网民的点赞、反对数。

（3）附件 4 是在留言信息的基础上，添加了答复时间与意见。

1.3 需要解决的问题

问题一：对群众留言进行分类，参考内容分类三级标签体系，对附件 2 所提供的留言数据建立关于内容的一级标签分类模型。

问题二：对热点问题进行挖掘，根据附件 3 将某一时段内反映特定特点或特定人群问题的留言进行分类，并定义合理的热度评价指标，得出评价结果。

问题三：对答复意见进行评价，针对附件 4 中各相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套定性或定量的评价方案。

二、数据预处理

因为留言质量良莠不齐，而模型的效果非常依赖于琐碎的数据清洗工作，所以本文首先对题设数据进行一定的处理，以增加程序的效率和鲁棒性。

2.1 自定义辅助分词词典

首先利用 python 中处理中文的文本分析包 jieba 对带有附件进行初步的分词处理；通过对比分词前后的文本，可以看出有些文本分词存在误差，比如在“水一方大厦”这个 A 词本来是一个整体，但是分词的时候会误认为分成“在水一方 B 词”和“大厦 C 词”。某个词 A 被切分为更小的错误词汇 B 和 C，导致我们文本处理存在误差.

为了提高处理的精度，本文在原本的自定义辅助分词词典的基础上，通过浏览整体数据，人工操作加入了辅助分词的自定义词汇,多数为小区名、街道名

一类的地名和政策条例。因此。在后续工作中，程序分词时会自动加载该词典完成准确的分词任务，使结果更加精确。

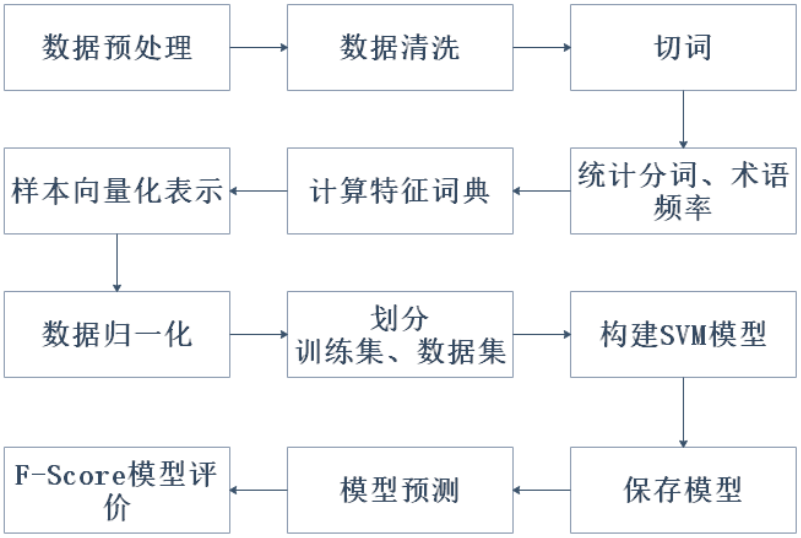
2.2 自定义停用词词典

同时，因为留言的人写作能力良莠不齐，导致文本的格式、口语化、正规化程度不同，留言文本较为琐碎。一些琐碎的、无意义的词语会增加程序处理数据的负担并且降低效率，影响准确度。本文自定义停用词词典加入一些口语化的、无意义、没有区分度的词，例如“呢”、“了”、“　此致敬礼”、“尊敬的胡书记”等。

分词处理会过滤掉这些词汇。因此，后续工作会更加简便，对于特征提取更加准确有效。

三、群众留言分类

3.1 流程图



图一：问题 1 解题流程图

3.2. 一级标签分类模型的建立

针对本问题，本文采用基于 NLP 问题的 SVM 机器学习方法，该方法是常见的监督学习方法，该算法建立在统计学理论的基础上，在计算机视觉、自然语言处理及生物信息学等方面有广泛的应用。

SVM[1]的主要思想是基于训练集数据在空间中找到一个使得类别之间间距最大的“超平面”，这样即可认为分类效果达到最佳，用数学语言表示为：

训练集样本为：

$$D=\{(x_1,x_{11}),(x_2,x_{12}),\cdots,(x_m,x_{1m})\}$$

其中 x_i 表示训练集中第 i 条留言的留言详情； x_{1i} 代表训练集中第 i 条留言的一级标签类型。

划分的超平面可以通过线性方程描述为：

$$\omega^T x+b=0$$

其中，$\omega=(\omega_1,\omega_w,\cdots\omega_d)$为法向量系数，决定该超平面的方向，位移项$b$决定超平面的具体位置，该平面记为$(\omega,b)$。

样本空间中任意的样本x到该平面的距离则可以表示为：

$$\gamma=\frac{\left|\omega^T x+b\right|}{\|\omega\|}$$

超平面(ω,b)如果能将所有的训练样本正确地分类，则有：

$$\omega^T x_i+b\geq+1, x_{1i}=+1$$

$$\omega^T x_i+b\geq-1, x_{1i}=-1$$

距离超平面最近的训练样本x_i，即使得上式成立的训练样本称之为“支持向量”，两个属于不同类别的支持向量的距离定义为“间隔”，数学公式表示为：

$$\gamma=\frac{2}{\|\omega\|}$$

支持向量机的基本原理就是使得“间隔”最大的超平面，找到满足分割条件的参数ω和b，使得γ最大，即：

$$\max_{\omega,b}\frac{2}{\|\omega\|}$$

$$s.t.\ x_{1i}\left(\omega^T x+b\right)\geq 1, i=1,2,\cdots,m$$

其等价于：

$$\max_{\omega,b}\frac{1}{2}\|\omega\|^2$$

$$s.t.\ x_{1i}\left(\omega^T x+b\right)\geq 1, i=1,2,\cdots,m$$

显然，这是一个凸二次规划最优解求解问题，本文接下来对这一公式进行具体的推导演算。对等价式使用拉格朗日乘子法可得到其“对偶问题”，每条约束添加拉格朗日乘子$\alpha_i\geq 0$，则该问题可写成：

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2+\sum_{i=1}^{m}\alpha_i\left(1-x_{1i}\left(\omega^T x_i+b\right)\right)$$

其中$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_m)$，$\alpha_i$为 Lagrange 系数，其数值必须为正数或零，令$L(\omega,b,\alpha)$对ω和b的偏微分为零，得到：

$$\frac{\partial L}{\partial\omega}=0\Rightarrow\omega=\sum_{i=1}^{m}\alpha_i x_{1i}x_i$$

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{m}\alpha_i x_{1i}=0$$

将上式代入拉格朗日乘式，消去ω和b，得到如下对偶问题：

$$\begin{cases} \max\limits_{\alpha} \sum\limits_{i=1}^{m} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{m} \alpha_i \alpha_i x_{1i} x_{1j} x_i^T x_j \\ s.t. \sum\limits_{i=1}^{m} \alpha_i x_{1i} = 0, \\ \alpha_i \geq 0. i = 1, 2, \cdots, m \end{cases}$$

解出 α 后，进而可解得模型如下：

$$f(x) = \omega^T x + b$$
$$= \sum_{i=1}^{m} \alpha_i x_{1i} x_i^T x$$

3.3 分类模型的评价

3.3.1 评价方法

通过查阅分类模型的评估方法，多数分类情况下，与分类模型得到的结果是否准确相比，大家更关心预测分类的出错率，因为错误事件对现实需求影响更大，例如本题对留言进行一级标签分类，若分类错误，那么相应的二、三级标签也会相应出错，错误的留言分类将会导致后期更多繁冗的工作，因此引入了“精确率（Precision）”，在本题中即为：

$$\frac{\text{分类正确的留言数}}{\text{分类正确的留言数+分类错误的留言数}}$$

然而，当分类错误的留言数为零时，精确率就会达到 100%，显然这是不合理的，因为“精确率”没有考虑到“未能成功分类的留言”中是错误分类的情况，因此就需要用到“召回率（Recall）”，在本题中即为：

$$\frac{\text{分类正确的留言数}}{\text{分类正确的留言数+未分类的正确留言数}}$$

F-Score[2]则同时兼顾了分类模型的精确率和召回率，可看作是模型精确率和召回率的一种调和平均，是衡量特征类别间分辨能力的有效方法，F_1 分数则是认为召回率和准确率同等重要，可以表示为：

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

其中，i 表示第 i 类一级标签，P_i 表示第 i 类一级标签的精确率，R_i 表示第 i 类一级标签的召回率。

3.3.2 评价结果

根据 SVM 模型的机器训练学习，本文发现附件 2 有着一级标签的留言文件中的留言只涉及到了城乡建设、党务政务、国土资源、环境保护、纪检监察、交通运输、教育文体共七类一级标签，用上述评价方法中提到的 F_1 分数计算出各类一级标签所对应的分类准确率、召回率以及 F-Score 指标，如下表所示：

表 1：分类模型的评价结果表

	分类准确率	召回率	F-Score 指标	支持向量数量
城乡建设	0.54	0.48	0.51	601
党务政务	0.60	0.49	0.54	293
国土资源	0.42	0.42	0.42	170
环境保护	0.67	0.73	0.70	499
纪检监察	0.59	0.62	0.61	602
交通运输	0.52	0.47	0.49	346
教育文体	0.44	0.61	0.51	252

表 2：分类模型的评价结果汇总表

	分类准确率	召回率	F-Score 指标	支持向量数量
汇总	0.54	0.55	0.54	2763

由上述表格可以看出，SVM 分类模型得到的结果在“分类准确率”和“分类召回率”两个方面是较为相近的，说明了此分类模型的客观合理性。从各类一级标签所应各指标的情况，可以看出“准确率”和“召回率”总体成反比，F-Score 指标的值刚好平衡了二者的偏差，给出较为准确的评价值，纵向观察，可以看出不同的一级标签的 F-Score 指标多数在 0.5 左右，说明 SVM 分类模型的分类情况较于稳定；从分类模型的评价结果的汇总表中可以发现，与上表的特征是较为吻合的，分类准确率、召回率、F-score 指标均是在 0.5 附近，更进一步印证了 SVM 分类模型的稳定性。

四、热点问题挖掘

4.1 问题分析与模型准备

4.1.1 问题分析

某一时段内群众集中反映的某一热点问题往往是政府急需尽快解决的问题，及时发现网络留言中的热点问题，有助于相关部门进行有针对性地处理，进而提高为广大人民的服务效率。

因此，需要对留言中反映特定地点、特定时间或特定人群问题的留言进行较为准确的分类，由于附件内留言数据将为庞大，分类模型仅仅可以对其进行较为宽泛的聚类，本文将采取基于卷积神经网络的 TextCNN 模型。

为了进一步得到准确的热点问题，本文将会定义“热度评价指标”，初步想法是，将“留言个数”、“留言反对数”、“留言点赞数”暂定为评价指标，后期聚类后，再根据聚类结果适当的增加合适的指标。最后，可以通过层次分析法计算各评价指标的权重，进而求出每条留言所对应的热度指数值。

4.1.2 留言信息预处理

本题同样要对大量的文本信息进行处理，本文将基于问题一确定的“用户自定义词典”，继续采用 jieba 分词，对留言信息进行分词处理，将一段文本序列划分为合理的词(字)序列,最后去除干扰页面有效信息的停用词。

4.1.3 留言信息的向量化

本文使用 Word2Vec 的 CBOW 训练模型[3]对留言信息进行向量化。该模型

是由输入层、投影层、输出层三层构成的神经网络，其中输入层是指当前词前后各 c 个词向量，$v\left(Context\left(w\right)_1\right)$、$v\left(Context\left(w\right)_2\right)$、$\cdots v\left(Context\left(w\right)_{2c}\right)$，在模型训练前初始词向量是随机赋值的，通过多次训练实现不断地更新。投影层把输入层地所有词向量放在一起执行简单地加法：

$$x_w=\sum_{i=1}^{2c}v\left(Context\left(w\right)_i\right)$$

接下来是输出层，输出最有可能的当前词 ，实际上可以表示成一棵二叉树，把文本集里所有地词汇作为叶子结点，每个词出现地频率当作权值创建 Huffman 树。

把输入层所有词向量执行向量加法到投影层 ，因为输出层可表示为一棵二叉树，再将 X_w 传入输出层时，每个分支可以看作二分类问题，二分函数即为 sigmoid 函数。（θ 向量是一个未知参数）

$$\sigma\left(\theta^Tx\right)=\frac{1}{1+e^{-\theta^Tx}}$$

结点被化到正类的概率：

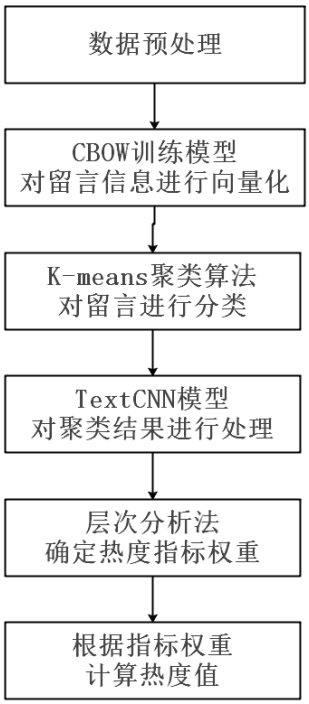
$$\sigma\left(x_w^T\theta\right)=\frac{1}{1+e^{-x_w^T\theta}}$$

而被化为负类的概率就是 $1-\sigma\left(\theta^Tx\right)$，将各个分支概率累乘，求得概率函数 $P\left(W\mid Context\left(W\right)\right)$ 将上述地概率函数取对数得到：

$$\sum_{w\in C}\log p\left(w\mid Context\left(w\right)\right)$$

分别对上式中 X_w 和 θ 求偏导，使得对数概率函数值最大。

4.2 流程图



图二：问题 2 解题流程图

4.3K-Means 初步聚类

为了对留言进行热点问题提取，本文先采用 K-means 聚类算法[4]对留言进行初步聚类分析，具体步骤为：

（1） 选择初始化的 K 个样本作为初始聚类中心：$c=\{a_1,a_2,\cdots,a_k\}$

（2） 针对数据集中每个样本 x_i 计算它到 K 个聚类中心的距离并将其分到距离最小的聚类中心所对应的类别中；

（3） 针对每个类别 ，重新计算它的聚类中心：$a_j=\frac{1}{|c_i|}\sum_{x\in c_i}x$

（4） 重复上述两步操作，直到聚类中心不再变化。

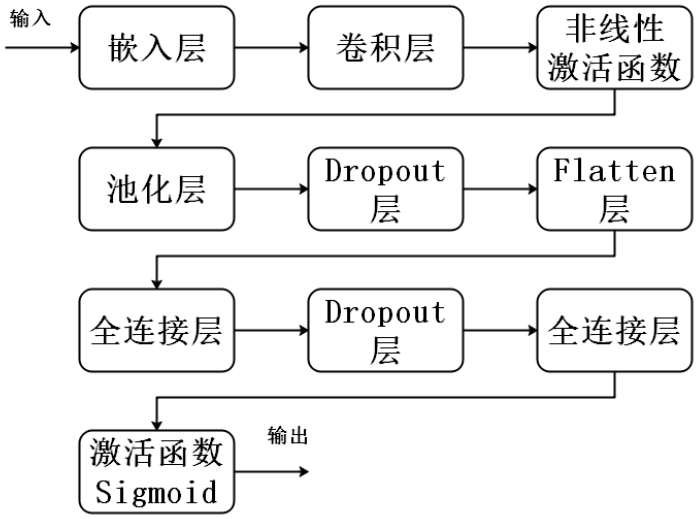
按照上述聚类约束准则对 4326 条留言进行聚为 30 类，通过观察所得聚类的留言的特点，本文引入时间衰减机制，并将五个热度评价指标定义为平均时间衰减值、留言个数、平均留言文本长度、反对数、点赞数。

由于 K-means 聚类分析是无监督聚类，聚类结果过于粗糙，未得到更为精准的分类，进而得到更为合理的热点问题，下文使用基于卷积神经网络的 TextCNN 模型对 30 类留言进行更加精细的聚类。

4.4 基于卷积神经网络的 TextCNN 模型

4.4.1 模型简介

卷积神经网络(convolutional neural network, CNN)是一类包含卷积计算且具有深度结构的前馈神经网络(Feedforward Neural Networks)，是深度学习(Deep Learning)的代表算法之一。首次应用于文本分类是在 2014 年纽约大学 Yoon Kim 提出的 TextCNN 模型。



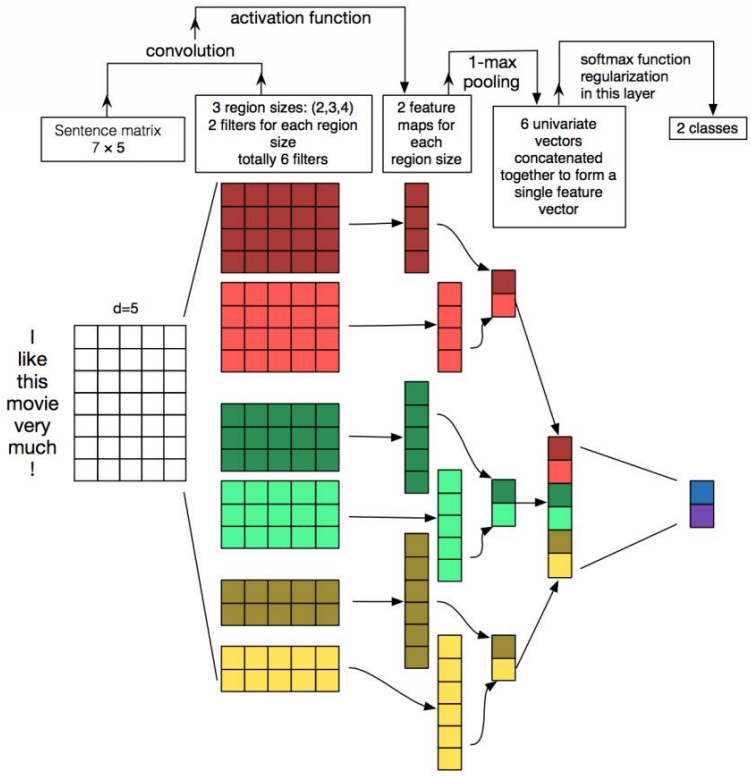
图三：基于卷积神经网络的 TextCNN 模型的流程图

基于卷积神经网络的 TextCNN 模型[3][5]首先将词向量化后的训练数据经过嵌入层，由加权模型将已训练好的的词向量作为初始二维矩阵输入到嵌入层，然后此二维矩阵在之后模型迭代过程中，也是不断的变化更新的。通过卷积层，由特定卷积核对之前嵌入层输入的词向量执行卷积操作，来提取多组局部特征。使用激活函数，来实现非线性映射。池化层使用最大池化，来实现降

维，从卷积层获得的特征中提取出某些局部最优特征。接着 dropout 层模型训练时随机舍掉部分神经元，防止出现过拟合现象。Flatten 层将之前的结果展平成一维数据。全连接层能够输出结果的维度，最后一个全连接层用 sigmoid 激活函数，实现将结果转化成[0,1]区间，进而用来预测类别概率。

4.4.2 模型的建立

卷积神经网络(convolutional neural network, CNN)由特征提取层和特征映射层构成。特征提取主要包含每个神经元的输入与其上一层局部接受域。每个特征映射层是一个平面，每个特征映射层上所有神经元的权值相等。



图四[3]：TextCNN 卷积神经网络结构示意图

Step1:嵌入层

对输入深度神经网络的文本数据进行词向量化训练，向量化的模型即为模型准备中的 Word2Vec 的 CBOW 训练模型。

Step2:卷积层

$$h_{ij}^{k}=g\left(\left(w^{k}*x\right)_{ij}+b_{k}\right)$$

其中，$*$表示卷积操作；x_i为词w_i的表示；w^k表示权重；b_k表示偏移量；$g(\bullet)$为激活函数。

得到若干个隐藏层后，卷积神经网络常会采用最大池化，将不定长度的隐藏层压缩到固定长度的隐藏层中。

Step3:池化层

池化层在卷积层之后，主要作用是进一步采样处理卷积层输出的特征图。类似于卷积层，池化层将提取到的特征看作一个矩阵。采用最大池化来对矩阵进行处理，输入区域分割成若干个子区域，输出每个子区域的最大值。最大池化特征点的计算公式为：

$$h_{m,j} = \max_{i \in N_m} \alpha_{i,j}$$

其中，$\alpha_{i,j}$ 是卷积操作后输出的特征点，$h_{m,j}$ 是最大池化后输出的特征值。

Stept4:全连接层

该层的输入为池化操作后形成的一维向量，经过激活函数输出，再加上 Dropout 层防止过拟合，dropout 设定为 0.5，并在全连接层上添加 12 正则参数化。该层的输入作为全连接层的输出，最后经过 Sigmoid 激活函数后输出留言信息的分类。

4.5 层次分析确定指标权重

1. 建立层次结构模型[6]

本文建立了基于“时间、内容、关注度”的热度评价模型。对于留言时间的分析，加入时间衰减机制，即留言事件发生越久，随着时间迁移对当前的影响会越小；一个问题衰减后的值越高，该问题相应的影响力也越大，在计算热度指数时贡献的权重也越大。

t 为间隔天数；*init* 为初始衰减值；*m* 为时间衰减长度；*finish* 为完成衰减值；*decay* 为时间衰减值。

$$\alpha = \frac{1}{m} \ln \frac{init}{finish}$$

$$\beta = -\frac{1}{\alpha} \ln init$$

$$decay = e^{-\alpha(t+\beta)}$$

对于留言内容，则分为留言数量和留言文本平均长度两个指标；关注度则是分为反对数和点赞数两个指标。

构建留言热度评估体系如下：

表 3：留言热度评估体系表

目标层	评估元素	评估指标
评价留言热度(Q)	留言时间(A1)	时间衰减值(B1)
	留言内容(A2)	留言数量(B2)
		留言文本平均长度(B3)
	留言关注度(A3)	反对数(B4)
		点赞数(B5)

2. 构造判断矩阵

层次分析法中构造判断矩阵的方法是一致矩阵法，即：不把所有因素放在一起比较，而是两两相互比较；对此时采用相对尺度，以尽可能减少性质不同因素相互比较的困难，以提高准确度。a_{ij} 为要素 *i* 与要素 *j* 重要性比较结果，有如下性质：

$$a_{ij} = \frac{1}{a_{ji}}$$

表 4：各因素的相对重要度表

因素 i 比因素 j	量化值
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2，4，6，8

根据上述重要度表，我们构造了相应的判断矩阵：

$$Q=\begin{bmatrix}1 & \frac{1}{2} & \frac{1}{3}\\ 2 & 1 & 3\\ 3 & \frac{1}{3} & 1\end{bmatrix}$$

$$A2=\begin{bmatrix}1 & 5\\ \frac{1}{5} & 1\end{bmatrix} \qquad A3=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$$

3. 判断矩阵的权值计算

假设判断矩阵为:

$$A=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix}$$

利用特征值法计算判断矩阵的权值，步骤如下：第一步计算矩阵 A 的最大特征值及其对应的特征向量；第二步对特征向量进行归一化，得到我们需要的权值；

4. 一致性检验

我们计算判断矩阵的最大特征值 $\lambda_{\max}$ 及其一致性指数 CI：

$$CI=\frac{\lambda_{\max}-n}{n-1}$$

其中 n 是矩阵的维数。$CI=0$，有完全的一致性；CI 接近于 0，有满意的一致性；CI 越大，不一致越严重。为衡量 CI 的大小，引入随机一致性指标 RI：

$$CR=\frac{CI}{RI}$$

如果 $CR<0.1$，判断矩阵的一致性是可以接受的。最后，我们计算各指标的权重如下:

表 5：A 指标权重

指标	权重
留言时间（A1）	0.19
留言内容（A2）	0.43
留言关注度（A3）	0.38

表 6：B 指标权重

指标	权重
时间衰减值（B1）	0.19
留言数量（B2）	0.36
留言文本平均长度（B3）	0.07
反对数（B4）	0.19
点赞数（B5）	0.19

按照权重加权求和即为热度值。每个问题对应的热度指数取值在[0，100]，热度指数值越大，表明该问题越应该受到重视。

4.6 五大热点问题

根据所建立的模型，得出热度指数排名前五的问题如下表：

表 7：热度指数排名前五的问题

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	69.61	2019.11.02 至 2020.01.26	A 市 A2 区丽发新城小区居民	小区附近搅拌站扰民
2	8	51.27	2018.07.07 至 2019.09.01	A 市伊景园滨河苑居民	小区强行捆绑销售停车位
3	4	41.79	2019.01.11 至 2019.12.15	A 市居民	58 车贷诈骗案无进展
4	3	37.6	2018.11.15 至 2019.12.15	A 市居民	购房资格与补贴的咨询
5	6	29.74	2019.07.21 至 2019.09.25	A 市 A5 区魅力之城小区居民	临街餐饮店油烟和噪音扰民

由表格可以得出，A 市小区附近扰民的问题较为严重，亟待解决，建议出台居民区噪声管理的有关政策；除了 58 车贷，还存在余利宝贷款平台诈骗的问题，建议有关部门关注贷款现象，及时采取有效的措施；在热度分析中，发现关于购房资格和购房补贴方面留言很多，除此还有关于 A 市住房公积金的政务问题咨询，建议 A 市完善住房的一系列相关政策。

排名前十的其他问题中，还存在渣土车扰民、学校收费情况不明、人才引进咨询的问题，建议有关部门加强管理。

五、答复意见质量评价方案

5.1 评价模型的整体介绍

针对相关部门对留言的答复意见，主要从 3 个方面构建答复意见的质量评

价模型：留言回复的相关性、完整性和及时性。附件 4 是非标记数据，基于此，模型采用无监督方式加权 3 个影响整体评价的方面，最终得到每条答复意见的质量得分。

1. 相关性评分

取值[0,1]。评价回复内容的优劣，其中一个重要的因素是回复是否跟留言相关以及相关性的大小。采用 LDA 主题模型(可根据给定的一段文档，反推其主题分布)来量化问答对之间的相关性大小，该算法可以有效捕捉文本背后隐含的深层次语义距离，从而挖掘出更多有意义的文本信息。使用给出的数据构建语料库进行去停用词、分词预处理，训练主题模型，留言详情和答复意见分别可以表示为主题分布向量，通过计算向量之间的余弦相似度从而可以得到文本相关性。

2. 完整性评分

取值[0,1]。通常来说，留言回复内容的详细程度与回复的文本长度有直接的关系。回复内容过于短小，其中包含的信息量不足，评分应该较低；同时，较长文本的回复评分不应该过高。在上述情况下可以使用 sigmoid 函数(一个有着优美 S 形曲线的数学函数，在逻辑回归、人工神经网络中有着广泛的应用)来建立完整性评分模型。

3. 及时性评分

取值[0,1]。答复及时性可以反映相关部门对留言的重视，答复越及时，评分越高，基于此，可以采用时间衰减机制来衡量答复意见的及时性评分。

5.2 评价流程

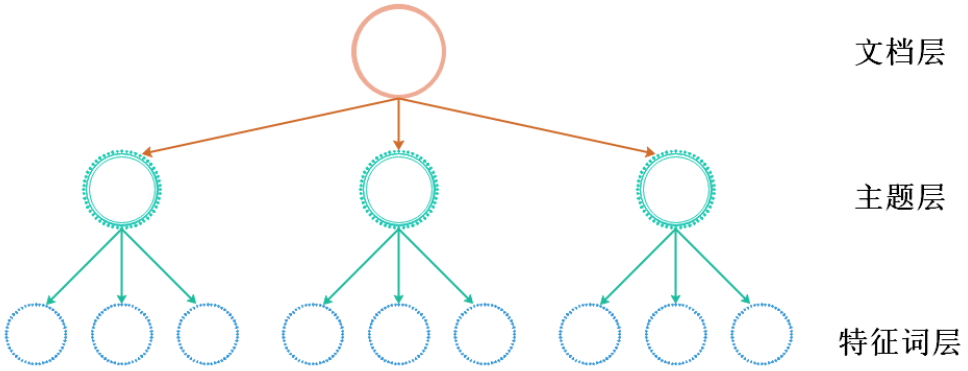
图五：答复意见质量评价流程图

5.3 LDA 主题模型的建立

5.3.1 模型简介

LDA 其实有两种含义，一种是线性判别分析(Linear Discriminant Analysis)，一般用来为后续的分类问题做数据的降维处理；另一种便是本文所使用的主题模型(就是一种自动分析每段文字，统计文档中的词语，根据统计的信息判断当前文档包含哪些主题以及各个主题所占比例各为多少)，它称作为：隐含狄利克雷分布(Latent Dirichlet Allocation，简称 LDA)[7][8]，是一种概率主题模型。

LDA 主题模型依次有文档层、主题层、特征词层三个非常清晰的层析，实质就是利用文本的特征词的共同特征来挖掘文本的话题，其结构如图所示：



图六：LDA 模型结构示意图

5.3.2 模型的建立

LDA 的主题思想是：整个文档集是基于主题的概率分布，而每个主题又是基于特征词的概率分布，则表示某词在，某答复意见中出现的概率公式为：

$$p(w_n \mid M_m) = \sum_{k \in K} p(w_n \mid K_k)\, p(K_k \mid M_m)$$

上式表示词 w_n 出现在答复意见 M_m 中的概率。其中，$n \in N$，N 表示特征词的总数，$m \in M$，M 表示答复意见的条数，$k \in K$，K 为主题的总数。

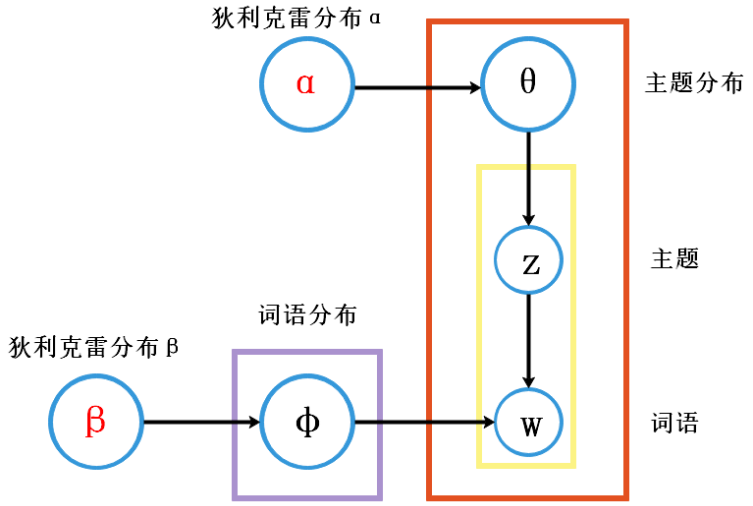
对于“文本-主题”分布过程来说，首先该分布需要服从参数为 α 的 *Dirichlet* 的先验分布，*Dirichlet* 分布是多维的 *Beta* 分布。这里给出 *Beta* 分布的密度函数，如下式所示：

$$\begin{aligned} f(p,\alpha,\beta) &= \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &= \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \end{aligned}$$

其中，$B(\alpha,\beta)$ 表示参数 (α,β) 的 *Beta* 分布。p 一般表示事件发生的概率，K 维的 *Dirichlet* 分布式如下式所示：

$$Dirichlet\left(\overline{p}\,|\,\overline{\alpha}\right)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

从上式可以看出，*Beta* 分布式 *Dirichlet* 分布在二维状态下的特殊形式，LDA 模型就是需要对参数(α,β)进行确定，进程如图所示：



图七：LDA 主题文档生成图

上图中，词分布用ϕ表示，主题分布用θ表示，主题分布θ的先验分布（即 *Dirichlet* 分布）参数α，词分布ϕ的先验分布参数β，文档的单词总数N，M表示文档的总数，且服从下式分布：

$$\vec{\phi}\sim Dirichlet\left(\vec{\alpha}\right)$$

$$\vec{\theta}\sim Dirichlet\left(\vec{\beta}\right)$$

LDA 主题模型的训练过程，主要是参数α和β进行训练，其典型代表是 *EM* 估计和 *Gibbs* 抽样。

5.3.4 模型的求解

通过 python 程序对 LDA 主题模型进行训练，计算得到留言详情与答复意见之间的相关性评分、答复意见的完整性评分、及时性评分和最终的答复评分，整体打分情况在命名为“答复评价结果表”的 Excel 表中呈现，现只截取一部分展示在下列表格中：

表 8：答复意见的三指标及最终评分表

留言编号	留言用户	相关性	完整性	及时性	答复评分
3684	UU008687	0.51	0.63	0.48	54.24
33343	UU0081646	0.94	0.65	0.95	94.98
119181	UU0081847	0.56	0.65	0.63	62.55
164165	UU0081141	0.5	0.64	0.52	55.55
177384	UU008548	0.61	0.66	0.91	75.23
177915	UU0082351	0.69	0.67	0.87	78.32

由上表可以看出留言答复意见的相关性、完整性、及时性大多都达到了

50%的概率，答复的评分更直观的展现了某条留言的答复情况，这样将便于对未能较好答复的留言进行二次答复，以彻底地解决社会问题。

对 2816 条留言及答复评价进行分析得出其相关性、完整性、及时性和答复评分的均值如下：

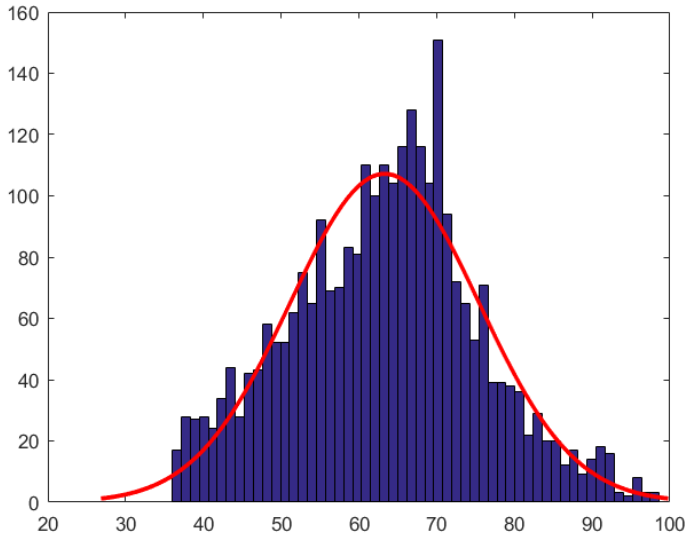
表 9：各指标评分均值表

相关性	完整性	及时性	答复评分
0.62	0.63	0.56	63.31

可以看出，在本文打分的模型标准下，政府有关部门的工作人员对留言的相关性、完整性的把握较好，在及时性方面欠佳，评分的平均分为 **63.91**，及格，客观的反映的政府对网络留言的答复情况。

5.4 结果分析

为直观的看出答复意见质量评价方案得到的评分情况，便于进一步的分析本文模型的合理性、客观性，本文展示了如下 2816 条留言的评分分布：



图八：答复意见评分的分布图

在上图中，横坐标表示留言答复意见得最终评分，纵坐标表示每个评分所有的数量，可以看出，留言及回复评价的得分基本符合正态分布。印证了本模型的稳定性。

六、结论

本文主要解决了智慧政务体系中的网络问政留言平台的留言分类、搜寻民意热点、留言答复评价的问题。结合基于 NLP 问题的支持向量机机器学习方法对留言进行分类；采用 textCNN 与 K-Means 聚类算法对留言归类，然后采用层次分析法计算指标权重，最后得出热点问题的热度值；针对答复意见，采用 LDA 主题模型来量化答复意见的质量，对指标加权求和得出评价得分。证明了上述算法对智慧政务体系的智能化改进具有出色的应用价值和重要意义，为智慧政务体系提供了新的研究方向。

然而，在本文中，利用 K-Means 聚类的无监督学习算法对热点问题的搜寻仍有一定的局限性，聚类结果不够纯粹；热度值的评价指标与评价得分指标的权重方面有一定的提高空间。

对于机器学习算法的应用上，可以尝试获取新的因子进行建模训练，进一步提升效果；咨询政务系统的工作人员，便于我们优化模型，能更有效地进行智慧政务体系的改良。

对于负责政务问题的工作人员，可以根据本文结果给出如下建议：

1、对政务信息和各类政策更加公开明确，让居民更容易地查询到有关信息，在街道、居委会加强宣传，可减少居民对一些政策的咨询留言；
2、住宅区应加强附近商业的管控，限制噪音的问题，可出台有关噪音分贝的规定，减少噪音扰民情况的投诉；
3、对于留言的及时性回复，可根据所反映问题的热度值来优先回复问题，热度较高的政务问题优先回复，提高热点问题回复的及时性；
4、可以根据留言的分类结果，对不同类型的留言设置不同部门或相关领域的工作人员来进行回复，提高留言回复的相关性和完整性；
5、可以根据留言的分类结果和热度，调整负责回复留言的工作人员的数量，就分类较多和热度值较高的政务问题，可以增多工作人员，实现人力资源分配的合理化；
6、对于热点问题，可以对留言者进行专访，增强对其所反映问题的了解程度，更好的处理相关及类似问题。

七、参考文献

[1]薛亮. 基于 SVM 的中文文本分类系统的设计与实现[D].重庆大学,2012.

[2]秦彩杰,管强.一种基于 F-Score 的特征选择方法[J].宜宾学院学报,2018,18(06):4-8.

[3]彭丹蕾. 商品评论情感分析系统的设计与实现[D].北京邮电大学,2019.

[4]毛郁欣,邱智学.基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J].中国信息技术教育,2020(08):99-101.

[5]刘春磊,武佳琪,檀亚宁.基于 TextCNN 的用户评论情感极性判别[J].电子世界,2019(03):48+50.

[6]Yoram Wind,Thomas L. Saaty. Marketing Applications of the Analytic Hierarchy Process[J]. Management Science,1980,26(7).

[7]程海琪. 基于情感分类的酒店评论短文本主题挖掘[D].浙江工商大学,2020.

[8]郭庆. 基于图与 LDA 的中文文本关键词提取算法[D].北京邮电大学,2019.