基于 NLP 的智慧政务系统管理研究

摘要

随着收集渠道的不断拓宽，待处理的社情民意文本数据量激增。为了提高政府相关部门的工作效率，由人工整理向机器智能识别的操作模式转型已成为当下的必然趋势。本文通过构建线性支持向量机分类模型以及基于 DBSCAN 的文本挖掘模型，有效提取关键信息，对市民反映的民生问题进行留言分类及热度排名，大幅减少人工操作的工作量。

在数据预处理阶段，首先针对数据原文件中重复的留言进行去重，将“留言主题”与”留言详情”中的自然句分词，并对文本中无意义的词语及标点符号进行去除停用词处理，使用 TF-IDF 权重将文本向量化表示，以此为基础构建留言内容的分类模型，并进一步进行关于热点问题的文本挖掘。

针对问题一，首先对一级标签进行编码，生成对应字典。其次，对“留言详情”文本进行词频统计，绘制词云图展示主要文本信息。接下来将数据集拆分为训练集与测试集，对所构建的 8 种分类模型进行评估，为解决数据不平衡问题，在拟合模型时通过设置参数 class_weight='balanced'更新不同类别样本权重，接着比较模型混淆矩阵反映的分类准确度及 F-Score 值的大小，最终得到线性支持向量机模型和逻辑回归模型的分类效果最佳，F-Score 值分别为 0.89、0.88，均处于较高水平。

针对问题二，首先对“留言主题”文本进行数据预处理得到词性矩阵。其次计算词频矩阵和 TF-IDF 权重矩阵，并将权重矩阵与词性矩阵相乘得到新权重矩阵，以此为根据建立 DBSCAN 聚类分析模型。在 DBSCAN 聚类分析模型中，将聚类半径和最小样本量均设定为 1，得到各留言的分类结果。最后将分类结果转化为数据框形式，以某一时间段内反映同一问题的留言数量作为热度评价指标，得到排名前五的热点问题。

针对问题三，首先通过观察了解数据特点，定义衡量答复意见质量的评价指标，分别从相关性、可解释性、可读性和答复效率四个方面对答复意见质量进行评判。将各变量进行标准化处理，根据实际情况分配以上四个指标的权重为(4,3,-2,1)，构建基于多元线性回归的答复意见质量评价模型，最终通过计算得到的数值大小衡量答复意见质量。

通过数据源文件对所建模型进行检验，问题一的线性支持向量机分类模型和问题二的 DBSCAN 聚类分析模型均得到较为准确的结果。

关键词：TF-ITF DBSCAN 聚类分析 线性支持向量机 多元线性回归

Abstract

With the continuous expansion of the collection channels, the amount of text data of social situation and public opinion to be processed increases sharply. In order to improve the work efficiency of relevant government departments, it has become an inevitable trend to transform the operation mode recognized by machine intelligence from manual sorting. By constructing the classification model of linear support vector machine and the text mining model based on DBSCAN, this paper effectively extracted the key information, classified the comments and ranked the popularity of the people's livelihood issues, and greatly reduced the workload of manual operation.

In data preprocessing phase, first of all, according to data repeat messages in the original file to heavy, details will message "theme" and "message" of the natural sentence segmentation, and meaningless words and punctuation in the text to remove the stop processing, use the TF - IDF weight to quantify the text said, on this basis, construct the message content classification model, and further on hot issues of text mining.

For question one, the first level label is coded to generate the corresponding dictionary. Secondly, the text of "message details" word frequency statistics, draw a word cloud map to show the main text information. Next data set into training set and testing set, to evaluate 8 kinds of classification model constructed, to solve the problem of unbalanced data, when the fitting model by setting the parameter update class_weight = 'balanced' different categories sample weight, by comparing the classification accuracy of the model reflects the confusion matrix and F - the size of the Score value, finally be linear support vector machine (SVM) model and the classification of the logistic regression model works best, F - Score values were 0.89, 0.88, are at a higher level.

For question two, the part of speech matrix is obtained by preprocessing the text of "message subject". Secondly, the word frequency matrix and tf-idf weight matrix are calculated, and the weight matrix is multiplied by the part of speech matrix to obtain a new weight matrix, on which DBSCAN clustering analysis model is established. In the DBSCAN clustering analysis model, the clustering radius and the minimum sample size were set as 1 to obtain the classification results of each message. Finally, the classification results were converted into the form of data box, and the number of comments reflecting the same problem in a certain period of time was used as the heat

evaluation index to get the top five hot issues.

For question three,firstly, the evaluation index to measure the quality of response comments is defined by observing the characteristics of data, and the quality of response comments is evaluated from four aspects: relevance, interpretability, readability and response efficiency. The variables were standardized, and the weight of the above four indicators was assigned as (4,3,-2,1) according to the actual situation. The quality evaluation model of the response based on multiple linear regression was constructed. Finally, the quality of the response was measured by the numerical value obtained through calculation.

The proposed model is verified by data source files. The classification model of linear support vector machine (SVM) in problem 1 and the DBSCAN clustering analysis model in problem 2 are both accurate.

Key word: TF-IDF DBSCAN cluster analysis Linear Support Vector Machine Multiple Linear Regression

目录

表录

图录

一、绪论

1.1 挖掘背景

在信息化时代，互联网的高速发展和大数据技术的日益成熟为人们的生活带来了巨大的便利，普及的领域涵盖金融、计算机、军事、生物医学、统计学、社会科学等多种学科，为人们获取海量多元的信息、作出更为科学精准的决策提供了技术上的支持。而自然语言处理和文本挖掘作为计算机科学领域和人工智能领域的一个重要分支，也正在不断发展，应用到各个领域的生产和工作中。

在政府的民生工作中，听取、收集民意是与民生产生直接沟通的重要一环，能否及时调查和处理市民反映的问题、如何进行更有效的反馈，直接反映了政府工作能力的优劣，也直接影响着市民的生活幸福指数。但在以往的民意收集工作中，大量的文本数据只能通过人工进行逐条筛选分类，巨大的工作量使得政府要耗费更多的劳动力和时间，进一步导致民意反馈的滞后性，引发更多民生问题。而在信息化的背景下，计算机的自然语言处理和文本挖掘模块能够提供极大帮助。因此，面对海量文本信息，计算机代替人工处理能够有效提高政府的工作效率，是未来民生服务结构优化升级的必然趋势。

构建自然语言处理和文本挖掘模型，目的在于使计算机在面对海量文本数据时，能够根据人们的具体需求进行筛选、分类及汇总，过滤无用信息，提取有效信息，大大减少时间及人力成本，提升用户体验。因此，本文基于自然语言处理和文本挖掘对民意收集模块进行改善，具有研究价值和重要意义。

1.2 挖掘目的

本文的目的为构建一个民意文本数据处理模型。该模型能够按照划分体系对市民留言进行分类，并且对市民集中反映的问题进行提取和排序，政府能够根据民生问题划分后的类别直接安排相应部门进行跟进处理，同时也能提取热点问题信息以便重点关注。

将该模型映射到二维空间中，可将问题简化为关于留言 A 和类别 B 的二元模型求解。即将留言 A 根据特定模型归到相应的类别 B 中去，并提取在特定标准下出现次数最多的留言 A 作为热点问题。

1.3 挖掘流程

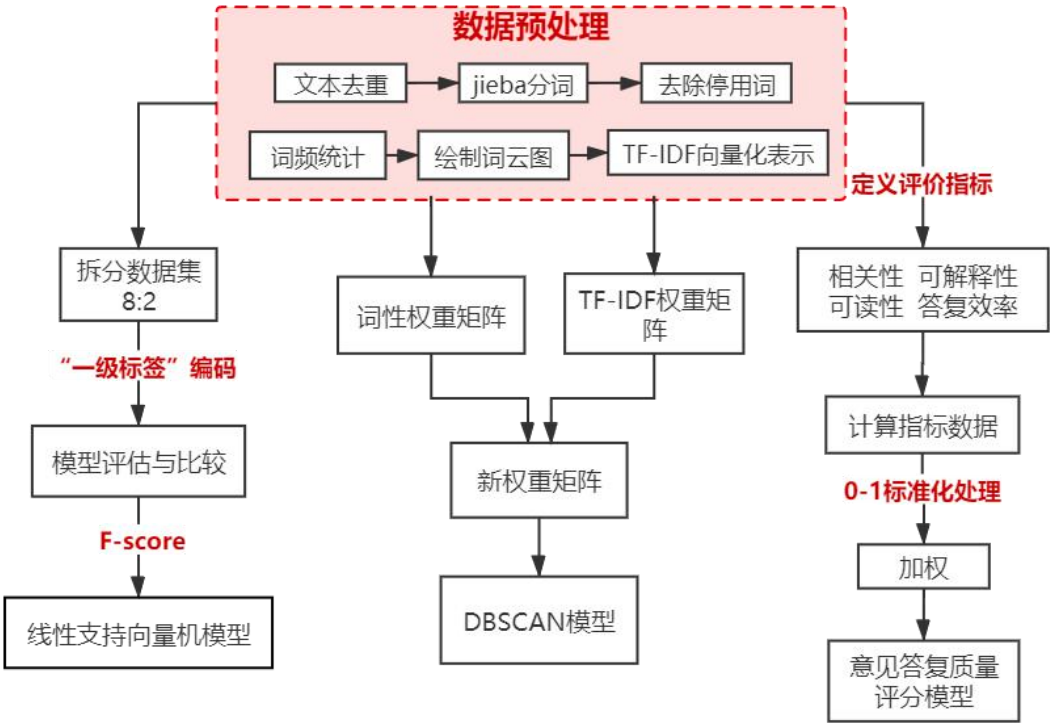

图 1 挖掘流程图

二、问题分析

2.1 问题一分析

问题一中将群众留言进行分类属于机器学习中的监督式学习，本题重点在于提取不同标签下的文本特征。对附件 2 文本进行分析，首先将一级标签进行数字编码，在对“留言详情”文本完成去重、分词、去除停用词等数据预处理后，建立 TF-IDF 模型，使文本数据向量化。为避免出现过拟合现象，本文将附件 2 数据集按各“一级标签”比例进行拆分，包括 80%的训练集以及 20%的测试集。对训练集分别建立朴素贝叶斯多项式模型、线性支持向量机模型、随机森林模型、逻辑回归模型、KNN 等模型，利用所建模型对训练集中的留言类别进行预测，计算各模型对应的 F-Score，最终确定 F-Score 值最高的模型作为本题的分类模型。

2.2 问题二分析

对附件 3 给出的留言数据进行分析。由于“留言主题”文本蕴含“留言详情”文本的主要信息，因此本题只对“留言主题”进行文本挖掘。“留言主题”中包含的标点及语气助词、

量词等没有具体含义，经过去除后对文本进行结巴分词，得到包含“留言主题”中词语及词性的字典，对字典中的短文本进行特征提取重新得到权重矩阵，并建立 DBSCAN 模型，以某一时间段内反映同一问题的留言数量作为热度评价指标，最终得到不同留言的类别以及数量排名前五的热点问题。

2.3 问题三分析

对附件 4 中的“答复意见”总体进行对比分析，了解各答复意见的特点，从而定义评价答复意见质量的四个指标，分别为相关性、可解释性、可读性和答复效率。其中，相关性通过留言详情文本和答复意见文本的样本相似度衡量，可解释性通过答复意见中是否引用书面材料来反映，可读性通过答复意见文本的有效字符数衡量，答复效率通过留言时间与答复时间的时间间隔长短反映。根据实际情况分配以上四个指标的权重，根据多元线性回归思想构建答复意见质量的评价模型，最终通过计算得到的数值大小衡量答复意见质量。

三、数据预处理

3.1 去重

在数据的的储存和提取过程中，由于技术和某些客观的原因，经常会出现部分文本重复。附件 1 数据文件中存在大量留言重复现象，留言信息冗杂会对分类效率产生影响，因此需对重复留言进行去重，仅保留重复文本中的一条记录，共删除重复留言 158 条。

3.2 分词

由于中文文本的词语之间没有明显的界限，因此从文本中提取词语时需要进行分词。本文采用 Python 中的 jieba 库进行分词，结巴分词(jieba)库支持简体分词、繁体分词以及自定义词典三种模式。对于问题一，对“留言详情”文本进行分词，将 jieba 分词不合理的地方，如”A1”“市”，通过自编函数进行合并修改，作为构建分类模型的基础。对于问题二，利用 posseg 词性标注分词器产生关于“留言主题”的字典。其中，字典的键为词语，值为词性。利用该字典可以得到权重矩阵，并进一步建立 DBSCAN 模型。对于问题三，对“留言详情”和“答复意见”分别进行分词处理，进一步计算文本相似度以反映相关性。

3.3 去除停用词

分词后文本中存在许多类似“的、了、呢、吗、也”等中文表达常用的功能性词语，以

及标点符号、语气助词、量词等，这些词语对于文本数据的研究并没有太多的实际含义，还有可能会对文本分析造成负面影响，因此需对文本进行去除停用词处理，减少无用信息并保留关键词，能有效提高模型运行速度及准确率。

3.4 绘制词云图

词云图是对文本结果展示的有力工具，通过词云图的展示可以对文本数据分词后的高频词予以视觉上的强调突出效果，使得阅读者及研究者一眼就可获取到主旨信息。

3.5 文本的向量化表示

对文本数据进行分析，需将文本转为数字形式。文本向量化指的是通过词向量获取到句子/段落向量的向量化方法，主要的方法有权重叠加法和模型法。其中模型法由于有着较高的计算成本[1]，而最简单的词袋模型虽简单易懂，但也存在问题，即没有对文本中出现的词赋予权重。本文采用 TF-IDF 模型，该方法考虑了文本的词频，文档中高频词应具有表征此文档较高的权重，一个词的权重由 TF*IDF 表示，其计算公式如下：

$$\text{词频}(TF)=\frac{\text{某个词在文章中的出现次数}}{\text{文章的总次数}} \tag{1}$$

$$\text{拟文档频率}(IDF)=\log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}\right) \tag{2}$$

四、实验评估

4.1 实验环境

本文的实验环境配置如表 1 所示：

表 1 实验环境配置表

配置	参数
CPU	Intel i7
显卡	GTX 1080
内存	16GB
操作系统	Win 10
Python	3.5.2
CUDA	8.0
CuDNN	5.1

4.2 实验评价指标

4.2.1 F-Score

首先定义混淆矩阵，如表 2 所示：

表 2 混淆矩阵定义

	预测值 0	预测值 1
真实值 0	TN	FP
真实值 1	FN	TP

其中 TP 为正类项目被判定为正类，FP 为负类项目被判定为正类，FN 为正类项目被判定为负类，TN 为负类项目被判定为负类。

由此得到：

精准率 P：预测为正类的项目中真实为正类的项目，公式如式（3）所示。

$$P = \frac{TP}{TP + FP} \tag{3}$$

召回率 R：真实为正类的项目中预测为正类的项目，公式如式（4）所示。

$$R = \frac{TP}{TP + FN} \tag{4}$$

F-Score：为综合考虑精准率和召回率的调和值，公式如式（5）所示。

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{5}$$

4.2.2 线性支持向量机模型

目前基于机器学习的文本分类方法得到了广泛应用，线性支持向量机是处理高维稀疏，数据的有效机器学习方法之一。1922 年，Thorsten Joachims 首次将支持向量机(Support Vector Machines,SVM)方法用于文本分类,相比于传统算法其分类性能有显著提高[2]。SVM 是在解决小样本、非线性、高维的分类和回归问题时具有特有优势的机器学习方法,在 SVM 基础上发展的线性支持向量机(Linear SVM)已成为处理文本分类等海量高维稀疏数据的一种有效机器学习方法[3,4]。

给定一组训练样本 $\{(x_i, x_i)\}_{i=1}^{l}$，$x_i \in R^n$，$y_i \in \{1,-1\}$，Linear SVM 可以表示为求解下式的

无约束优化问题，如式（6）所示：

$$\min_{w} f(w)=\frac{1}{2}w^{T}w+C\sum_{i=1}^{l}\xi(w,x_i,y_i) \tag{6}$$

其中$\xi(w,x_i,y_i)$是损失函数，$C>0$是惩罚因子，使用的损失函数如式（7）所示：

$$\xi(w,x_i,y_i)=\begin{cases}\max(0,1-y_i w^{T}x_i), & L1-SVM\\ \max(0,1-y_i w^{T}x_i)^2, & L2-SVM\\ \log(1+\exp(-y_i w^{T}x_i)), & LR\end{cases} \tag{7}$$

4.2.3 热度指数模型

由于特定时间段内反映某一问题的留言数量能最为直观地体现该问题的热度，且留言数量越大，说明问题热度越高，两者呈正向相关关系。因此在热度指数模型中，将留言条数L_i作为一个变量单独考量，且赋予正系数1。

同时，各留言的点赞数和反对数也反映了其他市民对该问题的看法和态度。其中，市民点赞说明对该留言提出的问题产生共鸣并持相同看法，能够侧面反映问题热度；市民反对则说明并不认同该问题或持其他看法，对问题热度影响较小。为了综合考虑点赞数和反对数对问题热度的影响，将点赞数Y_i与反对数N_i归纳为投票因子V_i，并以3:2的比例赋予两者权重，再对该因子进行归一化处理，得到范围在(0,1)间的热度指标，如式（8）所示。

$$V_i=\frac{0.6\times Y_i+0.4\times N_i}{0.6\times(Y_i+N_i)} \tag{8}$$

综合留言条数L_i和投票因子V_i，可以得到本题考察热点问题的热度指数模型，如式（9）所示。

$$R_i=L_i+\frac{0.6\times Y_i+0.4\times N_i}{0.6\times(Y_i+N_i)} \tag{9}$$

4.2.4 答复意见质量的评价模型

首先定义反映答复意见质量的四个指标，如表3所示：

表 3 答复意见质量评价模型指标

指标	定义	衡量标准
相关性	答复意见与留言详情的相关程度	“留言详情”文本与”答复意见”文本的文本相似度
可解释性	答复意见的说服力程度	“答复意见”中是否引用相关书面材料进行解释
可读性	答复意见的阅读体验感程度	“答复意见”文本的有效字符数
答复效率	答复意见的时效性	留言时间与答复时间的时间间隔

对于相关性，本文将数据预处理后的所有“留言详情”文本作为语料库，以此为基础建立 TF-IDF 模型，进一步利用 gensim 库中的 similarities 计算“留言详情”和“答复意见”文本的相似度，根据文本相似度的高低反映答复意见评价模型中的相关性因子。

对于可解释性，将“答复意见”文本作为分析对象，提取能够反映引用事实根据或书面材料的特征文本，如“《”、“经查”、“经调查”、“核实”、“据查”、“核查”等，统计答复意见中出现的特征文本数量。在同一答复详情中，某特征文本的数量最多只记作 1 次，重复次数不计。使用的特征文本越多，说明答复意见的根据越有说服力，进一步反映答复意见评价模型中的可解释性因子。

对于可读性[5]，将“答复意见”文本作为分析对象，使用 len 函数统计去除停用词后各答复意见中的有效字符数。有效字符数越多，说明解释得越详细，答复意见的可读性越强，两者呈正向相关关系。

对于答复效率，将“留言时间”与“答复时间”作为分析对象。计算两者差值，得到每条留言被答复的时间间隔（天）。时间间隔越短，说明回复速度越快，答复效率越高，进一步反映答复意见评价模型中的答复效率因子。

为了消除量纲带来的影响，便于最终结果的比较，在构建模型前先对四个指标数据进行标准化处理，标准化公式如式（10）所示：

$$x_{normalization} = \frac{x - Min}{Max - Min} \tag{10}$$

根据重要程度及相关的正负性赋予四个变量相应的权重(4,3,-2,1)，得到答复意见质量的评价模型如式（11）所示：

$$Y = 4X_1 + 3X_2 - 2X_3 + X_4 \tag{11}$$

其中，Y 为答复质量评分，X_1 为可解释性，X_2 为相关性，X_3 为答复效率，X_4 为可读性。

4.3 实验设置及结果分析

4.3.1 相关参数

（1）词性权重

在问题二的短文本特征提取中，通过赋予不同词性相对应的权重达到去除停用词的目的。各词性权重如表 4 所示：

表 4 词性权重表

词性	权重
n	1.5
vn	1.2
eng	1.2
v	1.3
m	0
r	0
y	0

（2）DBSCAN 聚类分析模型

进行 DBSCAN 聚类分析时相关参数设置如表 5 所示：

表 5 DBSCAN 聚类参数表

参数	参数值
聚类半径	1
最小样本量	1

4.3.2 相关结果

4.3.2.1 问题一

（1）词云图

在 Python 中导入 wordcloud 库，wordcloud 将文本中词语出现的频率作为参数绘制词云。词语出现的频率越大，在词云中呈现的字体越大。图 2 反映了不同一级标签下的词云图。

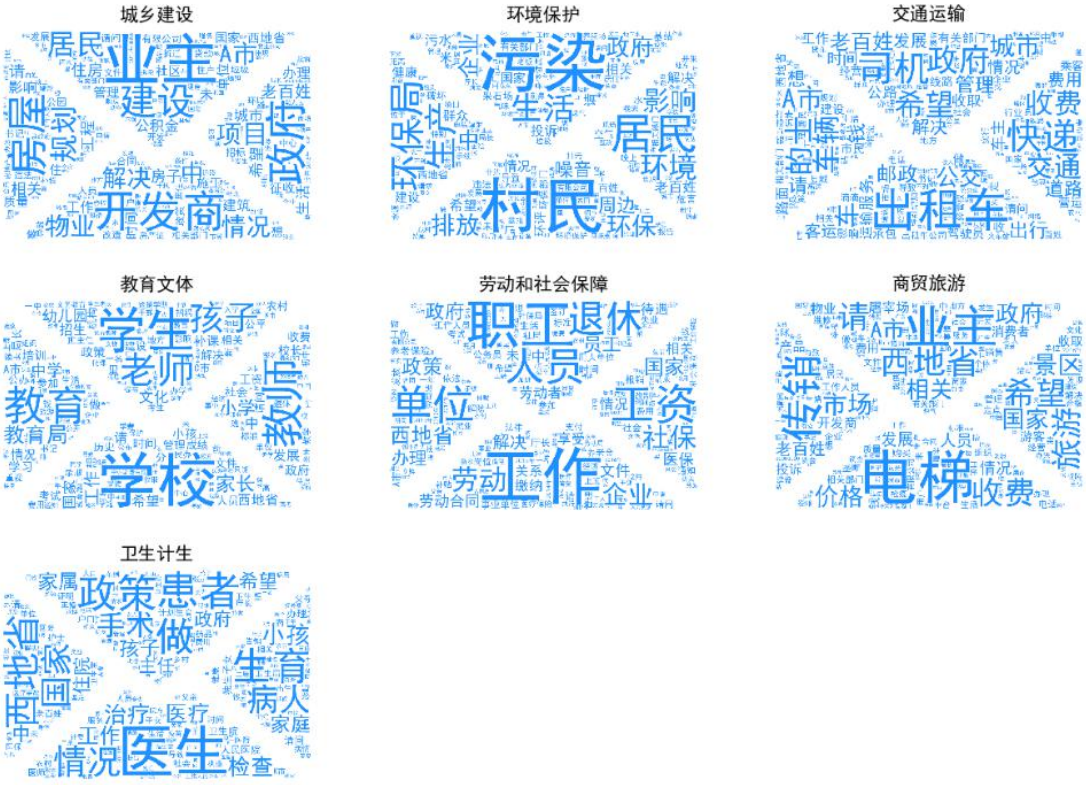

图 2 一级标签词云图

其中，城乡建设类留言中的高频词有“业主”、“开发商”、“房屋”等，环境保护类留言中的高频词有“污染”、“村民”、“环保局”等，交通运输类留言中的高频词有“出租车”、“司机”、“政府”等，教育文体类留言中的高频词有“学校”、“教育”、“老师”等，劳动和社会保障类留言中的高频词有“职工”、“工作”、“工资”等，商贸旅游留言中的高频词有“电梯”、“业主”、“传销”等，卫生计生类留言中的高频词有“医生”、“患者”、“手术”等。通过该词云图能够更为直观地把握各分类的文本特征。

（2）各模型 F-Score 值比较

本文尝试了 8 种分类模型对分词后的文本进行分类，由于样本数据中 7 种类别的留言占比存在差异，本文在拟合模型时通过设置参数 class_weight='balanced'更新权重，提高少数类的权重，以弥补数量上的差异，通过 sklearn 库中的 metrics.f1_score()函数，设置参数”average='macro',sample_weight=sw”，计算样本不完全均衡时的 F-Score，如表 6 所示：

表 6 各分类模型 F-Score 值

模型	F-Score
线性支持向量机	0.8940
逻辑回归	0.8811
支持向量机	0.8568
随机森林	0.8148
KNN	0.8140
Bagging	0.7709
决策树	0.7231
朴素贝叶斯多项式	0.6139

从 F-Score 表 6 可以看出，在实验的 8 种分类模型中，对测试数据预测类别时，线性支持向量机模型分类的 F-Score 值最高，逻辑回归模型次之，说明本题利用线性支持向量机模型对留言内容进行分类的效果最好，其混淆矩阵如图 3 所示：

$$\begin{pmatrix} 94 & 0 & 0 & 7 & 15 & 0 & 3 \\ 0 & 374 & 10 & 2 & 2 & 4 & 0 \\ 0 & 11 & 153 & 5 & 3 & 1 & 1 \\ 3 & 2 & 2 & 200 & 16 & 8 & 1 \\ 6 & 8 & 3 & 8 & 356 & 2 & 14 \\ 0 & 11 & 3 & 3 & 4 & 292 & 0 \\ 1 & 0 & 0 & 2 & 7 & 1 & 173 \end{pmatrix}$$

图 3 线性支持向量机模型混淆矩阵

在线性支持向量机模型的混淆矩阵中，主对角线上元素为该分类模型下回判正确的留言数，与同行其余错判的留言数相比，回判正确的留言占较大比例，因此也能够说明采用线性支持向量机模型进行分类的效果较好。

（3）线性支持向量机分类结果

利用线性支持向量机模型对附件 2 中的留言详情进行分类，选取前 10 条结果，如表 7 所示：

表 7 线性支持向量机模型分类结果（节选）

留言编号	留言关键词	一级标签	预测类别
1	强烈要求 西地省 加快 道路 运输 几年 经济 ...	交通运输	交通运输
2	市委 市政府 楚江 世纪 沿江 四公里 花 重金 建成...	城乡建设	城乡建设
3	M5 市 涟水 名城 55 栋 电梯 发生 下坠 颤抖...	商贸旅游	商贸旅游
4	企业 解除 劳动 关系 申请 职业病 诊断 工伤...	劳动和社会保障	劳动和社会保障
5	农村户口 购买 首套 商品房 补贴 实行...	城乡建设	城乡建设
6	借用 学校 场地 自习 收费 上课 收费 强制 补 学生 教育局...	教育文体	教育文体
7	单亲 妈妈 下岗 女职工 人民医院 做 无痛 人流 手术...	卫生计生	卫生计生
8	M9 县 旅游业 蓬勃 发展 无数 获益 县委 政府 英明 决策 管理者...	城乡建设	城乡建设
9	E4 县 朝阳 公园 几户 养 蜜蜂 散步 爬山...	城乡建设	城乡建设
10	计生委 夫妻 严格遵守 计划生育 政策 孩子...	卫生计生	卫生计生

可见，留言关键词的文本提取较为精准，与对应问题匹配，且预测类别与留言的真实标签一致，说明本题选用该模型的分类效果较好。

4.3.2.2 问题二

（1）热点问题表

根据 DBSCAN 聚类分析模型，以某一时间段内反映同一问题的留言数量作为热度评价指标，得到不同留言的类别以及数量排名前五的热点问题，如表 8 所示:

表 8 热点问题表（节选）

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	59.988	2019/11/2 至 2020/1/26	A 市 A2 区丽发新城小区	丽发新城小区旁边建搅拌站的合法问题、环境污染问题、噪音扰民问题
2	2	57.974	2019/7/7 至 2019/9/1	A 市伊景园滨河苑	伊景园滨河苑捆绑销售车位
3	3	21.833	2019/08/18 至 2019/9/4	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
4	4	12	2019/4/30 至 2019/10/16	泉星公园	泉星公园项目规划问题
5	5	10.989	2019/1/6 至 2019/7/30	A 市国家中心建设	国家中心建设存在的问题

可见，关于丽发新城搅拌站扰民留言集中在 2019 年 11 月-2020 年 1 月，说明该问题为该时段内的热点问题；伊景园滨河苑捆绑销售车位问题同样集中在两个月内，反映了该时段内捆绑销售车位问题严重；小区临街餐饮店油烟噪音扰民问题集中在半个月内，同样也为民生热点问题；而泉星公园项目规划问题和国家中心建设问题留言存在的时间跨度虽然较长，但反映该问题的留言数量较多，因此也位于热点问题前五的排行中。

（2）热点问题留言明细表

基于 DBSCAN 聚类分析模型的文本挖掘结果，按热度排名顺序分别选取两条留言展示在表 9 中：

表 9 热点问题留言明细表（节选）

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数

1	188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019/11/19 18:07	A市万家丽南路丽发新城居民区，开发商在小区旁50米处建搅拌站...	1	0
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019/11/13 11:20	我是A2区丽发新城小区的一名业主，我要投诉同发投资有限公司在未经小区业主同意的情况下，在离小区不到百米的地方建搅拌站...	0	0
…	…	…	…	…	…	…	…
2	251844	A909167	投诉伊景园滨河苑项目违法捆绑车位销售	2019/8/20 13:34	投诉广铁集团强制要求职工捆绑购买12万的车位费...	1	0
2	190337	A0009051 9	关于伊景园滨河苑捆绑销售车位的维权投诉	2019/8/23 12:22	投诉伊景园.滨河苑开发商捆绑销售车位....	0	0
…	…	…	…	…	…	…	…
3	189381	A0001098 15	A市万科魅力之城商铺无排烟管道，小区内到处油烟味	2019/12/4 16:25	A市万科魅力之城自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区大堂门口...	0	0
3	195095	A0003908 9	魅力之城小区临街门面油烟直排扰民	2019/9/5 12:29	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟烧烤熏死人...	3	0
…	…	…	…	…	…	…	…
4	193337	A0008034 3	请问A7县泉星公园何时开工，工期多长？	2019/10/16 13:03	泉星公园项目已经筹备了5年之久，人民群众非常期待，请问何时开工...	12	0
4	226408	A0008034 2	A市经开区泉星公园项目规划需优化	2019/8/9 16:47	目前A市经济技术开发集团有限公司的泉星公园项目的规划拟批准内容进行了公示...	4	0
…	…	…	…	…	…	…	…
5	193678	A0001040 70	反映A市城市交通存在的诸多问题	2019/3/24 20:40	A市地铁2号线出入口很多未设置下行扶梯，带行李旅客极为不便...	8	0

5	201854	A0004235 1	对 A8 县西部修建高速公路的看法和建议	2019/1/6 14:14	西地省 A8 县西部为什么迟迟没有修快速路高速公路的打算...	0	0
…	…	…	…	…	…	…	…

4.3.2.3 问题三

（1）原始变量值

计算各变量取值，按附件 4 中的留言顺序选取前五条展示，如表 10 所示：

表 10 答复意见原始变量值（节选）

留言编号	留言详情	答复意见	X1	X2	X3	X4
A00077538	2019 年 4 月以来，位于 A 市 A2 区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查核...	3	0.237994	15	109
A00077538	潇楚南路从 2018 年开始修，到现在都快一年了...	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...	1	0.021570	14	85
A00077538	地处省会 A 市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	1	0.334591	14	105
A00077538	尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	1	0.278450	14	78
A00077538	建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	0	0.774881	15	34

其中， X_1为可解释性， X_2为相关性， X_3为答复效率， X_4为可读性。

（2）标准化的答复意见评价模型

计算各留言详情与对应答复意见的文本相似度，按附件 4 中的留言顺序选取前五条展示，如表 11 所示：

表 11 标准化的答复意见评分（节选）

留言编号	X1	X2	X3	X4	Y
A00077538	0.6	0.253202957	0.012931034	0.041381929	71.72569555
A00077538	0.2	0.022948045	0.012068966	0.032270311	39.86232372
A00077538	0.2	0.355972537	0.012068966	0.039863326	53.8195318
A00077538	0.2	0.296243524	0.012068966	0.029612756	51.19302132
A00077538	0	0.824398749	0.012931034	0.012908125	61.81395343

可见，标准化后的变量值均位于(0,1)范围内，答复质量评分 Y 位于(0,100)范围内。

（3）答复意见质量排名

对（2）中的答复质量评分 Y 进行降序排列，按答复质量从高到低的顺序选取前五条展示，如表 12 所示：

表 12 答复意见质量排名表（节选）

留言编号	留言详情	答复意见	答复质量评分 Y
486	尊敬的易书记您好!我们是新河街道协管队伍里的一员，工作也快 2 年多了...	网友“UU0081971 您好！您的留言已收悉。现将有关情况回复如下：一、关于您所反映的...	100
427	请求政府及时接收“恒大城”小区配套幼儿园并尽快开办公立园报告...	网友“UU008144” 您好！您的留言已收悉。经 A5 区教育局调查了解，现将有关情况回复如下...	96.34973401
226	本人邱林松，家住 A3 区望岳街道金星村横岭塘组。2011 年本人与龙平和合伙	网友“UU008898”您好！您的留言已收悉。现将有关情况回复如下：据查...	93.90772781

	承包...		
430	请求政府及时接收“恒大城”小区配套幼儿园并尽快开办公立园报告...	网友“UU0081831”您好！您的留言已收悉。经A5区教育局调查了解，现将有关情况回复如下...	92.94158393
961	请求A8县市政府刹住殡仪服务乱收费的歪风尊敬的A8县市委周晖书记...	网友“UU0081656”您好！您的留言已收悉，我们立即进行了调查核实，现将有关情况答复如下...	90.2875769

可见，排名前五的答复评分均高于90，分布合理，说明答复意见评价模型有效。

五、模型优化

5.1 字典自定义

在分词的过程中，没有对字典进行重新定义，而直接使用jieba库里的posseg词性标注分词器产生字典，并进行分词、标注词性，这可能导致部分词语被拆分，最终无法准确提取特征文本，分类效果欠佳。

5.2 完善词性矩阵

本文使用的词性矩阵为自定义下的词性矩阵，并无区分良性词性和劣性词性。若能尽可能详细地列举良性词性和劣性词性，并赋予其合理的权重，将会使词性矩阵更加完善，分类效果更好。

5.3 增加地点分词

在聚类文本中，地点类型的词语多种多样。若在算法中加入地点分词，可以达到优化聚类效果的目的。

5.4 因子权重定义

答复意见质量的评价模型中，四个指标的权重为人为根据实际情况分配所得，若先能先对部分随机抽取的答复意见进行评分，将得到的分数与指标数据进行回归分析得到各指标的回归系数，则评价结果会更加准确。

六、参考文献

[1] 关满祺.一种快速的短文本相似度检测方式[J].通讯世界,2020,27(01):29-30.

[2] 杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术与信息化,2020(03):146-148.

[3] V.N.Vapnik.The Nature of Statistical Learning Theory[M].NY:Spring-Verlag,1995.

[4] 解洪胜.Linear SVM 在大数据分类中的应用[J].信息技术与信息化,2017(09):81-83.

[5] 郭卫丽. 文本评论数据质量分析方法研究[D].重庆大学,2016.