C 题：“智慧政务”中的文本挖掘应用

摘要

本文旨在设计“智慧政务”中的自然语言处理技术的应用，根据附件提供的群众留言记录及相关部门的答复意见等数据，实现了对群众留言内容的一级标签分类，并完成了对群众留言热点问题的挖掘和对相关部门答复意见的质量评价。对提升政府的管理水平和施政效率具有重大的意义。

针对问题一，利用深度学习**ERNIE模型**算法，构建了关于留言内容的一级标签分类模型。将预处理后得到的训练样本进行模型的训练，然后将训练好的模型对测试集数据进行测试，得到关于一级标签的分类结果，计算得到混淆矩阵，以及模型的Acc值（准确率）为98.37%，通过F-Score评价模型对七类的F1结果进行总评价，得到**F-Score为0.9820**。

针对问题二，定义了三个指标：簇内留言条数、簇内留言总点赞数和簇内总反对数，建立关于群众留言问题的热度评价模型：**热度指数=**w_1***簇内留言条数+**w_2***簇内留言总点赞数+**w_3***簇内总反对数**。首先通过**TF-IDF算法**对预处理后的数据进行特征提取，其次利用**DBSCAN聚类算法**进行聚类，得到留言聚类结果，最后统计每一簇内的留言数据总量、点赞数和反对数，得到了热度排名，并给去排名前5热点问题，其中**小区旁修建搅拌站扰民**排名第一。

针对问题三，从答复意见的相关性、完整性和可解释性三个角度，建立了基于**多特征融合**的答复意见质量的评价模型：**质量评价指标=**w_1***相关性+**w_2***完整性+**w_3***可解释性**。首先通过构建每条答复意见的**字向量**文本，其次通过**编辑距离**求文本相似度，得到相关性指标，然后定义了完整性和可解释性两个特征，最后基于多特征融合的评价模型即可得到答复意见质量的综合评价指数。

关键词：ERNIE；TF-IDF；DBSCAN聚类；编辑距离；多特征融合

Abstract

Issus in the "wisdom" of the purpose of this paper is to design the application of natural language processing technology, according to the attachment for the masses to the message data such as record and relevant departments to reply opinions, implements the message content classification of level 1 label to the crowds, and completed a message to the crowds hot problems of mining and of quality evaluation for the opinions of relevant departments to reply.It is of great significance to improve the management level and efficiency of government.

Aiming at problem 1, a level 1 label classification model for message content is constructed by using deep learning ERNIE model algorithm.After preprocessing the training model of training samples, and then the trained model test set of test data, on the classification of the level 1 label as a result, the confusion matrix is calculated, and the model of the Acc value (accuracy) of 98.37%, by F - Score evaluation model to evaluate total seven classes of F1 results, get the F - Score of 0.9820.

For question 2, three indexes were defined: the number of messages left in the cluster, the total number of messages left in the cluster, and the total number of objectionable messages left in the cluster. A heat evaluation model for crowd message was established: heat index = * the number of messages left in the cluster + * the total number of thumb up messages left in the cluster + * the total number of objectionable messages left in the cluster.First through the TF - IDF to feature extraction algorithm after preprocessing of the data, then use DBSCAN clustering algorithm for clustering, message clustering results are obtained, and the final amount of message data statistics within each cluster, count and opposed thumb up and get the heat number, and to go to the top 5 hot issues, intrusive or community built by mixing station ranked first among them.

Aiming at question 3, an evaluation model of response quality based on multi-feature fusion is established from the three perspectives of relevance, integrity and interpretability of response opinions: comprehensive evaluation index = * relevance + * integrity + * interpretability.Firstly, the character vector text of each reply is constructed; secondly, the text similarity is calculated by editing distance to obtain the correlation index; then, the two characteristics of completeness and interpretability are defined; finally, the comprehensive evaluation index of response quality can be obtained based on the evaluation model of multi-feature fusion.

Key word：ERNIE；TF-IDF；DBSCAN clustering；Edit distance；Multi-feature fusion

目录

1 问题重述

1.1 问题背景

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。同时，随着大数据、云计算、人工智能等技术的发展，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，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 要解决的问题

1、针对附件 2 数据，建立关于留言内容的一级标签分类模型。由划分的数据集，对测试集留言内容的所属一级标签进行预测，且根据 F-score 评价指标对分类方法进行评价。

2、针对附件 3 数据，定义合理出合理的热度评价指标，并给出评价结果，并且按照题目中表 1、表 2 的格式给出排名前 5 的热点问题及相应热点问题对应的留言信息，分别保存为文件“热点问题表.xls”、“热点问题留言明细表.xls”。

3、针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性和可解释性的角度，建立一套合理的评价方案对答复意见的质量进行评定。

2 问题一：群众留言分类

2.1 问题分析

题目要求根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。根据附件 2 给出的数据，其主要内容涵盖在网络问政平台下，群众的具体留言情况和内容以及对应一级标签，一级标签为离散型的 7 个类别，可通过留言主题和详情与对应类别，建立并学习有效的文本分类模型，对留言情况进行分类。

2.2 模型建立

2.2.1 ERNIE 模型

中文文本分类在自然语言处理中取得了最先进的成果。侧重于对模型的训练，通过几个简单的任务来掌握单词或句子的共现关系。训练语料库中还存在其他有价值的词汇、句法和语义信息，如命名实体、语义关系和相似度等。

为了更好的对附件二的留言数据分类，根据参考文献[1]和资料[2]，从训练语料库中提取词汇、句法和语义信息，采用连续学习的方法，充分利用训练数据当中词法结构，语法结构，以及语义信息去学习建模。从端到端的解决问题。通过 ERNIE 模型得到的文本特征，通过 TextClassifie 全连接层进行分类。**建立 ERNIE 模型**如图 1 所示：

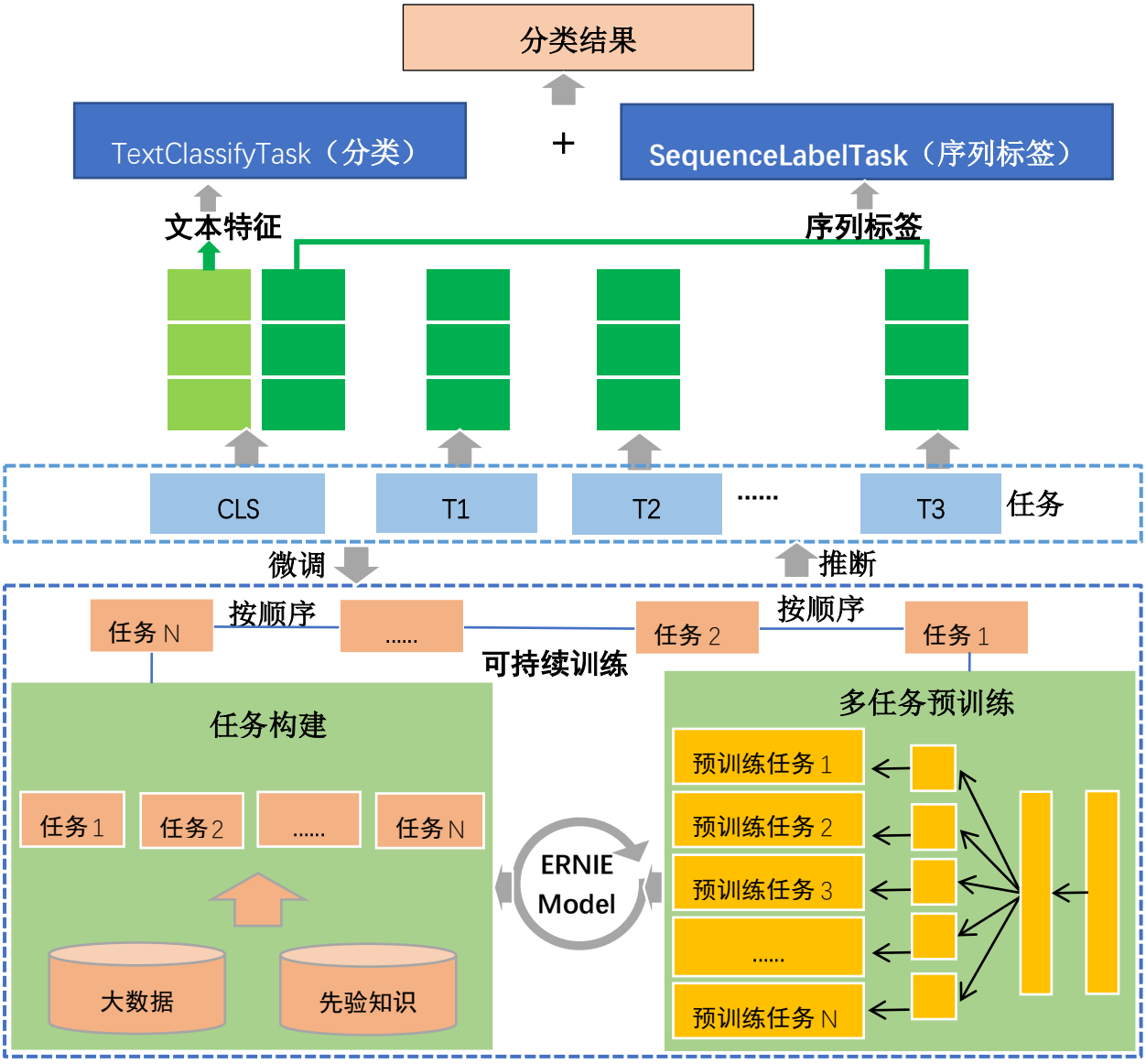

图 1 ERNIE 模型

2.2.2 评价模型

由资料[3]可知，F1-Score 是统计学中用来衡量二分类模型精确度的一种指标，同时兼顾了分类模型的精确率和召回率，可以看作是模型的精准率和召回率的一种加权平均。本题有七大类别，为计算建立的 ERNIE 模型的评价指标，综合每个类别的 F1-Score 求其均值，建立 F-Score 评价模型如下：

$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \qquad (1)$$

其中 $n=7$， $i=1,2,\cdots,7$ P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

2.3 模型框架

2.3.1 程序框架

ERNIE 模型运用程序的主要文件说明如下：

表 1 程序功能说明

run.py	程序的入口，实现整体框架及模型调用
ERNIE.py	数据入口，实现模型参数配置及模型定义
utils.py	对文本数据预处理及时间记录函数，以及 bug 修正
train_eval.py	模型的学习过程
pytorch_pretrained	模型的预训练代码（详细）

2.3.2 模型参数

ERNIE.py 文件的主要参数如下：

表 2 训练参数

参数	值
learning_rate：学习率	5e-5
pad_size：每句话处理成的长度(短填长切)	200
batch_size：mini-batch 大小	10
num_epochs：epoch 数	3
num_classes：类别数	数据所需分类数
require_improvement：若超过 1000batch 效果还没提升，则提前结束训练	1000

2.4 模型求解

2.4.1 数据预处理

1、数据简单分析

2、数据预处理：

（1）读取数据，数据合并。

（2）将合并的内容依次进行缺失值处理、去重、脱敏处理。

（3）将一级标签转化为连续的数值类别，并将一级标签保存到模型目录下的 THUCNews\data 中，命名为“class1.txt”

（4）将处理好的留言内容与对应的数值类别用制表符间隔合并。

（5）将合并后的数据以 8：1：1 切分，并保存到模型目录下的 THUCNews\data 中，依次命名为“train1.txt”、“dev1.txt”、“test1.txt”。

2.4.2 预训练连续学习

构建每条文本的词法级别，语法级别，语义级别的预训练任务。所有的这些任务，都是基于无标注或者弱标注的数据。如图 2 所示

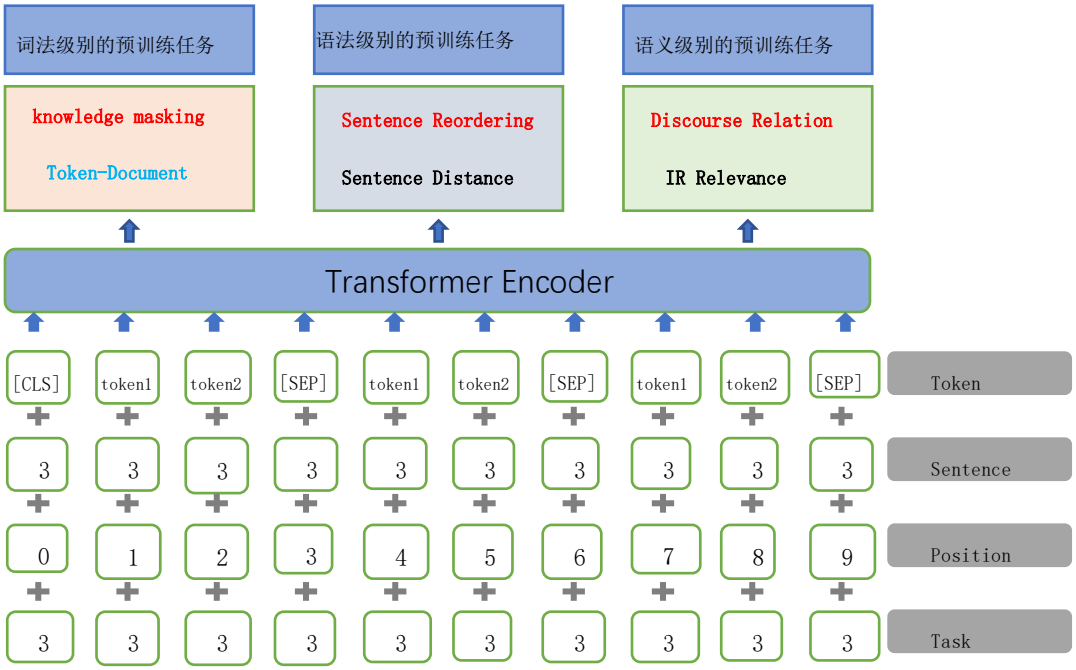

图 2 预训练

1、构建词法级别的预训练任务，来获取训练数据中的词法信息：knowledge masking task，捕捉输入样本局部和全局的语义信息；Capitalization Prediction Task，判断一个词是否大写； Token-Document Relation Prediction Task，通过识别这个文中关键的的词，增强模型去获取文章的关键词语的能力。

2、构建语法级别的预训练任务，来获取训练数据中的语法信息：（1）Sentence Reordering Task，在训练当中，将 paragraph 随机分成 1 到 m 段，将所有的组合随机 shuffle. 我们让 pre-trained 的模型来识别所有的这些 segments 正确的顺序. 这便是一个 k 分类任务；（2）Sentence Distance Task，构建一个三分类任务来判别句子的距离，0 表示两个句子是同一个文章中相邻的句子，1 表示两个句子是在同一个文章，但是不相邻，2 表示两个句子是不同的文章。通过构建这样一个三分类任务去判断句对（sentence pairs）

位置关系（包含邻近句子、文档内非邻近句子、非同文档内句子 3 种类别），更好的建模语义相关性。

3、构建语义级别的预训练任务，来获取训练数据中的语义任务：

（1）Discourse Relation Task，通过判断句对间的修辞关系，更好的学习句间语义。

（2）IR Relevance Task，通过类似 google-distance 的关系来衡量 两个句子之间的语义相关性，更好的建模句对相关性。

2.4.3 学习过程

1、对预训练的结果进行学习，大致分为两类，token_evel 就是词汇层次的任务，一类是 sentence_level 就是句子层次的任务，可以理解为填空题，判断这个词到底是哪一类、哪个词，是关键词还是非关键词，另一部分呢就是句子输出 cls 是 a 类还是 b 类，如图 3 所示。

2、tranfromer,总共有 12 层，每一层又包括自注意力机制、归一化、又包括求和与归一化、以及一些前馈神经网络的一些机制，最终构成了一个相对复杂的神经网络结构，这里面最核心的是自注意力机制，能学会每个词汇和词之间的关系。如图 4 所示。

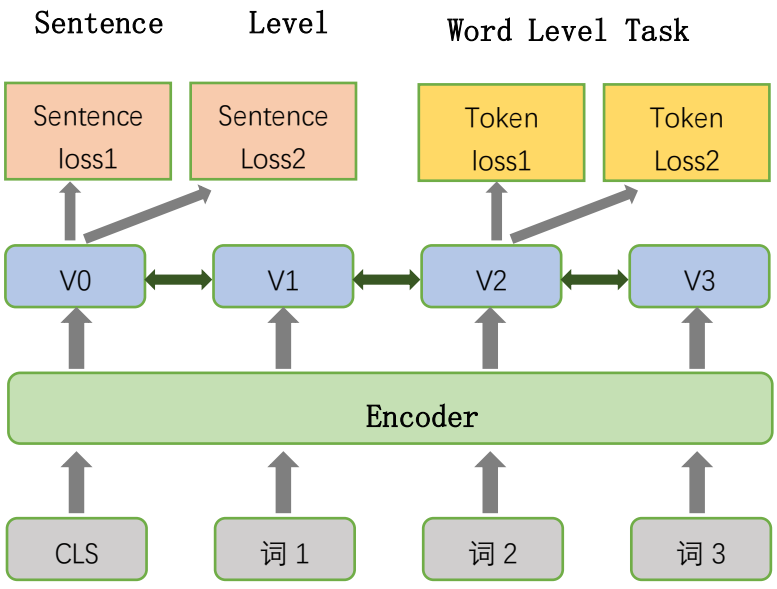

图 3 网络输出层

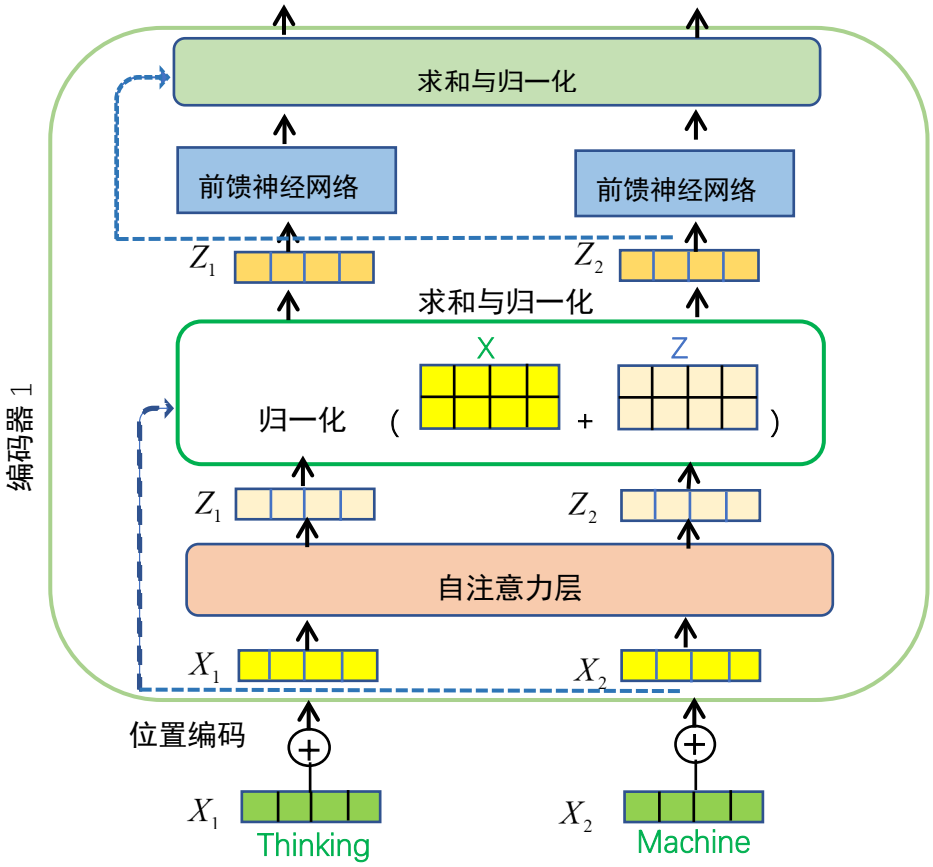

图 4 transformer 编码器

2.4.4 分类过程

1、通过预训练和学习过程得出的数据之后，获输入和输出的变量，从输出变量中找到用于分类的文本特征，将文本特征输入 TextClassifier 全连接层进行分类，如图 5 所示。

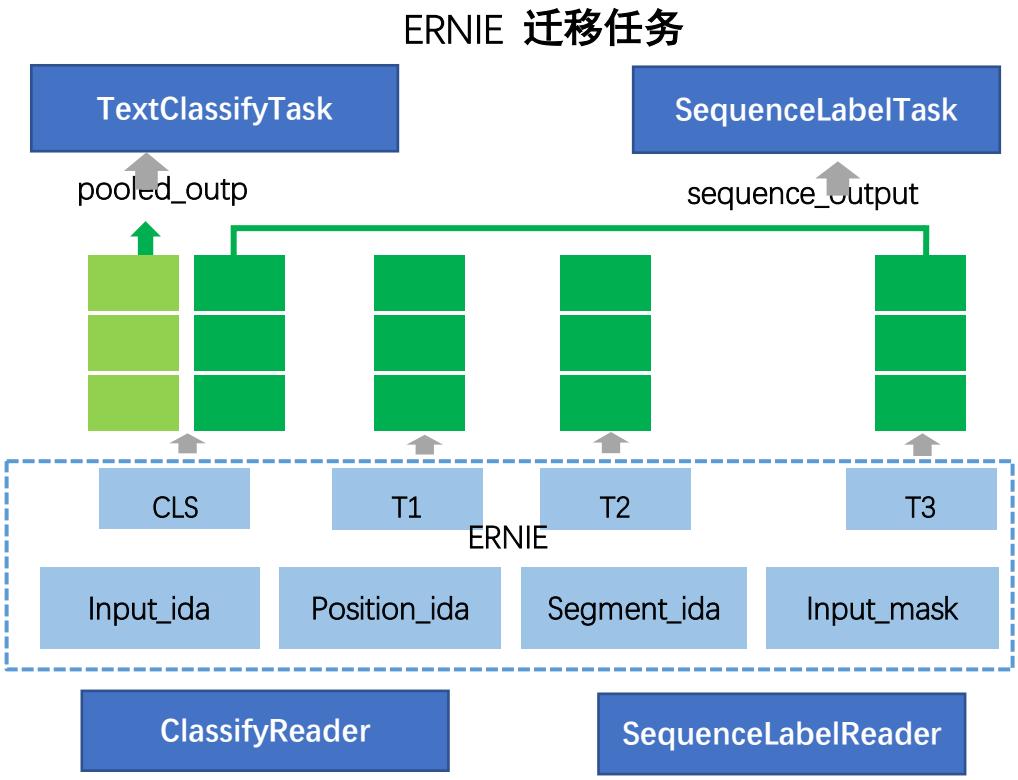

图 5 迁移任务

2.4.5 F1 值计算

1、引入概念

TP（True Positive）：把正的判断为正的数目，判断正确。

FP（False Positive）：把负的判断为正的数目，判断错误。

TN（True Negative）：把负的判断为负的数目，判断正确。

FN（False Positive）：把正的判断为负的数目，判断错误。

将上述结果用混淆矩阵表示：

表 3 混淆矩阵

	样本为正类	样本为负类
预测为正类	TP	FP
预测为负类	FN	TN

（1）准确率（accrracy_score）：所有预测正确的（正类负类）占总的比重。

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

（2）精准率（precision_score）：查准率，即正确预测为正的占全部预测为正的比例。

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

（3）召回率（recall_score）：查全率，即正确预测为正的占全部实际为正的比例。

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

（4）F 得分函数（F-score）：基于查准率与查全率的的调和平均定义，如下式：

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

其中本文 $n = 7$，$i = 1,2,\cdots,7$，P_i 为第 i 类的精准率，R_i 为第 i 类的召回率。

2.5 结果与分析

2.5.1 模型的训练

1、训练过程展示如下图：

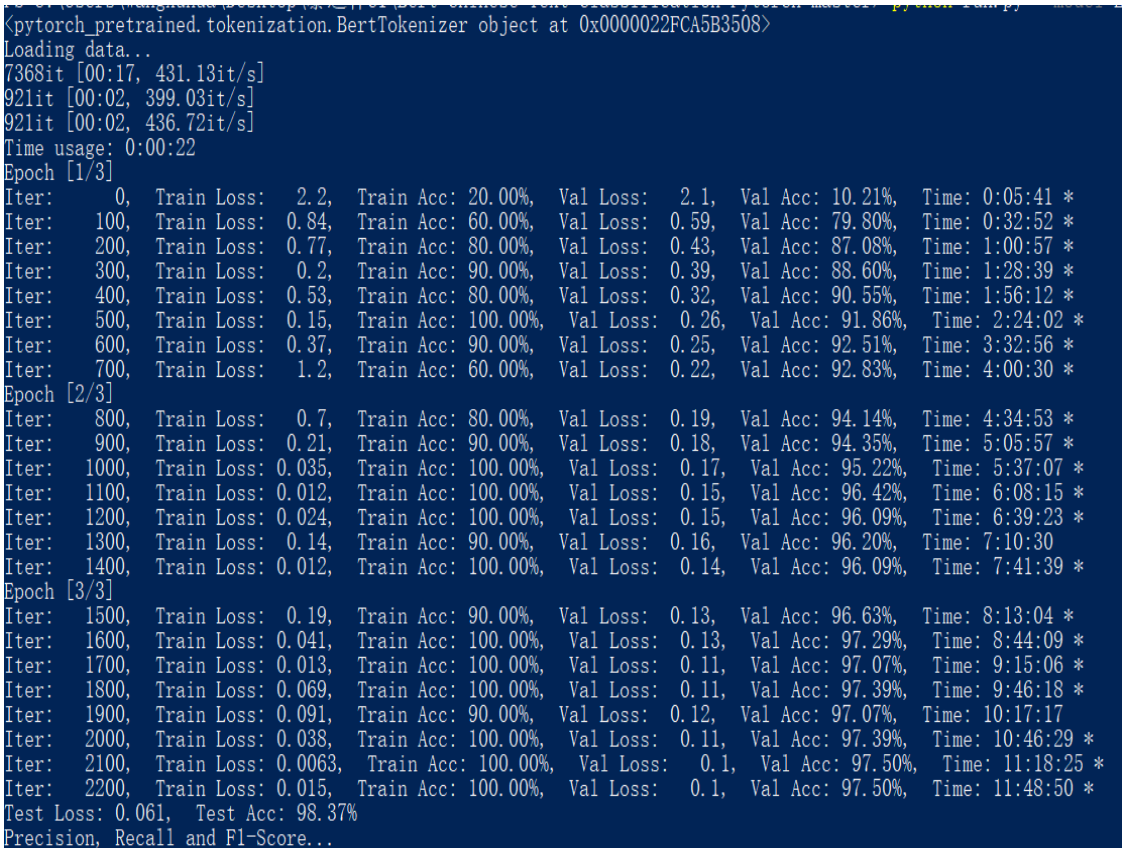

图 6 完整数据训练过程

Epoch 表示训练次数；Iter 表示使用一个部分数据对模型进行一次参数更新的过程，每次训练更新七次。Train Loss 代表训练集损失值，Train Acc 代表训练集准确率，Val Loss 代表测试集损失值 Val Acc 代表测试集准确率。

2.5.2 测试结果

1、对附件二处理后的数据进行测试，分别得到各个类别的精准率(precision)、召回率(recall)和 F1 得分(F1-score)，测试结果见表 4：

表 4 模型的测试结果

	Precision	Recall	F1-score	Support

A1	0.9854	0.9758	0.9806	207
A2	0.9902	0.9712	0.9806	104
A2	0.9630	1.0000	0.9811	52
A4	0.9941	1.0000	0.9971	169
A5	0.9848	0.9848	0.9848	169
A6	0.9744	0.9913	0.9828	115
A7	0.9733	0.9605	0.9669	76
accuracy			0.9837	921
Macro avg	0.9807	0.9834	0.9820	921
Weighted avg	0.9838	0.9837	0.9837	921

得到测试结果**混淆矩阵**见式（6）：

$$\begin{bmatrix} 202 & 1 & 1 & 1 & 1 & 1 & 0 \\ 2 & 101 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 52 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 169 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 195 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 114 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 & 73 \end{bmatrix} \quad (6)$$

2、ERNIE 模型分类最终 F 得分计算结果如下：

$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} = 0.9820 \quad (7)$$

2.5.3 模型评价

1、根据表 4，建立的一级标签分类模型对结果的预测效果很好，ERNIR 模型分类效果的评价指标 F 得分高达 0.9820，准确率高达98.37%，得到的模型性能的评估指标，精准率、召回率和 F1 得分均达到 95%以上，该模型对七个类别的预测都达到了很好的效果，F1 得分几乎都达到了 98%以上。

2、由式（6）混淆矩阵可以看出，921 条测试集数据，906 条数据能够被正确分类，15 条数据被错误分类。

3、综上，于是认为 ERNIE 模型用于处理分类问题的性能高，对该数据的预测效果很好。

2.6 模型比对

1、近年来，复杂文档和文本的数量呈指数级增长，需要对机器学习和深度学习的方法有深刻的理解，才能在许多应用中准确地分类。许多深度学习的方法取得了卓越的成绩。本文将深度学习理论应用于文本分类中，提高文本分类算法的精确度和效率。本题通过用不同的深度模型进行计算，得出最佳的文本分类效果为 ERNIE 模型。针对本题附件二数据采用多个模型分类效果如下：

表 5 模型比对

模型	Acc	备注	程序执行过程
Transformer	0.7362	效果较差	python run.py --model Transformer
TextCNN	0.9642	Kim2014 经典的 CNN 文本分类	python run.py --model TextCNN
TextRNN	0.9359	BiLSTM	python run.py --model TextRNN
FastText	0.9598	bow+bigram+trigram， 效果很好	python run.py --model FastText
Bert	0.9653	单纯的 bert	python run.py --model Bert
Bert_RNN	0.8675	bert + RNN	python run.py --model Bert_RNN
ERNIE	**0.9837**	**官方解释处理中文文本效果优于 bert，本题数据实测分类效果最佳**	python run.py --model ERNIE

3 问题二 ：热点排行

3.1 问题分析

本题根据附件三将某一段时间内反应特定地点或者特定人群问题的留言进行归类，定义合理的热度评价指标，给出评价结果。其中数据预处理采用 NLP 传统的方法，对数据进行脱敏、去重、结巴分词、去停用词等处理；采用人工标注手段对留言主题进行特定的人名、地名、特殊词汇进行标注；采用 TF-IDF 方法对每一个留言主题进行特征提取并输出词频矩阵；用余弦计算两两之间的相似度，采用 DBSCAN 密度聚类方法，将同类数据进行归类，最后进行热点指标排行。

3.3 求解流程图

根据对问题的分析，确定问题的解决思路和具体求解过程，其中对问题二的求解总流程图如图 7 所示：

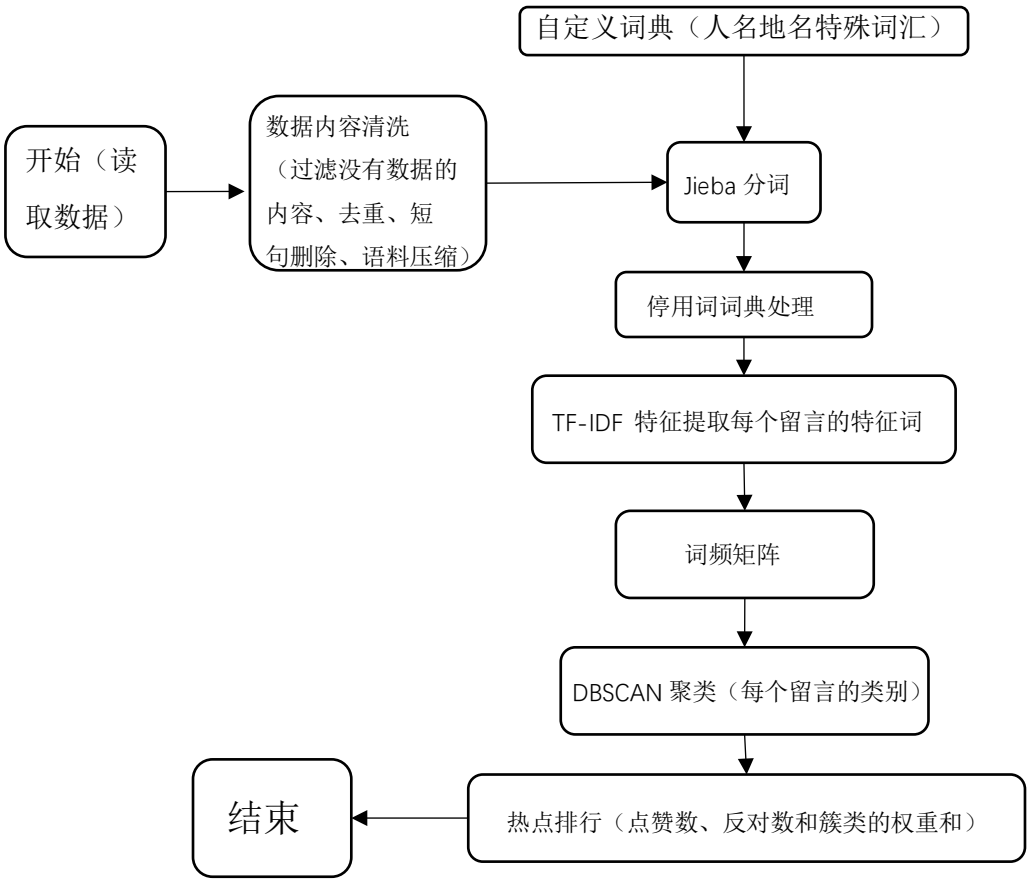

图 7 问题求解流程图

3.4 数据预处理

由于题目所给数据集存在大量的群众留言信息，其中包括一些重复的留言，而且留言内容也可能存在一些对所反映问题没有作用的字词，所以需要对其进行数据过滤，再进行分词，提取出关键词进行特征提取。留言内容数据预处理流程图如图 8 所示：

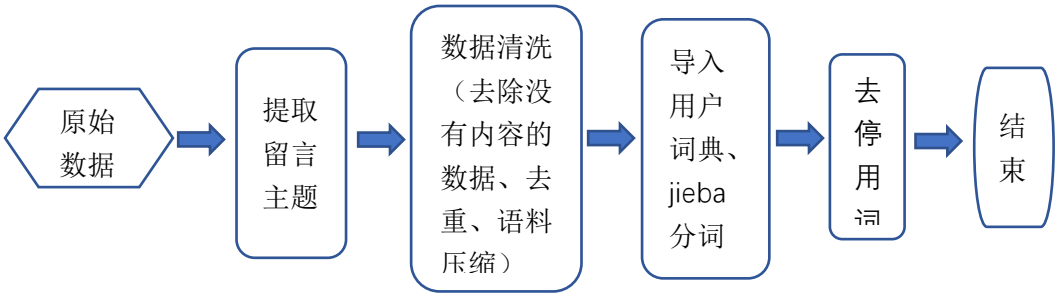

图 8 留言主题数据预处理流程图

Step1：读取数据，提取主题信息。由于数据过于冗长，这里提取主题内容，对它进行进一步处理和热点问题挖掘。

Step2：数据清洗。首先剔除掉没有内容的留言主题，其次对重复的主题内容进行过滤，最后对留言主题信息进行语料压缩。

Step3：对于清理后的留言信息，基于 jieba 分词的主题内容分词。在分词的时候考虑使用分词工具进行分词效果会有较大的误差，于是这里通过观察留言主题内容，自定义一个用户词典，将词典原来无法识别的特殊的人名、地名、机构名等特殊词汇加入到用户词典，这样，在分词前让分词工具先对用户词词典进行分析，最后分词的时候便能够达到我们想要的分词效果。

Step4：载入停用词典，去停用词。由于文本内容存在大量的语气词、标点符号等对文本信息量的提取无贡献的词，需要载入停用词典，将其去除，只保留想要的部分有用的文本信息量。

3.5 模型求解

3.5.1 TF-IDF 特征提取

基于 TF-IDF 的留言主题内容特征提取。经数据预处理后，用剩下的这些词来做特征提取，可以更好的反映留言的特征。本文使用 TF-IDF 方法来提取特征，通过 TF-IDF 特征提取之后，每一条留言数据都有一个向量标识，向量的每一个值都是一个词的 TF-IDF 值。该向量的获取方式为首先统计出所有的词，把

每个词当成向量的一个维度，如果该条留言数据有这个词，就在这个词的维度上计算它的 TF-IDF 值；如果不存在这个词，那么在这个词的维度上的值就为 0。用这种算法对所有的留言数据进行特征提取，提取的结果就是一个稀疏矩阵。

TF-IDF 原理

TF-IDF 即词频-逆文档频率，用来表示一个特定的词在数据集中出现的次数和它在文档中的重要性。根据资料[7]可知，主要有两部分构成：

（1）TF（词频）：$TF(\omega)=\frac{\text{词条}\omega\text{在文档中出现的次数}}{\text{该文档的总词条数}}$ （9）

（2）IDF（逆文档频率）：$IDF(\omega)=\lg(\frac{\text{文档总数}}{\text{包含词条}\omega\text{的文档数}})$ （10）

计算 TF-IDF，即将上述两式相乘：$TF-IDF=TF\times IDF$ （11）

3.5.2 基于 DBSCAN 文本聚类

1、BDSCAN 原理

DBSCAN 算法是一种通过数据对象密度进行查找相似属性的聚类算法。该算法不需要提取确定聚类簇的数量，不仅能够对任意数据进行聚类，还能识别数据中的噪声点。由参考文献[5]可知，DBSCAN 算法的定义如下：

（1） ε（Eps）邻域：给定对象半径 ε 内的区域称为该对象的 ε 邻域。

（2） $MinPts$ 算法中设定的参数值，表示 ε-邻域中的数据对象的最小值。

（3）核心对象：如果给定对象 ε 邻域内的样本点数大于等于 $MinPts$，则称该对象为核心对象。

（4）直接密度可达：给定一个对象集合 D，如果 p 在 q 的 ε 邻域内，且 q 是一个核心对象，则我们说对象 p 从对象 q 出发是直接密度可达的。

（5）密度可达：对于样本集合 D，如果存在一个对象链 $P_1,P_2,\cdots,P_n$，$p=P_1,q=P_n$，对于 $P_i\in D(1\le i\le n)$，P_{i+1} 是从 P_i 关于 ε 和 *MinPts* 直接密度可达，则对象 p 是从对象 q 关于 ε 和 *MinPts* 密度可达的。

（6）密度相连：如果存在对象 $o\in D$，使对象 p 和 q 都是从 o 关于 ε 和 *MinPts* 密度可达的，那么对象 p 到 q 是关于 ε 和 *MinPts* 密度相连的。

（7）噪音点：数据集 D 中既不是核心点也不是边界点的点则为噪音点。

（8）簇。所有密度相连的点组成的集合。

2、DBSCAN 流程

通过以上定义可知，DBSCAN 算法的核心在于参数 *Eps* 和 *MinPts*，通过这两个参数确定每个点的领域和核心对象，继而通过核心对象寻找密度可达点，从而实现数据对象的聚类。根据参考文献[6]可知，DBSCAN 算法流程如下：

Step1：输入的数据集 $D=\{x_1,x_2,\cdots,x_n\}$，ε 为半径参数，*MinPts* 为最小对象参数，将数据集 D 中的所有对象标记为未读。

Step2：从数据集 D 中取包含任意个数据对象 p 的数据集 D_i，其中 $D_i\in D,i=1,2,\cdots$，并将 D_i 标记为已读。

Step3：通过 ε 和 *MinPts* 参数对 p 进行判断，如果 p 为核心对象，找出 p 的所有密度可达数据对象，并标记为已读，若 p 不是核心对象，且没有哪个对象对 p 密度可达，将 p 标记为噪声数据。

Step4：在满足 $D_i\cap D_{i+1}\in\varnothing$ 的条件下重复 Step2 和 Step3，直到所有数据都标记为已读。

Step5：将其中一个核心对象作为种子，将该对象的所有密度可达点都归为一类，形成一个较大范围的数据对象集合，也称为聚类簇。

Step6：不断循环 Step5 直至所有核心对象都遍历完，剩下没有归为一类的数据便为噪声点。

DBSCAN 算法流程图如下所示：

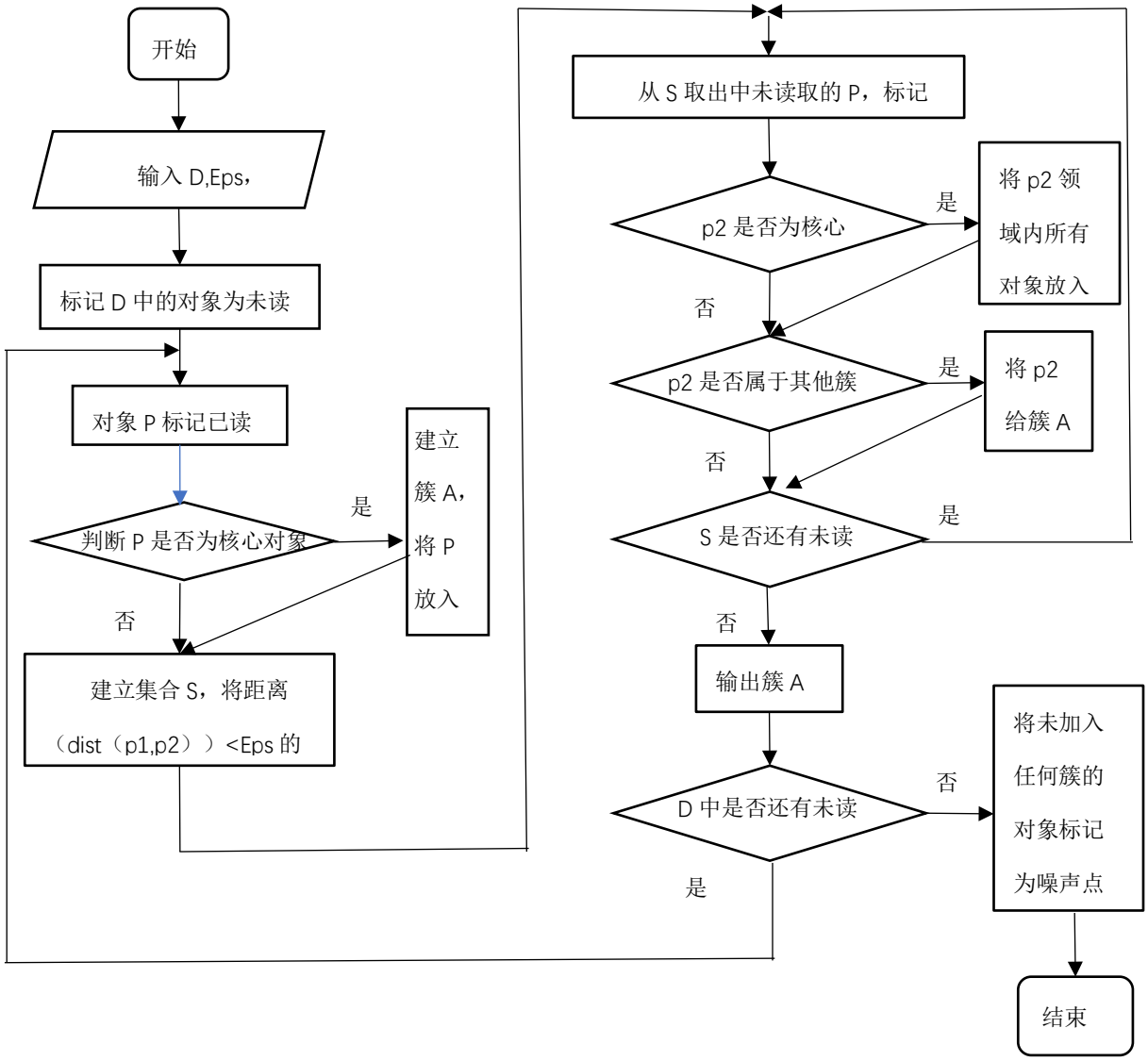

图 9 DBSCAN 算法流程图

通过 DBSCAN 聚类算法，对附件 4 的数据进行聚类，得到每一类的簇。

3.5.3 热点排行

将留言问题进行归类，根据热度评价指标，对留言问题热点排行的过程如图 10 所示：

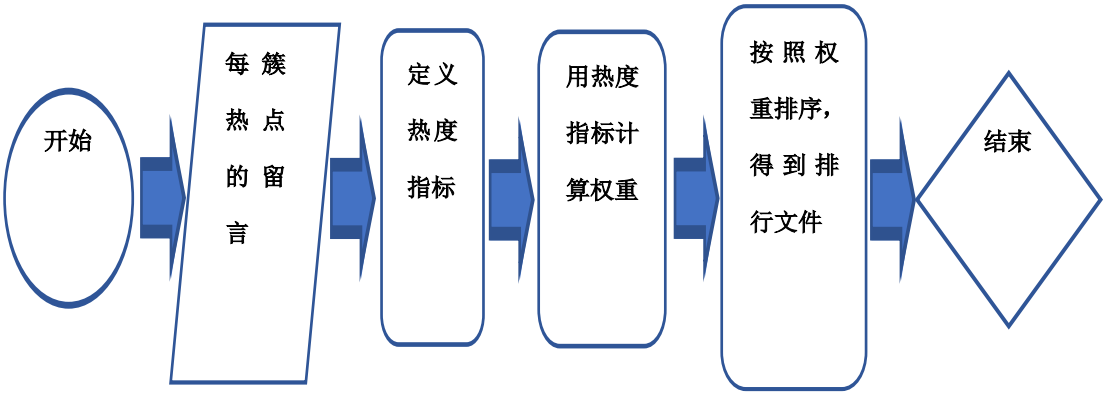

图 10 热点问题排行计算流程

（2） 按照题目表 1 的格式给出排名前 5 的热点问题，见表 7：

表 7 热点问题表

热度排名	热度 ID	热度指数	时间范围	地点/人群	问题描述
1	1	61	2019/11/02 至 2020/01/06	A 市 A2 区丽发新城小区	小区旁修建搅拌站扰民
2	2	40	2019/07/07 至 2019/09/01	A 市伊景园滨河苑	车位捆绑销售问题
3	3	38	2019/08/09 至 2019/08/26	A 市经开区泉星公园	对公园项目规划的建议
4	4	34	2019/01/08 至 2019/07/30	A 市	关于推动国家中心城市建设的意见
5	5	20	2017/06/08 至 2019/11/05	A 市经济学院学生	学校强制学生去定点企业实习

结果分析：

其中前五的热点问题分别为 A 市丽发小区旁修建搅拌站扰民现象严重、A 市伊景园滨河苑捆绑销售车位的问题、A 市经开区泉星公园项目规划优化建议的问题、推动 A 市加快国家中心城市建设的意见、以及关于讨论 A 市某经济学院强制学生外出学习的问题。

（3） 按照题目表 2 的格式给出的相应热点问题的对应留言信息部分数据展示见表 8：

表 8 热点问题留言明细部分数据展示

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	214282	A909209	A 市丽发新城小区附近搅拌站噪音扰民和污染环境	2020/1/25 9:07:21	你们管不管A2……	0	0
1	217700	A909239	丽发新城小区旁的搅拌站严重影响生活	2019/12/21 2:33:21	开发商把特大型搅拌……	0	1
1	244512	A000947 06	搅拌站丽发新城小区粉尘大的孩子生活不了	2019-12-05 20:57:50	我是暮云街道……	0	1
...	...	...	...	...	...	..	...
2	218709	A000106 692	A 市伊景园滨河苑捆绑销售车位	2019/8/1 22:42:21	伊景园滨河苑作为……	0	1
2	258037	A909190	投诉伊景园滨河苑捆绑销售车位问题	2019-08-23 11:46:03	尊敬的领导：我是广……	0	0
2	279941	A909177	广铁集团职工商品房竟然捆绑销售	2019-08-28 09:30:20	领导好！A 市广铁集……	0	0
...	...	...	...	...	...	..	...
3	278545	A000368 41	给 A 市经开区泉星公园项目规划进一步优化的建议	2019/8/26 13:00:06	目前A市经济技术开……	0	13
3	289574	A000803 42	对 A 市经开区泉星公园项目规划再进一步优化……	2019/8/16 13:33:59	目前 A 市经济技术……	0	0
3	238692	A000803 42	建议 A 市经开区泉星公园项目规划进一步优化	2019/8/12 13:15:05	目前A市经济技术开……	0	16
...	...	...	...	...	...	..	...
4	266031	A000316 18	请 A 市加快申报国家中心城市的建设	2019/2/27 9:36:24	现在国家在提都市圈……	0	8
4	270056	A000316 18	请加快 A 市国家中心城市建设刻不容缓	2019/1/11 14:26:48	作为内陆大省省会常……	0	5
4	287532	A000316 18	请 A 市加快国家中心城市建设力度	2019/1/8 10:04:45	作为内陆大省的 A 市……	1	2
...	...	...	...	...	...	..	...
5	255719	A000608 10	A 市商贸旅游职业技术学院强制学生实习	2019/4/28 17:32:51	各位领导干部大家好……	0	1

5	360114	A0182491	A 市经济学院体育学院变相强制实习	2017/6/8 17:31:20	书记您好，我是来自……	9	0
5	233759	A909118	A 市涉外经济学院强制学生实习	2019/04/28 17:32:51	各位领导干部大家好……	0	0

结果分析：

4 问题三 ：质量评价

4.1 问题分析

针对相关部门对留言的答复意见，这里主要从答复的相关性、完整性和可解释性的角度对其进行质量评价，引入这三个一级指标并确定综合指标，给出一套综合评价方案。

4.2 模型建立

4.2.1 质量评价模型

针对答复意见评价，我们从答复的相关性、完整性和可解释性的角度对答复意见的质量给出一综合评价方案。由参考文献[12]知，建立**答复意见的质量评价模型**如下：

$$Q_i = w_1 \cdot COR_i + w_2 \cdot JS_i + w_3 \cdot WZ_i \tag{12}$$

上式中Q_i为第i簇的留言信息的综合评价指标，w_1、w_2和w_3分别为各个特征的权重系数，其值满足$w_i > 0$并且$\sum_{i=1}^{3} w_i = 1$，COR_i、JS_i和WZ_i分别代表相关性、可解释性和完整性，融合这三个特征线性加权求和进行计算。

4.3 总流程图解

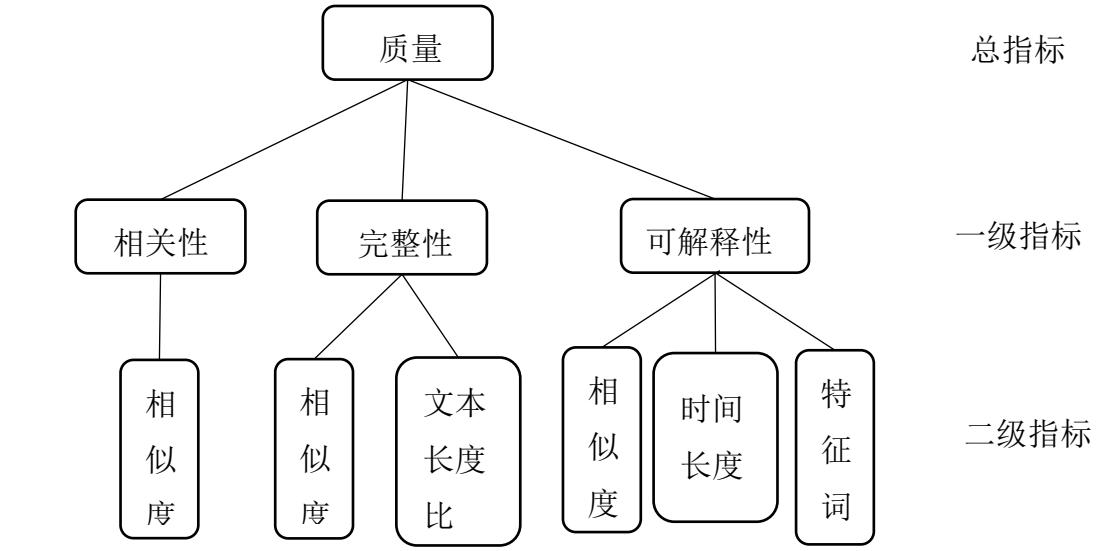

图 11 质量评价流程图

4.4 数据预处理

根据附件 4 的数据，数据内容比较多，且形式多样化，字段内容大小不一。避免一些重复数据和不成句的句子，造成训练结果的误差。时间序列格式多样，为了方便处理时间内容，对时间序列进行标准化。数据预处理流程图如图 12 所示：

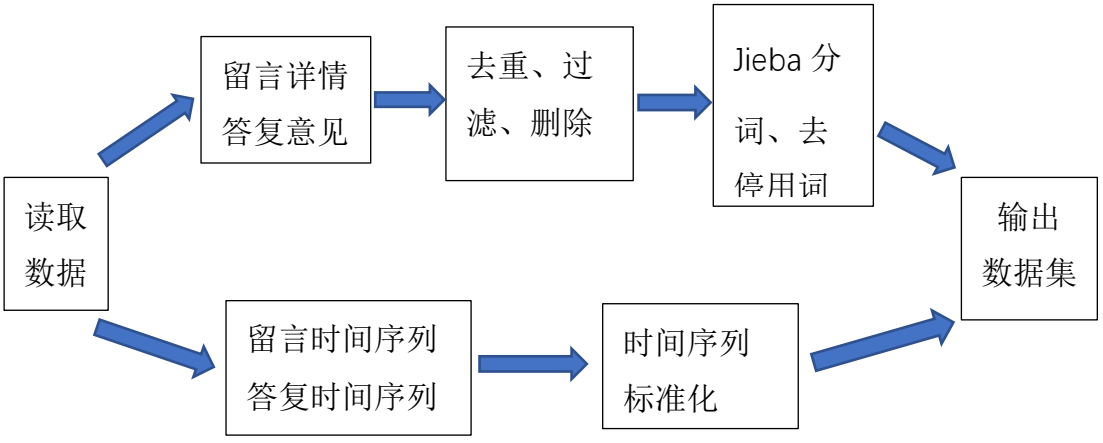

图 12 数据预处理

Setp1：先读取附件 4 数据

Setp2：提取留言详情和留言主题，留言时间和答复时间等数据

Setp3：对留言和答复数据进行去重、过滤掉一些少的数据，定义删除除字母,数字。

Setp4：采用传统的结巴分词，对每一条留言和答复数据进行分词后去停用词处理。

Setp5：对留言时间序列和答复时间序列进行标准化。

Setp6：把数据处理后的结果整理成一个新的数据集。

4.5 模型求解

4.5.1 相关性

1、字向量原理

字作为中文文本最小的组成单位，字包含了丰富的语义信息，特别是成语、人名、地名这样的超短文本，基于字向量进行分析能够获得更好的语义信息。

由参考文献[11]知，根据预处理后留言详情和答复意见的文本，为其建立字频表，为每一个字设置统一标识符，得出每条句子的字向量文本，流程如所图 13 所示：

```
# 最大长度
max_clean_review_comments_length
# 建立字频表
vocabulary_comments = {}
for clean_review_comments in X:
      ……
      “循环遍历每一留言”
# 按照字频排序
  vocabulary_comments_list
# 设置字频词典，为每一个字设置唯一标识符
  vocab = dict(……)
# 为每一个字建立向量
comments_vec = []
for 变量 in 留言主题：
```

图 13 字向量

对留言详情和答复意见进行向量表示后得到如下文本向量：

$$a_i = [v_{i1}, v_{i2}, v_{i3}, \cdots, v_{in}] \tag{13}$$

$$b_i = [v_{i1}, v_{i2}, v_{i3}, \cdots, v_{in}] \tag{14}$$

其中 a_i, b_i 分别代表留言详情词向量和答复意见词向量；i 表示留言条数或者是答复条数，$i = 1,2,\cdots,2815$；n 表示最大向量数，为保证向量维度一样，计算过程中取最大数，$n = 1,2,\cdots,3066$。

2、编辑距离相似度

Levenshtein Distance 是用来度量两个序列相似程度的指标，通俗的讲，编辑距离指的是在两个词 $< w_1, w_2 >$ 之间，由其中一个词 w_1 转换为另一个词 w_2 所需要的最少单字符编辑操作次数。

由参考文献[9]和参考文献[10]知，根据附件 4，将留言详情设为a_i，答复意见设为b_i，$i=0,1,2,\cdots,2815$。

我们将两个字符串a_i,b_i的 levenshtein Distance 表示为$lev_{a_i,b_i}(|a_i|,|b_i|)$ $i=0,1,2,\cdots,2815$，其中$|a_i|$和$|b_i|$分别对应a_i,b_i的长度的 Levenshtein Distance。

即$lev_{a_i,b_i}(|a_i|,|b_i|)$可用如下的数学语言描述：

$$lev_{a_i,b_i}(i,j)=\begin{cases}\max(i,j) & if \quad \min(i,j)=0,\\ \min\begin{cases}lev_{a_i,b_i}(i-1,j)+1\\ lev_{a_i,b_i}(i,j-1)+1 & otherwise\\ lev_{a_i,b_i}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases}\end{cases} \tag{15}$$

（1）定义$lev_{a_i,b_i}(|a_i|,|b_i|)$指的是$a_i$中的前$i$个字符和$b_i$中前$j$个字符之间的距离，为了方便理解，这里的$i,j$可以看作是$a_i,b_i$的长度。这里的字符串的第一个字符 index 从 0 开始，因此最后的编辑距离便是$i=|a_i|,j=|b_i|$时的距离：

$$lev_{a_i,b_i}(|a_i|,|b_i|)=lev_{a_i,b_i}(i,j) \tag{16}$$

（2）当$\min(i,j)=0$的时候，对应着字符串a_i中的前i个字符和b_i中前j个字符，此时的i,j有一个值为 0，表示字符串a_i和b_i中有一个为空串，那么从a_i转换到b_i只需要进行$\max(i,j)$次单字符编辑操作即可，所以它们之间的编辑距离$\max(i,j)$，即i,j中的最大者。

（3）当$\min(i,j)\neq 0$的时候，$lev_{a,b}(|a_i|,|b_i|)$为如下三种情况的最小值：

①. $lev_{a_i,b_i}(i-1,j)+1$表示删除a_i

②. $lev_{a_i,b_i}(i,j-1)+1$表示插入b_i

③. $lev_{a_i,b_i}(i-1,j-1)+1_{(a_i\neq b_j)}$表示替换$b_i$

（4）$1_{(a_i \neq b_j)}$为一个指示函数，表示当$a_i = b_j$的时候取 0；当$a_i \neq b_j$的时候，其值为 1。

综上所述，**相关性计算**如下：

$$COR_i = lev_{a_i,b_i}(|a_i|,|b_i|) \tag{17}$$

4.5.2 完整性

相关性角度主要代表了相关部门给的答复意见与群众留言内容相关强度的大小，即问答是否一致。从完整性角度而言，与相关性是具有密切联系的，在相关性的基础上，考量回复的信息量是否完整，即对留言内容谈及的每一个问题都有进行答复，考量回复是否面面俱到。

在这里引入文本长度这个二级指标，首先计算每条留言数据的详情长度，以及每条答复意见的文本长度，其次将两个文本长度做比值，再引入相关性指标得到一个误差，最后引入相关性联系。对于**完整性计算**如下：

$$WZ_i = \sqrt{COR_i \times \sqrt{\frac{len_{huifu-i}}{len_{liuyan-i}}} + COR_i} \tag{18}$$

上式$i = 0,1,2,\cdots,2815$，中WZ_i代表第i条留言信息完整性指标，COR_i代表第i条留言信息的相关性指标，len_{dafu-i}代表第i条答复意见文本长度，$len_{liuyan-i}$代表第i条留言详情长度。

4.5.3 可解释性

针对可解释性，这里引入了相关性指标，问答时间差，特征词长度三个特征作为二级指标，通过对这三个特征进行线性加权求和来计算可解释性。

根据留言时间至答复时间，先对时间序列进行标准化，计算问答**时间差**如下：

$$time_i = t_{di} - t_{li} \tag{19}$$

其中 $i = 0,1,2,\cdots,2815$，$time_i$ 表示第 i 条留言答复时间差，t_{di} 表示第条 i 答复时间，t_{li} 表示第 i 条留言时间。

根据答复意见文本表示，提取出特殊词汇，例如：《XX 条例》、《XX 方案》、答复情况、《XX 法》、XX 政府调查、“第一、第二、第三”等等具有代表性的词汇，制作特殊词汇字典，**特征长度**计算如下：

$$A_i = (A_{i1}, A_{i2}, \cdots, A_{in}) \tag{20}$$

$$word_i = sum(A_i) \tag{21}$$

其中 $i = 0,1,2,\cdots,2815$，A_i 表示第 i 条答复的特殊词，$word_i$ 表示第 i 条特殊词长度。

综上所述，**可解释性计算**如下：

$$JS_i = w_1 \cdot COR_i + w_2 \cdot time_i + w_3 \cdot word_i \tag{22}$$

上式中，$i = 0,1,2,\cdots,2815$ 为第 i 条留言信息的可解释性特征，$COR_i, time_i$ 和 $word_i$ 分别表示第 i 条留言信息的相关性、留言与答复时间差和特殊词长度这三个特征，为各个特征的权重系数，其值满足 $w_i > 0$ 并且 $\sum_{i=1}^{3} w_i = 1$，至此，将可解释性评价体系转换成了定量的计算过程，实现了可解释性特征的度量。

4.5.4 各级评价体系特征权值的计算

根据本文构建的答复意见质量评价体系层次结构，根据参考文献[8]和[12]可知，本文使用层次分析法(AHP)计算各个权重的权值，最后通过对各个特征进行

线性加权求和得到上一级特征值，本文采用几何平均法对权向量进行计算。具体计算步骤如下：

（1）特征判断矩阵的构造

层次分析法引入数字 1-9 以及其倒数作为衡量因素之间重要性的标度，来定义判断矩阵 $A=(a_{ij})_{n\times n}$，如表 9 所示判断矩阵标度进行定义。本文在定义可解释性特征的时候，认为其评价体系中二级指标的重要程度高至低依次为相关性、时间差、特征词长度；在定义综合评价指标的时候，认为其综合评价体系中一级指标的重要程度由高到低为相关性、可解释性、完整性。根据以上分析和层次分析法判断矩阵的定义规则，给定判断矩阵初始值，并通过迭代和一致性检验得到如式（23）、（24）所示的判断矩阵。具体迭代和一致性检验将在如下进行介绍。

表 9 判断矩阵标度定义

标度	含义
1	两个因素相比，具有相同的重要性
3	两个因素相比，前者比后者稍重要
5	两个因素相比，前者比后者明显重要
7	两个因素相比，前者比后者强烈重要
9	两个因素相比，前者比后者极端重要
2、4、6、8	表述上述相邻判断的中间值

倒数	若因素 i 与因素 j 的重要性之比为 a_{ij}，那么因素 i 与因素 j 重要性之比为 $a_{ji}=1/a_{ij}$

$$A_{JS}=\begin{bmatrix}1 & 5 & 3\\ 1/5 & 1 & 1/4\\ 1/3 & 4 & 1\end{bmatrix} \tag{23}$$

$$A_{JS}=\begin{bmatrix}1 & 3 & 4\\ 1/3 & 1 & 2\\ 1/4 & 1/2 & 1\end{bmatrix} \tag{24}$$

其中 A_{JS} 可解释性特征判断矩阵，A_{ZH} 为综合指数特征判断矩阵。

（2）求解判断矩阵的最大特征向量并进行一致性检验

本文采用几何平均法计算特征权值向量 W，计算公式如下：

$$W_i=\frac{(\prod_{j=1}^{n}a_{ij})^{\frac{1}{n}}}{\sum_{i=1}^{n}(\prod_{j=1}^{n}a_{ij})^{\frac{1}{n}}},\quad i=1,2,3,...,n \tag{25}$$

其中 W_i 为第 i 个特征的权值。

为了对判断矩阵进行一致性检验，需要对判断矩阵最大特征值 $\lambda_{\max}$ 进行计算，计算方法如下：

a、判断矩阵乘以特征权值向量 V，得到矩阵 N；

b、再对矩阵 N 中的每一个元素除以 n 与权值向量 V 的乘积；

c、矩阵 N 中最大的值即为判断矩阵最大值 $\lambda_{\max}$。

其中 n 为特征个数，本文取值为 3。

在得到特征权值向量和判断矩阵最大特征值后，接下来需要对判断矩阵进行一致性检验，具体过程如下：

① 一致性指标的计算：

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{26}$$

其中为CI一致性指标，$\lambda_{\max}$为判断矩阵最大特征值，n为特征个数，本文取值为3。

② 查表确定平均随机一致性指标$RI(RandomIndex)$：

平均随机一致性指标RI为一致性指标CI的期望值，表示CI的集中程度。查表当$n=3$时，$RI=0.58$。

③ 计算一致性比例

当$CR<0.10$时，认为判断矩阵的一致性是可以接受的，否则就应该对判断矩阵进行修改。CR的计算如下式：

$$CR=\frac{CI}{RI} \tag{27}$$

通过对矩阵（23）、（24）进行一致性检验，计算得$CR_{JS}=0.0739, CR_{Q}=0.0760$都小于0.10，因此这两个判断矩阵都满足一致性检验的要求。于是权值向量$V_{JS}=(0.63,0.28,0.09), V_{Q}=(0.63,0.25,0.13)$分别为（22）、（28）式中对应的$(w_1,w_2,w_3)$。

4.5.5 综合评价体系特征确定

完整性和可解释性与相关性指标有一定的联系，衍生于相关性。在确定特征权值时，本文认为留言内容与答复意见相关性的重要程度最高，因此如果问答之间的相关程度越高，就越能代表答复意见的质量越高；其次为可解释性特征，这里不仅引入了特征词向量之间的距离，而且还增加了问答时间差因素，当回复时间越快，相关政府部门就越能够采取相应的措施快速解决群众的问题。所以，根据综合评价体系确定了特征的重要性排行：相关性>可解释性>完整性；通过层次分析法的计算，最终确定出各个特征的权值w_i按照重要程度由高到低的顺序依次为(0.63, 0.24, 0.13)。

通过解决相关性、完整性、可解释性以及权重划分的问题，计算质量评价模型：

$$Q_i = w_1 \cdot COR_i + w_2 \cdot JS_i + w_3 \cdot WZ_i \tag{28}$$

上式中Q_i为第i簇的留言信息的综合评价指标，w_1、w_2和w_3分别为各个特征的权重系数。

参考文献

[1] Yu Sun, ShuohuanWang, Yukun Li. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding[R]. 北京. 百度公司. 2019 年

[2] 飞桨 PaddlePaddle. 一文读懂最强中文 NLP 预训练模型 ERNIE[EB/OL]. https://baijiahao.baidu.com/s?id=1648169054540877476&wfr=spider&for=pc.2020/05/05

[3] 百度百科.F1 分数[EB/OL]. https://baike.baidu.com/item/F1%E5%88%86%E6%95%B0/13864979?fr=aladdin .2020/05/05

[4] 叶建成.利用文本挖掘技术进行新闻热点关注问题分析[D].广州.广州大学.2018 年

[5] 黄静，官易楠.基于改进 DBSCAN 算法的异常数据处理[A].浙江理工大学.2019 年

[6] 王安瑾.一直基于 MimHash 的改进新闻文本聚类算法[A].东华大学.2019 年

[7] 百度百科.tf-idf[ER/OL]. https://baike.baidu.com/item/tf-idf/8816134?fr=aladdin.2020/05/05

[8] 闫威，陈长怀，陈燕.层次分析法一致性指标的临界值研究[J].数理统计与管理.2011,30(3):414-423

[9] 周伟杰.面向问答领域的语义相关性计算的研究[D].广州.华南理工大学.2017 年

[10] 王宝勋.面向网络社区问答对的语义挖掘研究[D].哈尔滨工业大学.2013 年

[11] 王宇.一种基于字向量和 LSTM 的句子相似度计算方法[J].长江大学学报.2019 年

[12] 胡鹏辉.基于多模型的问答社区答案质量评价研究[D].南京师范大学.2019 年