基于自然语言处理的“智慧政务”

摘要

本文旨在设计基于自然语言处理技术的**智慧政务**系统，以群众问政留言记录及相关部门答复意见为原始数据，实现了对群众的留言的一级标签分类、留言的热点问题挖掘与相关部门答复意见的评价体系的构建这三个功能。

针对问题一，本文利用基于BERT预训练语言模型对附件2数据构建一级标签分类模型。该分类模型主要由文本预处理、文本向量化和分类器等三部分构成。利用双向Transformer编码器中的Encoder结构实现文本向量化，利用Softmax回归模型搭建分类器。其中，Encoder结构主要运用Self-attention算法实现，Softmax回归模型主要用梯度下降算法训练实现。最后，该模型在测试集上测试的分类整体F-score值可达到**0.9194**，查准率可达到**0.9131**，查全率可达到**0.9198**。

针对问题二，定义了三个热度评价指标：同类话题贴子数，该类话题帖子总赞数，该话题总反对数。建立热点问题评价模型：热度=w1*同类话题贴子数+w2*该类话题帖子总赞数+w3*该话题总反对数。首先对附件3的留言主题与留言详情文本进行jieba分词，去停用词等预处理，再利用TF-IDF算法,将其转化为固定维度的向量，再基于DBSCAN 聚类算法，对语句向量进行聚类，得到同话题帖子数，再基于热点问题评价模型进行热度排名，得到排名前5的热点问题和热点留言明细。

针对问题三，本文定义了四个指标：jaccard系数，余弦相似度，答复意见的文本长度，和答复时间差，建立答复意见评价模型：质量评分=w1*余弦相似度+w2*jaccard系数+w3*答复时间差+w4*答复文本长度。利用余弦相似度公式，word2vec词向量化算法与jaccard公式求解余弦相似度和jaccard系数，最后按答复意见频数将答复意见划分为底、中等、高三个层次。

关键词：智慧政务； Bert ；DBSCAN聚类；TF-IDF ；Jaccard系数 ；

Abstract

The purpose of this paper is to design an intelligent government system based on natural language processing technology, which takes the record of the public's political message and the reply of relevant departments as the original data, and realizes the three functions of the first level label classification of the public's message, the hot topic mining of the message and the construction of the evaluation system of the reply of relevant departments.

To solve the first problem, this paper uses the pre training language model based on Bert to build the first level label classification model for the attachment 2 data. The classification model consists of three parts: **text preprocessing**, **text vectorization** and **classifier**. The encoder structure of the bidirectional transformer encoder is used to realize the text vectorization, and the softmax regression model is used to build the classifier. Among them, encoder structure is mainly realized by self-attention algorithm, and softmax regression model is mainly realized by gradient descent algorithm training. Finally, the overall F-score of the model can reach **0.9194**, the precision can reach **0.9131** and the recall can reach **0.9198**.

To solve the second problem, three heat evaluation indexes are defined: **the number of posts on the same topic**, **the total number of likes** and **objections on this topic**. **Establish the evaluation model of hot issues: heat = W1 * number of posts on the same topic + W2 * total likes - W3 * total objections on the topic**. Firstly, the message subject and message detail text in Annex 3 are preprocessed by Jieba word segmentation to remove stop words, and then **TF-IDF** algorithm is used, It is transformed into a fixed dimension vector, and then based on the **DBSCAN** clustering algorithm, the statement vector is clustered to get the number of posts on the same topic, and then based on the hot issue evaluation model, the hot issues and hot message details in the top 5 are obtained.

To solve the third problem, this paper defines four indexes: **Jaccard coefficient**, **cosine similarity**, **text length of reply opinion**, and **reply time difference**. Using cosine similarity formula, word 2vec vectorization algorithm and Jaccard formula to solve cosine similarity and Jaccard coefficient. Finally, according to the response frequency, the response is divided into three levels: bottom, middle and high.

Keywords: smart government; Bert; DBSCAN clustering; TF-IDF; Jaccard coefficie.

目录

第 1 章 问题重述

1.1 问题背景

近年来，随着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，建立基于自然语言处理技术的智慧政务系统，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. 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，并预测测试集的留言内容的一级标签类别和对预测结果进行评价。
2. 根据附件3给出的数据，定义合理的热度评价指标，并给出评价结果，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并分别保存在两张表内。
3. 根据附件4相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

第 2 章 群众留言分类

2.1 问题分析

已知附件 2 的留言消息，根据这些数据建立关于留言内容的一级标签分类模型，对所给预测数据进行预测，并对分类结果进行评价，大致思路可以分如表 1 以下四点。

表 1 问题分析

1.怎么处理原始数据

2.运用什么算法实现文本向量化

3.运用什么分类器进行分类

4.怎么提高模型评价结果

2.2 模型建立

本文建立的基于 BERT 的一级标签分类模型如图 1，主要由短文本预处理、短文本向量化以及分类器分类等三部分构成，先对数据进行 BERT 预处理切分成词向量，接着使用该模型对预处理后的留言短文本进行句子层面的特征表示，最后将获得的特征向量接入 Softmax 分类器进行留言分类，实验结果表明了本文所提方法的有效性。短文本预处理的目的是将留言句子切分和转化为词向量，随后，对预处理后的短文本进行向量化表示，形成特征向量，最后，将特征向量输入搭建好的分类器中，进而实现对留言短文本的分类，输出留言的一级标签分类结果。

输入： 初始中文文本训练集$T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$,其中$x_i$为每条中文短文本， y_i为每条短文本所属的类别，i =1,2,...,N。

输出：中文文本分类模型M。

步骤1：使用2.3.2节中方法对训练集T 进行预处理， 得到预处理后的训练集$T'=\{(x_1',y_1'),(x_2',y_2'),...,(x_N',y_N')\}$,其中$x_i'$为预处理后的中文短文本，$y_i'$为预处理后每条短文本所属的类别，i =1,2,...,N。

步骤2：使用2.2 节中介绍的BERT 预处理语言模型在训练集T'上进行微调，采用如图6所示的BERT 模型输出，得到训练集T'对应的特征表示为$V=(v_1,v_2,...,v_{N'})$，其中v_i是每条短文本x_i'对应的句子级别的特征向量, i =1,2,..., N'。

步骤3：将步骤2 中得到的特征表示V 输入2.3节中介绍的Softmax回归模型进

行训练，输出短文本分类模型M。

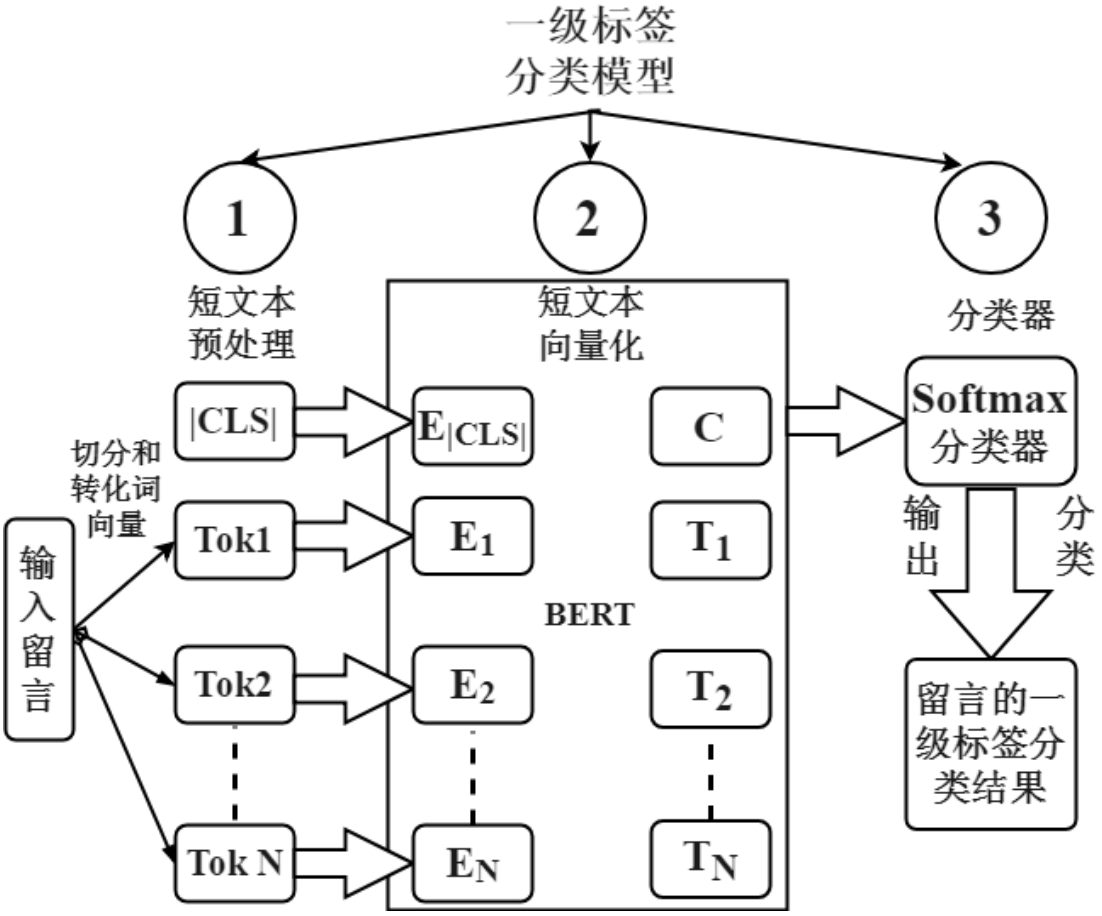


图 1 一级标签模型结构图

2.2.1 BERT 数据预处理

如图 2 为附件 2 某一留言的预处理，在其首尾分别添加[CLS]、[SEP]。对于本模型的输入，每一个词语的表示都由词语向量（Token Embeddings）、 段向量（Segment Embeddings）和位置向量（Positional Embeddings）相加产生。

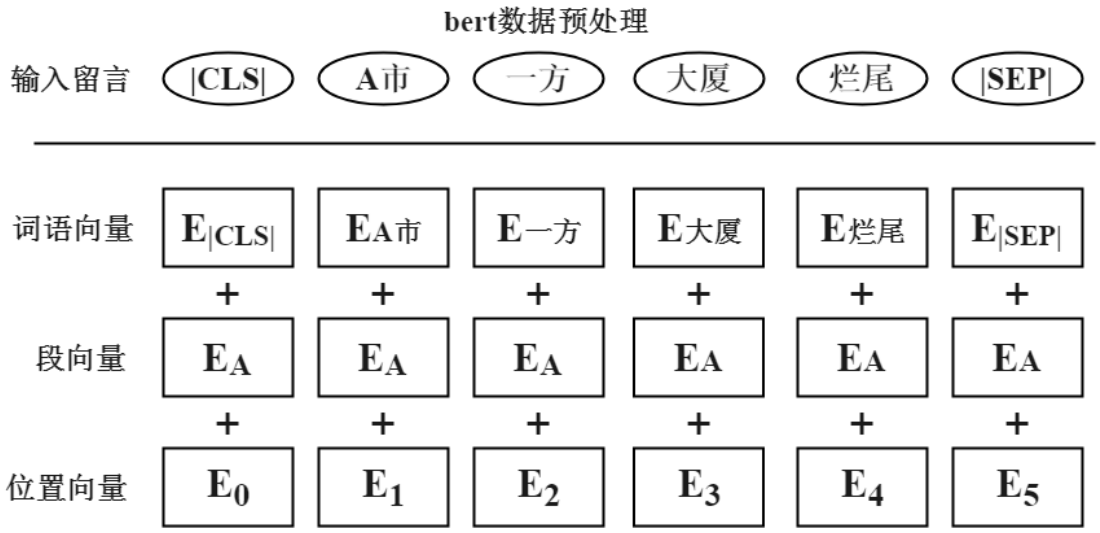


图 2 留言预处理

其中，每个输入句子的第一个标记都是[CLS]，对应的是 Transformer 的输出，是用来表示整个句子的，可用于下游的分类任务。标记[SEP]是用来分隔两个句子的，对于句子分类任务，只需对一个留言句子进行输入，即对于单句仅使用一个段向量。

2.2.2 短文本向量化

如图 3 所示，输入留言短文本预处理后切分的词向量进行句子层面的特征表示，最后将获得的特征向量。本模型的文本向量化是通过双向 Transformer 编码器实现的，其模型结构[1]。图中的 $E_1, E_2, ..., E_N$ 表示词的文本输入，经过双向的 Transformer 编码器，就可以得到文本的向量化表示。

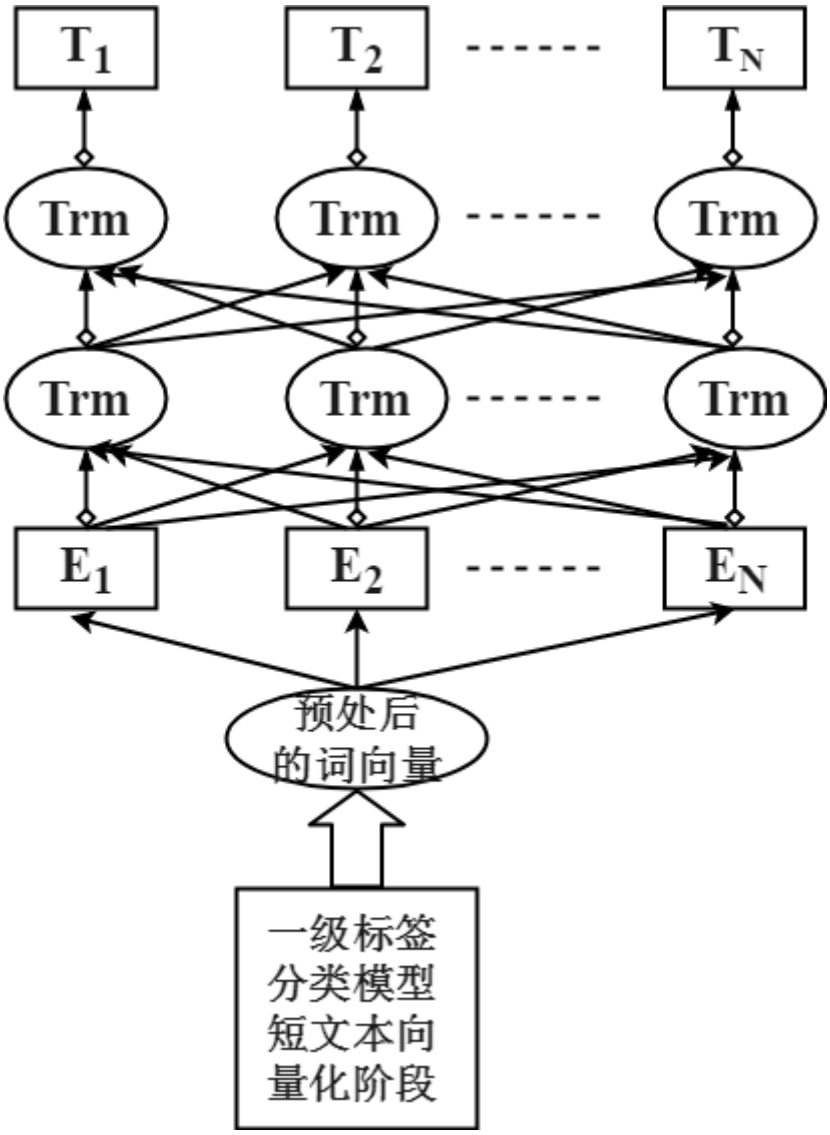


图 3 BERT 文本向量化

Transformer 由文献[2]提出，它是一个基于 Self-attention 的 Seq2seq 模型，Seq2Seq 是一个 Encoder-Deocder 结构的模型，即输入是一个序列，输出也是一个序列，其中，Encoder 将一个可变长度的输入序列变为固定长度的向量,Decoder 将这个固定长度的向量解码成可变长度的输出序列，模型结构可简化如图 4 所示：

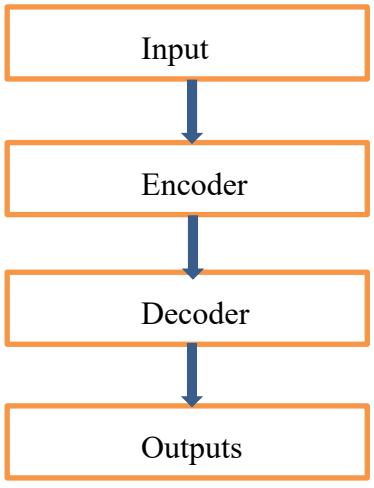


图 4 Seq2Seq 模型结构

通常，解决这种序列问题的 Encoder-Decoder 结构的核心模块是基于 RNN 实现的，但是 RNN 存在无法并行、运行慢的缺点，为了改进这个不足，Transformer 使用 Self-attention 来替代 RNN。BERT 模型中主要使用的是 Transformer 的 Encoder 部分，下面就着重介绍一下 Transformer 模型中 Encoder 的结构，如图 3 所示：

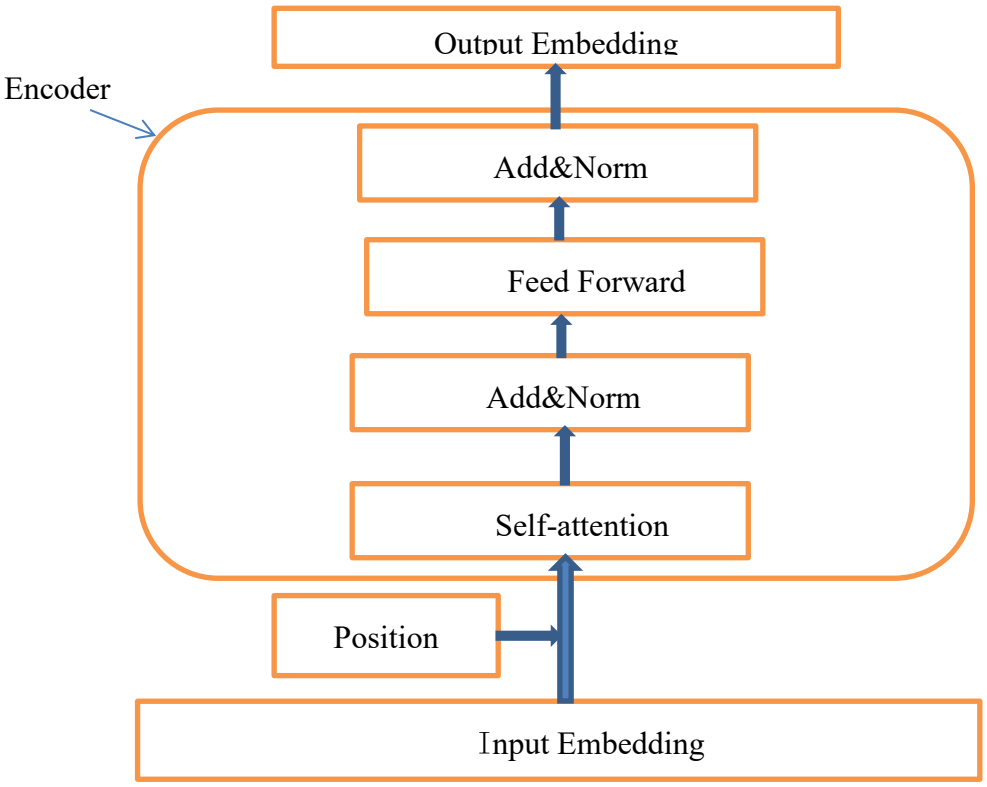


图 5 Transformer Encoder 结构

从图 5 中可以看出，Encoder 的输入是一句话的字嵌入表示，随后加上该句话中每个字的位置信息，之后经过 Self-attention 层，帮助 Encoder 在编码每个字的时候可以查看该字的前后字的信息，它的输出会再经过一层 Add & Norm 层，Add 表示将 Self-attention 层的输入和输出进行相加，Norm 表示将相加过的输出进行归一化处理，使得 Self-attention 层的输出有固定的均值和标准差，均值为 0，标准差为 1，归一化后的向量列表会再传入一层全连接的前馈神经网络，同样的，Feed Forward 层也会有相应的 Add & Norm 层处理，之后输出全新的归一化后的词向量列表。

Encoder 部分中最主要的模块是 Self-attention，其核心思想是去计算一句话中的每个词与这句话中 所有词的相互关系，再利用这些相互关系来调整每个词的权重就可以获得每个词新的表达，这个新的 表达不但蕴含了该词本身的语义，

还蕴含了其他词与这个词的关系，因此和传统的词向量相比是一个更加全局的表达。Self-attention 的计算步骤如下：

假设有输入句子，将其按照字粒度进行分字可表示为 $X=(x^1,x^2,...,x^N)^T$，N 表示该输入句子中字的个数，现将每个字采用 One-hot 向量表示，设维数为 k，则对应的字嵌入矩阵为 $A=(a^1,a^2,...,a^N)^T$ ，其中 a^i 是对应 x^i 的向量表示，是一个 k 维向量，显然 A 是一个 $N*K$ 的矩阵，每一行对应于该输入句子中的一个字的向量表示，

综上，可以得出，BERT 模型使用双向的 Transformer 的 Encoder 部分可以学习每个单词前后两边的信息，获得更好的词向量表示。

BERT 模型创新性的提出了两个新的预训练任务：Masked LM (MLM) 和 Next Sentence Prediction (NSP)：

一、MLM 任务：给定一句话，随机抹去其中的一个或者几个词，用剩余的词去预测这几个词分别是什么。这个任务在业界被称为 Cloze 任务（完形填空），它是为了让 BERT 模型能够实现深度的双向表示，具体做法是：针对训练样本中的每个句子随机抹去其中 15%的词汇用于预测，取附件 2 的一个留言来分析，例如：“A 市在水一方大厦烂尾”，被抹去的词是“尾”，对于被抹去的词，进一步采取以下如表 2 策略：

表 2 MLM 任务

1.80%的概率真的用[MASK]去替代被抹去的词：

- “A市在水一方大厦烂尾”→A市在水一方大厦烂[MASK]”

2.10%的概率用一个随机词去替代它：

- “A市在水一方大厦烂尾”→A市在水一方大厦烂掉”

3.10%的概率保持不变：

- “A市在水一方大厦烂尾”→A市在水一方大厦烂尾”

这样做的主要原因是：在后续微调任务中语句中并不会出现[MASK]标记，若总是使用[MASK]来替代被抹去的词，就会导致模型的预训练与后续的微调不

一致。这样做的优点是：当预测一个词汇时， 模型并不知道输入的词汇是否为正确的词汇，这就使得模型更多地依赖于上下文信息去预测词汇，赋予了模型一定的纠错能力。另外，这里只随机替换了 1.5%的词为其他词，整体上不会损害模型的语言理解能力。

二、NSP 任务：给定一篇文章中的两句话，判断第二句话在文章中是否紧跟在第一句话之后。许多重要的自然语言处理下游任务，如问答（QA）和自然语言推理（NLI）都是基于理解两个句子之间的关系，因此这个任务是为了让 BERT 模型学习到两个句子之间的关系。具体做法是：从文本语料库中随机选 择 50%正确语句对和 50%错误语句对，即若选择 A 和 B 作为训练样本时，B 有 50%的概率是 A 的下一个句子，也有 50%的概率是来自语料库中随机选择的句子，本质上是在训练一个二分类模型，判断句子之间的正确关系。在实际训练中，NSP 任务与 MLM 任务相结合，让模型能够更准确地刻画语句乃至篇章层面的语义信息。

2.2.3 Softmax 回归模型

经过 2.2.1，2.2.2 节的处理后，有了短文本的向量表示，本文引入 Softmax 回归模型作为分类器来进行留言短文本的分类见图 6。

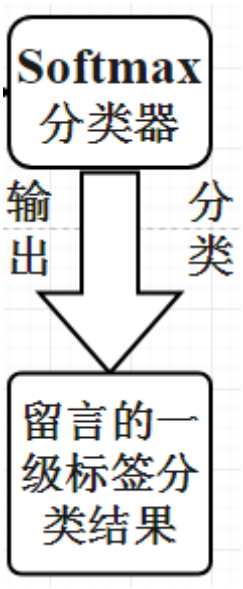


图 6 分类输出过程

Softmax 回归模型是 Logistic 回归模型在多分类问题上的推广，均属于广义线性模型。假设有训练样本集 $\{(x^1,y^1),(x^2,y^2),...,(x^m,y^m)\}$，其中 $x^i \in R^n$，表示第 i 个训练样本对应的短文本向量，维度为 n，共 m 个训练样本，$y^i \in \{1,2,...,k\}$，

表示第 i 个训练样本对应的类别，k 为类别个数，本文研究的是短文本多分类的问题，因此类别数 $k \geq 2$。给定测试输入 x，Softmax 回归模型做分类判别的分布函数为条件概率 $p(y = j \mid x)$,即计算给定样本 x 属于第 j 个类别的概率，其中出现概率最大的类别即为当前样本 x 所属的类别，因此最终分布函数会输出一个 k 维向量，每一维表示当前样本属于当前类别的概率，为了保证归一化，模型将 k 维向量的和做归一化操作，即向量元素的和为 1。因此，Softmax 回归模型的判别函数 $h_\theta(x)$ 为[3]：

$$h_\theta(x^i) = \begin{bmatrix} p(y^i = 1 \mid x^i;\theta) \\ p(y^i = 2 \mid x^i;\theta) \\ \vdots \\ p(y^i = k \mid x^i;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^i}} \begin{bmatrix} e^{\theta_1^T x^i} \\ e^{\theta_2^T x^i} \\ \vdots \\ e^{\theta_k^T x^i} \end{bmatrix} \quad (1)$$

其中 $\theta_1, \theta_2, \ldots, \theta_k \in R^n$ 是模型的参数。

$h_\theta(x^i)$ 是一个向量，向量中任一元素 $p(y^i = k \mid x^i;\theta)$ 是当前输入样本 x^i 属于当前类别 k 的概率，并且向量中各个元素之和等于 1，其中 θ 为模型的总参数，$\theta_1, \theta_2, ..., \theta_k$ 是各个类别对应的分类器参数，具体有如下关系：

$$\theta = [\theta_1^T, \theta_2^T, ..., \theta_k^T]^T \quad (2)$$

该矩阵中的每行为一个类别对应的分类器参数。Softmax 回归模型的参数估计可用极大似然法来求解，似然函数为：

$$L(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{k} \left(\frac{e^{\theta_j^T x^i}}{\sum_{j=1}^{k} e^{\theta_j^T x^i}} \right) I\{y^i = j\} \quad (3)$$

其中 $I\{\cdot\}$ 是示性函数：

$$\mathrm{I}\{y^i = j\} = \begin{cases} 1, y^i = j \\ 0, y^i \neq j \end{cases} \quad (4)$$

对数似然函数为：

$$\mathrm{I}(\theta) = \ln(L(\theta)) = \sum_{i=1}^{m}\sum_{j=1}^{k} I\{y^i = j\} \cdot \ln \frac{e^{\theta_j^T x^i}}{\sum_{j=1}^{k} e^{\theta_j^T x^i}} \quad (5)$$

一般情况下，Softmax 回归模型是通过最小化损失函数求得θ的数值，从而预测一个新样本的类别。定义 Softmax 回归模型的损失函数[4]为：

$$\mathrm{J}(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} I\{y^i = j\} \ln \frac{e^{\theta_j^T x^i}}{\sum_{j=1}^{k} e^{\theta_j^T x^i}}] \quad (6)$$

其中，m 为样本个数，k 为类别标签的个数，i 表示某一个样本，x^i 是第 i 个样本的向量表示，j 表示某一个类别标签。

为优化上述损失函数，本文使用随机梯度下降法，由于在 Softmax 回归模型中，样本 x 属于类别 j 的概率为：

$$\mathrm{p}(y_i = j \mid x^i; \theta) = \ln \frac{e^{\theta_j^T x^i}}{\sum_{j=1}^{k} e^{\theta_j^T x^i}} \quad (7)$$

因此损失函数的梯度为：

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}(\sum_{i=1}^{m}[x^i(I\{y^i = j\} - \mathrm{p}(y_i = j \mid x^i; \theta))]) \quad (8)$$

梯度下降法的实现过程中，每一次迭代需要按照下列公式更新参数[5]：

$$\theta_j = \theta_j - \alpha \nabla_{\theta_j} J(\theta), (j = 1, 2, ..., k) \quad (9)$$

由此方法求出θ，得到判别函数$\mathrm{h}_\theta(x)$，即可以实现对新的输入留言数据进行预测分类。

2.3 模型求解

2.3.1 实验条件

表 3 实验条件

所需要的库	Tqdm sklearn tensorboardX
GPU	Nvidia GTX 1080Ti
编程语言	Python3.6
深度学习框架	Pytorch1.0
配置文件（预训练模型）	pytorch_model.bin bert_config.json vocab.txt

2.3.2 数据预处理

去除没语义特征的留言，让短文本的特征表示尽可能地只关注短文本自身词汇的特征和语义本身，降低其他符号对分类准确率的影响，方便 BERT 算法后续的文本向量化表示。如下图 7 是没语义特征的留言如韦什模：

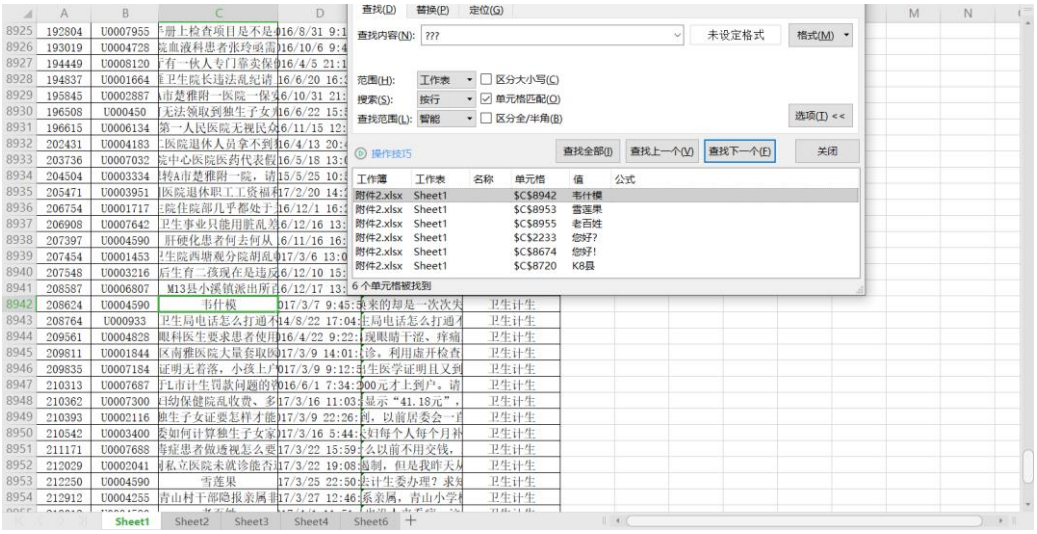


图 7 没语义特征留言

根据附件 2 对数据集按 7:2:1 划分训练集，验证集，测试集，并整理到 txt 文本里，训练集与验证集每行有留言内容和对应类别标签如下图 8，

```
A8县市花明楼镇花明楼村贵庭住工毁我池塘，生产沥青毒气！    1
C4市“鑫典和府”违规建筑为何没人管？    0
投诉K8县桂冠中学违规招生    3
K8县一中高二老师在外私自办培训机构    3
K9县非法造纸厂严重污染楚江水系舜水河    1
G1区泽云广场内星瑞国际打着免费体验的幌子，给老年人洗脑卖保健品    6
B8县乡镇外线工人待遇最低，公司按社会最低工资标准发其工资    4
A市在水一方大厦人为烂尾多年，安全隐患严重    0
西地省D市技师学院领导贪污堕落道德败坏故意滥用权力欺骗上级    3
```

图 8 数据预处理

2.3.3 算法流程图

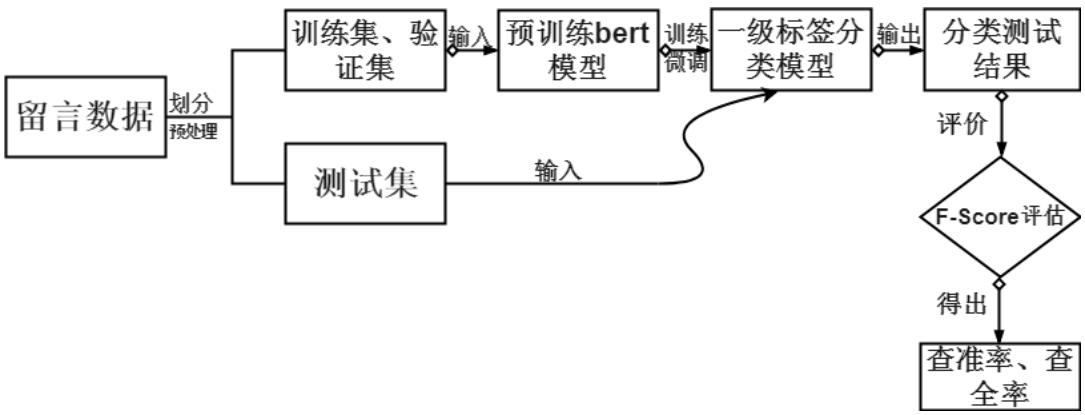


图 9 算法流程图

2.3.4 程序框架

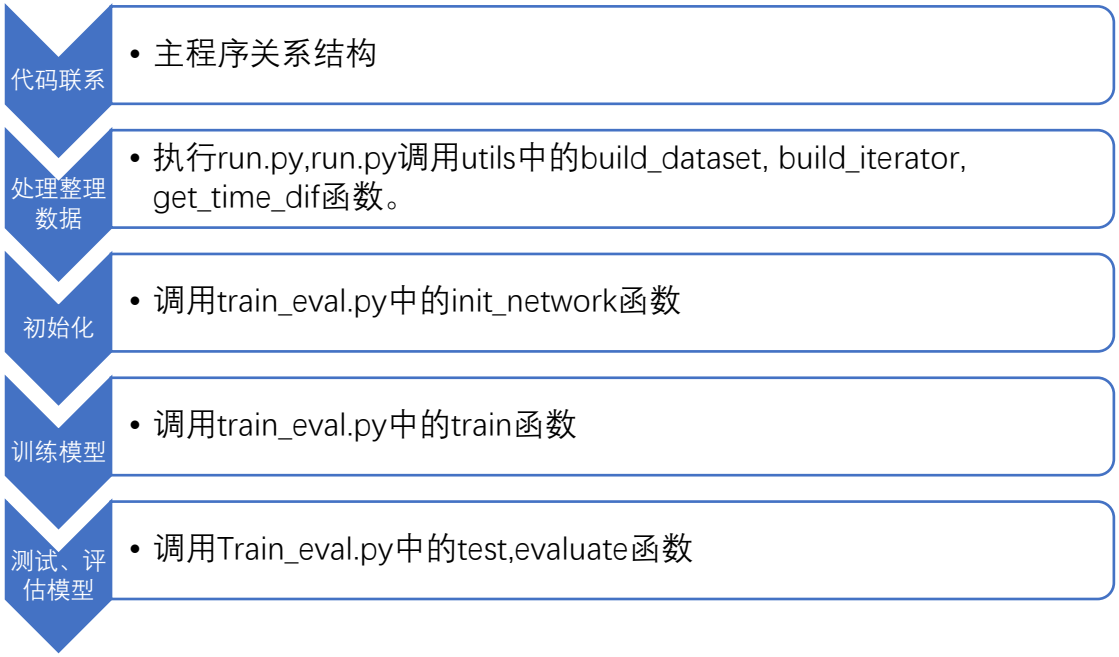
1.程序框架表如表 4

表 4 程序框架

Bert 预训练语言模型放在 bert_pretrain	主程序	模型程序
• pytorch_model.bin • bert_config.json • vocat.txt	• run.py • train_eval.py • utils.py	• bert.py

2.主程序关系结构如表 5

表 5 主程序关系结构



3.执行步骤：

在 run.py 文件目录启动命令行执行：

```
python run.py –model bert
```

便可执行整个 bert 模型程序。

2.3.5 求解步骤

第一步：将附件 2 的留言数据按 7:2:1 划分训练集、验证集、测试集。
第二步：将训练集、验证集输入预训练 BERT 模型，微调参数，以训练 5 个周期，学习率设为 2e-5，每批训练大小设为 32，运行 run.py 文件训练一级标签分类模型，反复调参试验达到最高的 f1 得分，最终得出网络结构一共 12,层，隐藏层有 768 维，采用 12 头模式，共有 110M 个参数，其他训练参数如表 6 所示。
第三步：将测试集输入一级标签分类模型的 Softmax 分类器，输出测试数据分类结果，fi 得分、精准率、召回率还有混淆矩阵。

表 6 BERT 模型参数

BERT 模型参数	值
layer	12
模型训练周期(epoch)	5
学习率(learning_rate)	2e-5
每批训练集数据大小(batch_size)	32
每句话处理的长度(pad_size)	40
hidden	768

2.3.6 模型的训练

• 调节至如表 6 的参数训练模型，模型训练过程展示如下表 7：

表 7 训练过程

Epoch [1/5]
Iter: 0, Train Loss: 2.1, Train Acc: 0.00%, Val Loss: 1.9, Val Acc: 15.60%, Time: 0:00:04 *
Iter: 100, Train Loss: 0.54, Train Acc: 81.25%, Val Loss: 0.61, Val Acc: 81.20%, Time: 0:00:31 *
Iter: 200, Train Loss: 0.31, Train Acc: 90.62%, Val Loss: 0.43, Val Acc: 87.12%, Time: 0:00:58 *

```
Epoch [2/5]
Iter:    300,  Train Loss:  0.37,  Train Acc: 87.50%,  Val Loss:  0.41,  Val Acc: 88.10%,  Time: 0:01:25 *
Iter:    400,  Train Loss: 0.081,  Train Acc: 96.88%,  Val Loss:  0.38,  Val Acc: 88.42%,  Time: 0:01:52 *
Epoch [3/5]
Iter:    500,  Train Loss:  0.13,  Train Acc: 90.62%,  Val Loss:  0.37,  Val Acc: 89.57%,  Time: 0:02:19 *
Iter:    600,  Train Loss:  0.02,  Train Acc: 100.00%,  Val Loss:  0.39,  Val Acc: 89.24%,  Time: 0:02:44
Epoch [4/5]
Iter:    700,  Train Loss: 0.011,  Train Acc: 100.00%,  Val Loss:  0.38,  Val Acc: 89.62%,  Time: 0:03:11
Iter:    800,  Train Loss: 0.029,  Train Acc: 100.00%,  Val Loss:   0.4,  Val Acc: 88.97%,  Time: 0:03:38
Epoch [5/5]
Iter:    900,  Train Loss: 0.045,  Train Acc: 96.88%,  Val Loss:   0.4,  Val Acc: 89.40%,  Time: 0:04:04
Iter:   1000,  Train Loss:  0.16,  Train Acc: 96.88%,  Val Loss:  0.41,  Val Acc: 89.57%,  Time: 0:04:29
Test Loss:  0.31,  Test Acc: 91.94%
```

2.3.7 模型的测试

将测试集导入模型进行测试，测试结果如表 8

表 8 测试结果

测试精准率	测试损失值
91.94%	0.31

预测类别标签如图 9 所示：

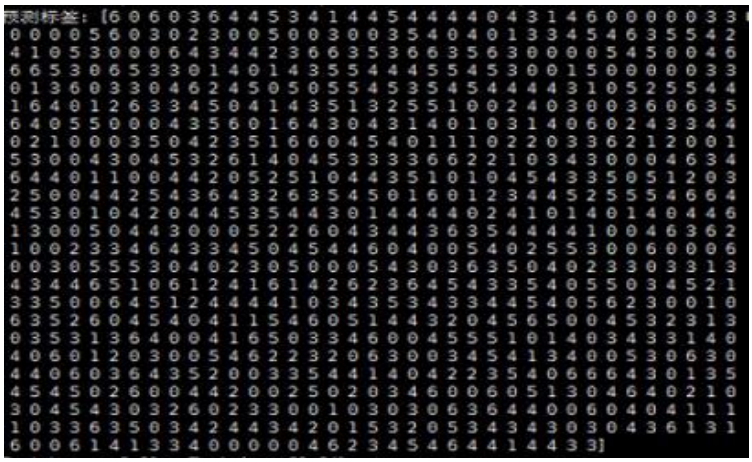


图 9 预测标签

2.4 模型评价

2.4.1 评价指标

1．F1 得分的指标

本题研究的问题属于分类问题，分类问题最常用的评价指标包括精确率 P、召回率 R 以及 F1 值，它们的计算公式需要用到混淆矩阵[21]，混淆矩阵如表 10 所示：

表 10 混淆矩阵

混淆矩阵	预测为正例	预测为反例
真实为正例	TP（真正例）	FN（假反例）
真实为反例	FP（假正例）	TN（真反例）

(1)精准率 P 是指分类器预测为正且预测正确的样本占所有预测为正的样本的比例，计算公式如下：

$$P=\frac{TP}{TP+FP} \quad (10)$$

(2)召回率 R 是指分类器预测为正且预测正确的样本占所有真实为正的样本的比例，计算公式如下：

$$R=\frac{TP}{TP+FN} \quad (11)$$

(3) F1 值是综合了 P 和 R 的一个指标，一般计算公式[22]如下：

$$F1=\frac{2*P*R}{P+R} \quad (12)$$

其中 $0\le F1\le 1$，当 P =1,R =1 时，F1 值达到最大为 1，此时精确率和召回率均达到 100%，这是最完美的情况，这种情况在实际应用中是很难实现的，P 和 R 是一对矛盾的度量，当 P 高时，R 往往会偏低；当 R 高时, P 往往偏低，因此，在使用 F1 值评估分类器性能时，其值越接近 1，说明分类器的性能越好。由于 F1 值是对 P 和 R 两个评价指标的一个统一，可以更加全面的反映分类性能，因此它是本文衡量实验效果主要评价指标。

2．混淆矩阵的指标

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。如表 11 中，就能得到如下这样一个矩阵，我们称它为混淆矩阵（Confusion Matrix）：

表 11 混淆矩阵

混淆矩阵		真实值	
		正例	反例
预测值	正例	TP	FP
	反例	FN	TN

观测值在第二、四象限对应的位置，即 TP 与 TN 的数值越多越好；反之，在第一、三象限对应位置出现的观测值肯定是越少越好。

2.4.2 评价结果

如图 10，将测试结果用 F-Score 评估得出查准率、查全率和 f1 得分如表 12，用混淆矩阵评估则得出混淆矩阵，

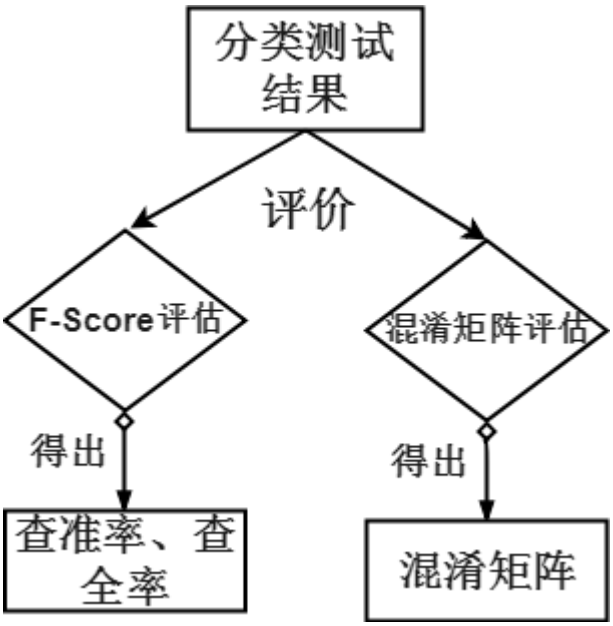


图 10 评估过程

表 12 F-Score 得分表

F1	精确率(Precision)	召回率(Recall)
91.94%	91.31%	91.98%

类别标签说明如表 13

表 13 标签说明

Chengxiangjianshe	**类 1**	**城乡建设**
Huanjingbaohu	**类 2**	**环境保护**
Jiaotonyunshu	**类 3**	**交通运输**
Jiaoyuwenti	**类 4**	**教育文体**
Laodonheshehuibaozhang	**类 5**	**劳动和社会保障**
Shangmaolvyou	**类 6**	**商贸旅游**
Weishengjisheng	**类 7**	**卫生计生**

对测试集用矩阵评估得出混淆矩阵 如表 14

表 14 测试集的混淆矩阵

混淆矩阵		预测						
		类 1	**类 2**	**类 3**	**类 4**	**类 5**	**类 6**	**类 7**
实际	**类 1**	**201**	**4**	**3**	**1**	**1**	**8**	**2**
	类 2	**5**	**71**	**0**	**2**	**0**	**2**	**0**
	类 3	**1**	**0**	**56**	**0**	**0**	**1**	**0**
	类 4	**6**	**1**	**0**	**151**	**2**	**3**	**1**

	类 5	**2**	**1**	**1**	**2**	**186**	**1**	**3**
	类 6	**8**	**1**	**3**	**3**	**0**	**99**	**1**
	类 7	**0**	**0**	**0**	**0**	**3**	**2**	**80**

2.4.3 模型对比

为了和本文提出的基于 BERT 模型的短文本分类方法作对比，本文选择了 TextCNN 模型[6]作为对照模型进行了实验，它是利用卷积神经网络对文本进行分类的算法，2014 年由 Yoon Kim 在文献[6]中提出，该模型是文本分类最常用的方法，执行效率高且分类效果较好。本文使用 TextCNN 模型进行短文本分类的训练过程中经过模型输入，embedding 层，卷积层，池化层，拼接，全连接和预测。其他的一些训练参数设计和 BERT 比较如表 15 所示：

表 15 模型参数对比

BERT 模型参数	值	TextCNN 模型参数	值
模型训练周期(epoch)	5	模型训练周期(epoch)	5
学习率(learning_rate)	2e-5	学习率(learning_rate)	2e-5
每批训练集数据大小(batch_size)	32	每批训练集数据大小(batch_size)	32
每句话处理的长(pad_size)	40	每句话处理的长度(pad_size)	50
hidden	768	卷积核数量(channels 数)	256

作为对比实验，测试集保持不变，预测时模型参数和训练时的模型参数保持一致，评价指标主要采用 F1 值，对比结果如图 11

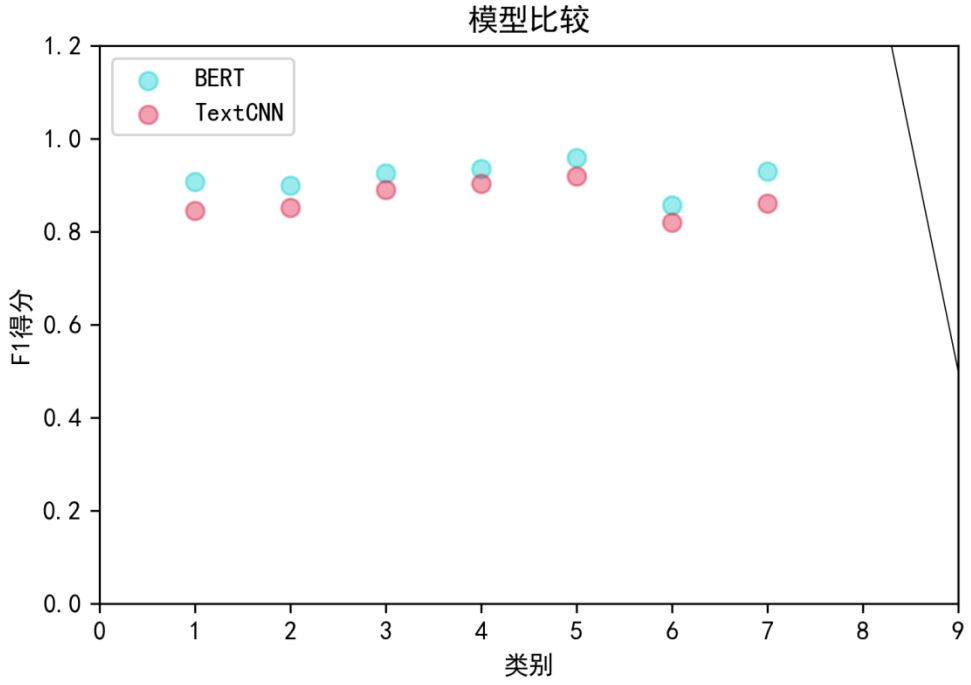


图 11 模型比较

从图中可以看出，BERT 模型在各个类别上的分类性能均比 TextCNN 模型效果好，说明本文提出的基于 BERT 模型的中文短文分类算法是可行的。

第 3 章 热点问题挖掘

3.1 问题分析

群众反映的问题分布在不同时间不同用户的留言中，而挖掘出其中的热点问题，首先需要将群众反映的同一类问题聚合起来，并且对这同一类问题建立合适的热点模型公式。计算每一类热点问题的热度，进行热度排名，并给出评价结果。

最后按表 1 的格式给出 排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

3.2 数据预处理

文本预处理是文本分类中至关重要的一步，中文分词的结果以及停用词的存在都会直接影响特征提取的结果，进而影响文本分类的效果。利用结巴库进行分词，去停用词，清洗数据，如图 12。

3.2.1 中文分词

中文分词技术是自然语言处理领域中很多关键技术的基础，包括文本分类、信息检索、信息过滤等。现如今中文分词的方法很多，大多数研究者的研究重点都是按照提升算法的精度、速度来的，常用的算法可总结如下：字典分词方法、理解分词方法和统计分词方法。本文采用业界比较知名的 Jieba 中文分词对留言文本进行分词，同时，考虑到如果直接使用留言内容作为语料库，可能会引起两个类别的留言分到同一类，即存在误分类情况。为了减少这种情况，我们根据换行符等特殊标记将整个留言内容进行分段，通过段落信息构建语料库。

3.2.2 过滤只保留名词

对于中文句子，关键信息都保留在了名词，而动词、形容词、副词等并没有保存太多信息，因此首先对所有词语进行词性过滤，只留下名词。

3.2.3 去除停用词

中文分词结果并不能直接用于特征选择，其中，某些词如果作为特征，会留言文本分类最后的结果。此时，需要引入停用词去除分词结果中无用的词，例如：“的”、“吧”、“啊”等。

预处理效果如下图 12 所示：

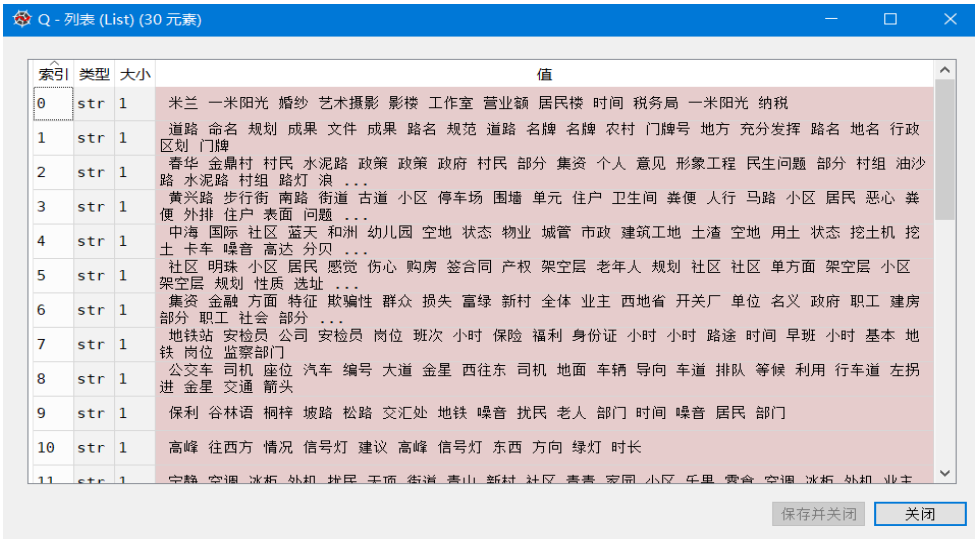


图 12 留言详情预处理效果图

3.3 模型建立

3.3.1 定义指标

同类话题贴子数：附件 3 中的相同时间，地点的同类话题贴子数

该类话题帖子总赞数：相同时间，地点，对于帖子群众的点赞总数

该话题总反对数：相同时间，地点帖子的反对数

对于热度的定义，一个热点问题如果有许多用户发帖留言反映，并且对该类热点问题点赞支持，那么就可以依此建立热度模型：

热度 = w1*该类话题帖子数 + w2*该类话题帖子总赞数 + w3*该类话题总反对数（13）

如果同一类热点问题，发帖和点赞的人越多，那么这个问题就拥有越高的热度。

图 13 热点问题挖掘模型图

3.3.2 中文分词模型

基于规则或词典的分词方法是切分词语序列的一种较为机械的分词方法，其基本思想是将词语中的字符串和字典逐个匹配，找到匹配的字符串则切分，不匹配则减去边缘的某些字符，从头再次匹配，直到匹配完毕或者没有找到词典的字符串结束，这里采用双向最大匹配法进行分词。

正向最大匹配法的思想是，假设有一个待分词文本和一个分词词典，词典中最长的字符串长度为 i，从左至右匹配待妇女此文本的前 i 个字符串，查找是否有和词典一致的字符串，若匹配失败，则删去该字符串的最后一个字符，仅留下前 i-1 个词，继续匹配余下待分词文本的字符串。如果匹配成功，则被切分下来的第二个文本序列成为新的待分词文本，重复以上操作直到匹配完毕，算法流程如图 14 所示：

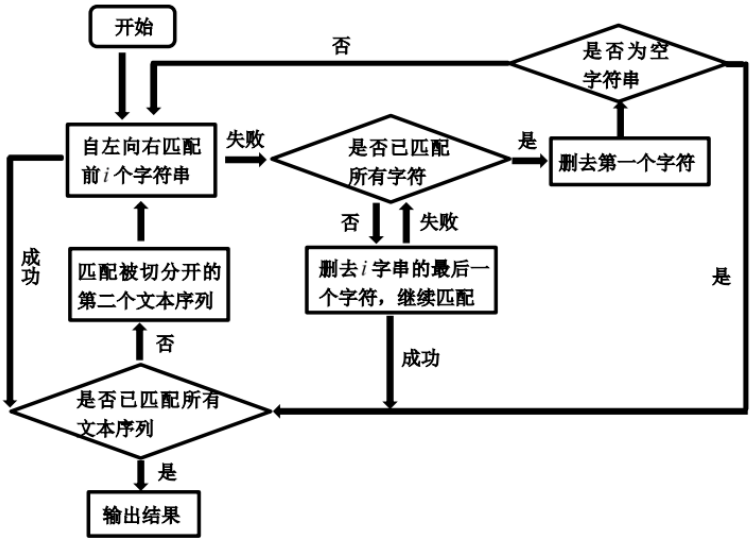


图 14 正向最大匹配法

逆向最大匹配法类似于正向最大匹配法，只不过方向相反，而双向最大匹配法的基本思想是将正向最大匹配与逆向最大匹配法的结果进行对比，根据最大匹配规则，选择两种方法中切分次数较少的，作为切分结果。

基于统计的分词语言模型，我们使用隐含马尔可夫模型，即 HMM 模型，用来描述一个含有隐含未知参数的马尔可夫过程。在马尔可夫模型中，模型的状态是可见的，而隐马尔科夫模型则是部分可见，是隐含未知参数的马尔可夫过程，HMM 描述观测变量和状态变量之间的概率关系。

利用 HMM 进行中文分词，首先设置每个字每个时刻的观测值，状态序列为标注的结果，每个时刻的状态值有四种情况：{B,M,E,S}，中文分词问题可以描述为：已知$O=o_1,o_2,....o_T$为待标注的字观测序列，$S=\{s_1,s_2,....s_N\}$为待标注的状态，求概率 P(Q|O)最大的状态序列$Q=q_1q_2....q_T$，既给定待标注的字序列，求最有可能对应的状态序列。

根据贝叶斯公式有：

$$P(Q|O)=\frac{Q(Q,O)}{P(O)}=\frac{P(O|Q)P(Q)}{P(O)} \qquad (14)$$

由于待标注的字观测序列$O=o_1,o_2,....o_T$已知，因此 P(O)为常数。只需要计算最大化情况下的$P(O|Q)P(Q)$。由 HMM 的两个基本假设：每一时刻的观测值只与对应时刻的状态值有关，每一时刻的状态值只与上一时刻的状态值有关，则有：

$$P(O|Q)=P(o_1|q_1)P(o_2|q_2)....P(o_T|q_T) \tag{15}$$

$$P(Q)=P(q_1)P(q_2|q_1)....P(q_T|q_{T-1}) \tag{16}$$

其中，$P(q_i|q_{i-1})$ 表示状态转移概率，$P(o_i|q_i)$ 表示发射概率。这些参数可以通过用语料库训练得到。这样中文分词问题可以描述为：已知 $O=o_1,o_2,....o_T$ 为待标注的字观测序列，求最有可能对应的状态序列 Q*

$$Q^*=\arg\max P(O|Q)P(Q) \tag{17}$$

利用维特比算法方法求解 $P(O|Q)P(O)$ 的最大值，得到标注的状态序列值。

一般在实际分词中，以基于规则和词典的分词方法为主，以基于统计的分词方法为辅，可以得到更好的分词效果。

3.3.3 TF-IDF 文本特征提取模型

从上一步分词并筛选重要词之后，用剩下的这些词来进行特征提取，能更好反映留言的特征。因为留言中普通的词很多，但其重要性可能都比较小，本文使用 TF-IDF 来提取特征，这种特征提取能更好的提取留言特点。通过 TF-IDF 特征提取之后，在所有的留言中，每个留言都有一个向量标识。向量上的每一个值都是一个词的 TF-IDF 值。其向量获得方式为首先统计出所有的词，把每个词当成向量的每一个维度，如果该文档中有某词，就在某词的维度上计算它的 TF-IDF 值；如果不存在某词，那么某词的维度上的值就为 0。用这种方式对所有的留言进行特征提取，提取的结果是一个稀疏矩阵。

在 TF-IDF 中，单词的重要性由两个因素共同决定，它与它在文档中出现的次数成正比，但它随着语料库中出现该词的频率越多而下降。在某一文档中，词频（TF）是指词在文档中出现的次数。TF 的值往往偏向于词汇量较大的文件，即长文件。（如果用词频来决定一个词是否重要，那么长文本中 的单词相同的频率往往会比短文本中的频率更高）。词的频率的计算可由此公式算得：

$$\text{词频(TF)}=\frac{\text{词在文档中的出现次数}}{\text{文档的总词数}} \quad （18）$$

TF-IDF 的另一指标，当然就是逆文档频率（IDF）。IDF 是衡量单词总体重要性的指标，其值等于文档总数除以包含该单词的文档数量的商再取其商的对数。词的逆文档频率的计算可由此公式算得：

$$\text{逆文档频率}(IDF)=\log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}}\right) \quad (19)$$

为了避免某词可能从来都没出现在所有的文档中，而导致被除数为零，一般分母用（包含该词的文档数+1）代替。

某词在某文档中是高词频，而在整个文档集中，该词又是低文档频数，那么该词可以得到一个较高权重的 TF-IDF 值。因此，TF-IDF 有助于降低常见的词语特征。

3.3.4 DBSCAN 密度聚类模型

特征提取完毕后，需要进行聚类，本文采用的聚类算法是 DBSCAN 聚类。

DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种基于密度的空间聚类算法。该算法将具有足够密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，它将簇定义为密度相连的点的最大集合。与 K-Means 算法不同，它不需要确定聚类的数量，而是基于数据推测聚类的数目，它能够针对任意形状产生聚类。

DBSCAN 算法需要首先确定两个参数：

1) ε -领域:在一个点周围邻近区域的半径
2) minPts:邻近区域内至少包含点的个数

根据以上两个参数，结合 epsilon-neighborhood 的特征，可以把样本中的点分成三类：

1. 核点（core point）：满足 NBHD(p,epsilon)>=minPts，则为核样本点。
2. 边缘点（border point）：NBHD(p,epsilon)<minPts，但是该点可由一些核点获得（density-reachable 或者 directly-reachable）。
3. 离群点（Outlier）：既不是核点也不是边缘点，则是不属于这一类的点。

注：边缘点 *density-reachable* 是指存在当前类中其他点作为核点所在的类中。例如，朋友的朋友（可以是 n 多个）也是朋友。如下图 15，黄圈右下角的点即为 *density-reachable*，*directly-reachable* 的点即为 NBHD 中的点。

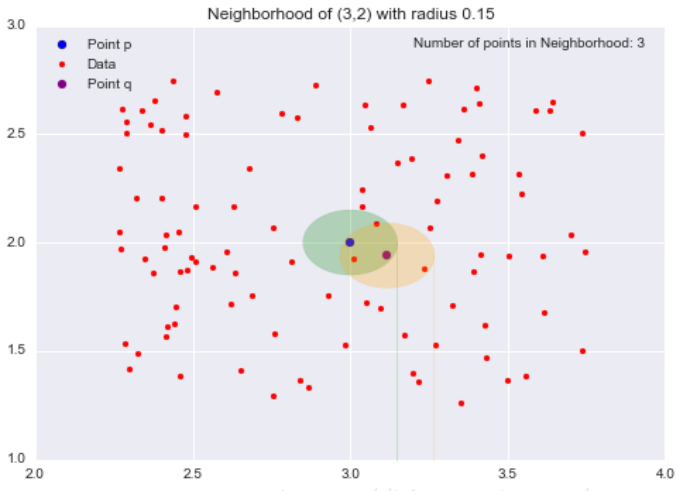


图 15 DBSCAN 聚类示意图

算法流程：

输入：文本特征矩阵数据集 $D=(\overrightarrow{x_1},\overrightarrow{x_2},.....,\overrightarrow{x_N})$

领域参数 $(\varepsilon,MinPts)$

输出：热点问题簇划分 $C=\{C_1,C_2,.....,C_K\}$。

算法步骤如下：

初始化核心对象集合为空集： $\Omega=\phi$。

寻找核心对象：遍历所有的样本点，计算密度可达距离

迭代：以任一未访问的核心对象为出发点，寻找有其密度可达的样本生成的聚类簇，直到所有核心对象都被访问为止。

3.4 模型求解

3.4.1 实验条件

Windows10 操作系统

Python 3.7

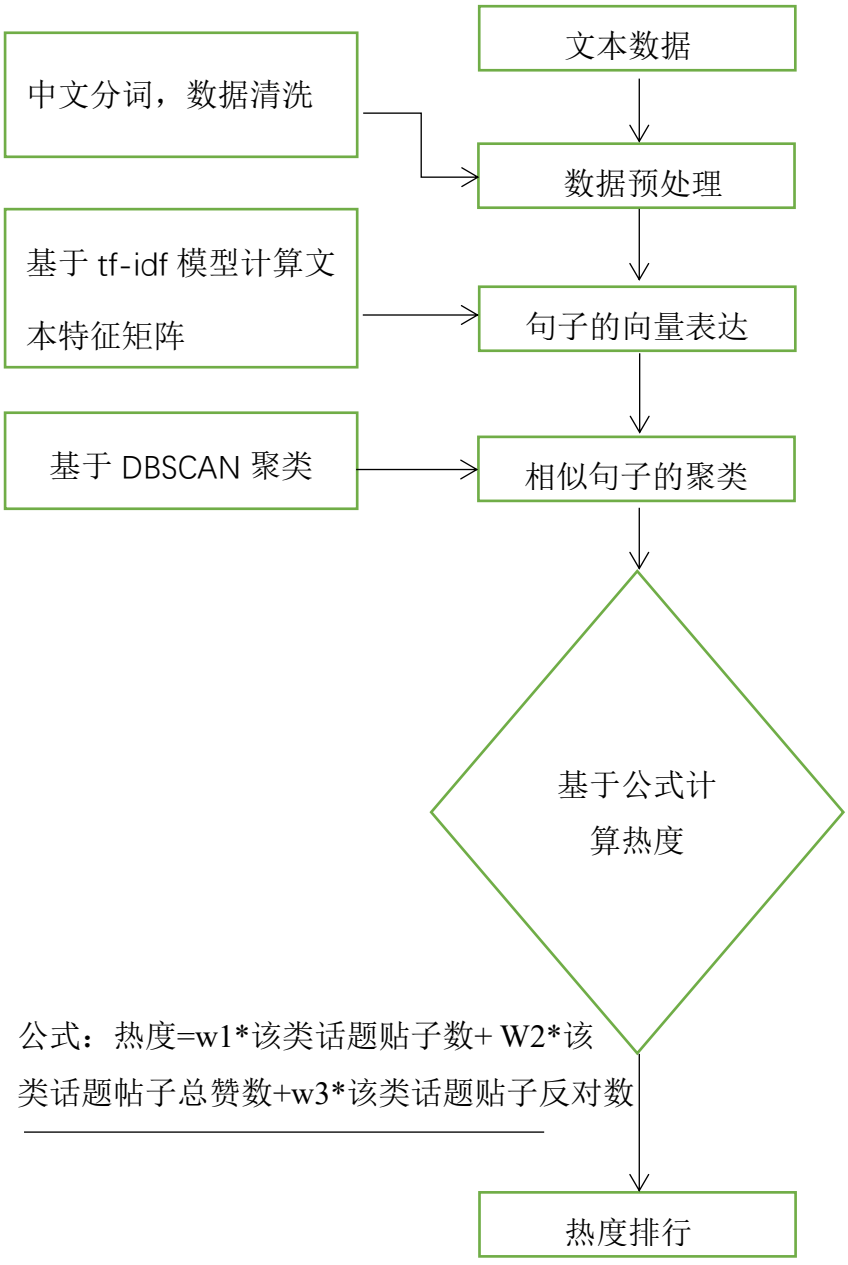
Tableau 2018.3

3.4.2 数据选择

附件 3 ：留言详情，留言时间，点赞数，反对数

3.4.3 算法流程

表 15 算法流程



3.4.5 模型训练

表 16 模型主要参数

参数	值
ε	0.5
MinPts	5

第一步：先经 3.2 进行数据预处理，去掉停用词等

第二步：再将预处理好的数据经过 3.3.3 的 TF-IDF 提取文本特征向量

第三步：再将提取到的文本特征向量放进 3.3.4 中的 DBSCAN 密度聚类模型中，划分数据类别。

第四步：通过累加，分别计算出该类话题贴子数，该类话题帖子总赞数和该类话题帖子的反对数。

第五步：分别将权值 0.5，0.25，0.25 赋给 w1,w2,w3 计算出热点问题热点指数。

3.4.6 实验结果

热度排名前 5 的热点问题如表 18 下：

表 18 前 5 的热点问题

热度排名	问题 ID	热度指数	时间范围	热点/人群	问题描述
1	1	428	2019/01/11 至 2019/07/08	A 市 A4 区 58 车贷案	西地省 A 市 58 车贷恶性退出诈骗案件，金额巨大，老板跑路，侦察不力
2	2	358	2019/04/18 至 2019/09/06	A 市 A4 区绿地海滩	A 市绿地海外滩小区会饱受附近高铁噪音困扰
3	3	110	2019/01/08 至 2020/01/06	A 市武广新城伊景园	A 市伊景园 购房捆绑销售车位，项目违规，欺诈消费者
4	4	103	2019/01/07 至 2019/12/20	A 市 A3，A7 区幼儿园	A 市 A3 区幼儿园 入学难，收费不合理，没有普惠性幼儿园
5	5	46	2019/11/02 至 2020/01/15	A 市 A2 区丽发新城	A2 区丽发新城附近建搅拌站噪音严重扰民，影响睡眠

				搅拌站	

相应热点问题对应的留言信息，保存在“热点问题留言明细表.xls”。

部分热点问题留言明细表如下表 19 所示：

表 19 部分热点问题留言明细表

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
1	214238	A00061787	请问A4区么	########	标题：恳请	1
1	218132	A00010609	再次请求这	########	尊敬的胡书	0
1	220711	A00031682	请书记关注	########	尊敬的胡书	0
1	223787	A00034861	西地省58车	########	原帖链接：	0
1	226265	A00010644	恳请A市经	########	唐局长，您	0
1	234320	A00010659	不要让A市	########	胡书记：您	0
1	240554	A00029163	A市58车贷	########	A4区经侦我	0
1	254532	A00010606	A市58车贷	########	背景：58车	0
1	272413	A00010606	西地省A市	########	西地省58车	0

3.5 模型评价

本文综合点赞数与留言出现频率对模型结果进行评价如图 16。

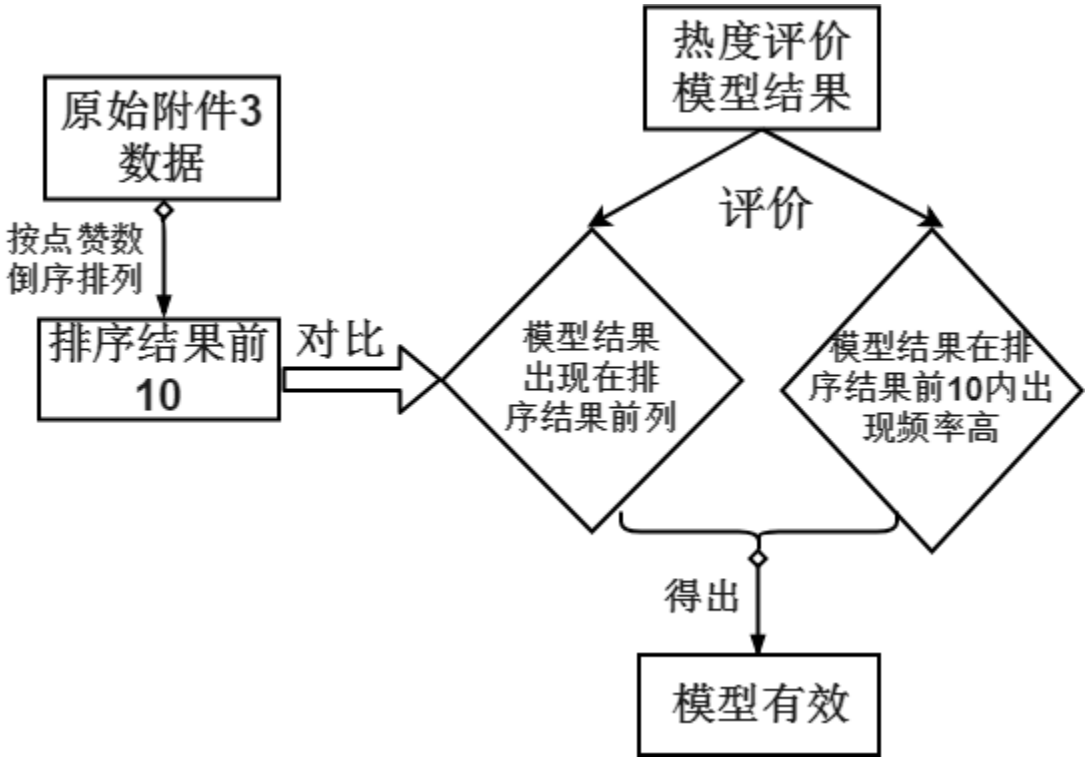


图 16 模型评价流程图

如下表 20 是对原始附件 3 数据按点赞数降序排名的结果：

表 20 降序排列的附件 3 数据

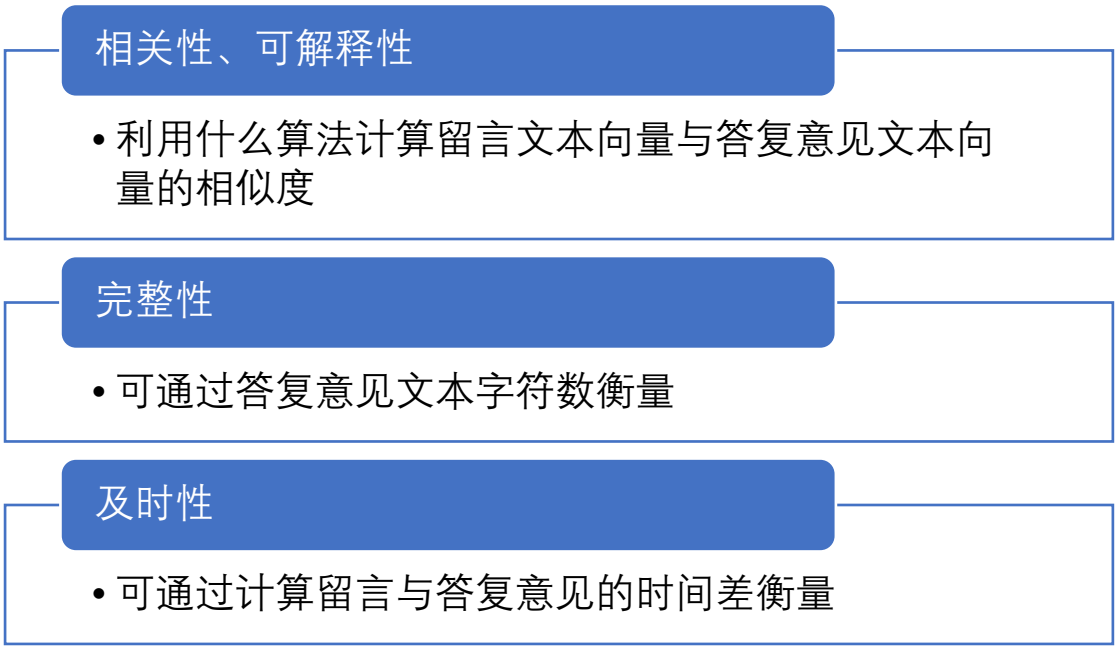
留言主题	留言时间	反对数	点赞数
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2019/8/19 11:34:04	0	2097
反映 A 市金毛湾配套入学的问题	2019/4/11 21:02:44	5	1762
请书记关注 A 市 A4 区 58 车贷案	2019/2/21 18:45:14	0	821
严惩 A 市 58 车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	0	790
承办 A 市 58 车贷案警官应跟进关注留言	2019/3/1 22:12:30	0	733
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？	2019/9/5 13:06:55	0	669
A 市富绿物业丽发新城强行断业主家水	2019/6/19 23:28:27	0	242
建议西地省尽快外迁京港澳高速城区段至远郊	2019/1/10 15:01:26	0	80
请问 A 市为什么要把和包支付作为任务而不让市场正当竞争？	2019/1/16 17:01:25	0	78
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	2019/3/26 14:33:47	0	78
A 市三一大道全线快速化改造何时启动？	2019/9/15 15:31:19	0	66
关于 A6 区月亮岛路 110kv 高压线的建议	2019/4/9 17:10:01	2	55

从点赞数和贴子数的数据来看，A 市 58 车贷案和丽发新城搅拌案等的点赞数和出现频率明显居多，验证了本题模型得出的热度问题留言前 5 的结果，而点赞排名第一的留言主题汇金路案，虽然点赞高，但是相关热点问题的帖子仅有一两个，难免不能保证是有人自己发帖子，刷点赞，不能代表广大用户，所以并未列入热点问题。模型结果理想，证实了本文所提的模型的有效性。

第 4 章 答复意见评价

4.1 问题分析

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、可解释性等角度对答复意见的质量给出一套评价方案，可分以下三点分析：



4.2 模型建立

余弦相似度：本文采用了 doc2vec 文本向量化算法将留言内容和答复内容转化为文本向量，并采用余弦相似度公式计算文本相似度。余弦相似度算法：一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值接近于 0，夹角趋于 90 度，表明两个向量越不相似。其中余弦相似度公式如下：

$$\text{余弦相似度}=\cos(\theta)=\frac{A\cdot B}{|A|\times|B|} \quad (20)$$

其中 **A** 为留言文本向量，**B** 为留言答复向量。两个向量的余弦值越大，余弦相似度就越大，文本相似度越大。

Jaccard 系数：jaccard 系数是计算留言内容和答复意见两者之间一起用到的关键词占两个文本长度的百分比。其计算公式如下

$$\text{Jaccard系数}=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (21)$$

其中 **A** 为留言文本向量，**B** 为留言答复向量。两个向量的 jaccard 系数值越大，相似度越高。答复质量越高

答复时间差：本文采用群众留言的时间和政府答复的时间的时间差作为衡量及时性的指标，其中时间差公式：

时间差=答复时间-留言时间（22）

答复文本长度：本文采用答复意见文本长度作为衡量完整性的指标：

为了消除数据的大小对模型的影响，还需要对数据进行 0~1 标准化，标准化公式如下：

$$F(X)=\frac{X-\min}{\max-\min} \quad （23）$$

最后，将以上三个指标加权和，得到质量评价方案公式如下：

评价质量 = w1*余弦相似度 + w2*jaccard 系数 + w3*答复时间差 + w4*答复文本长度（24）

4.3 模型求解

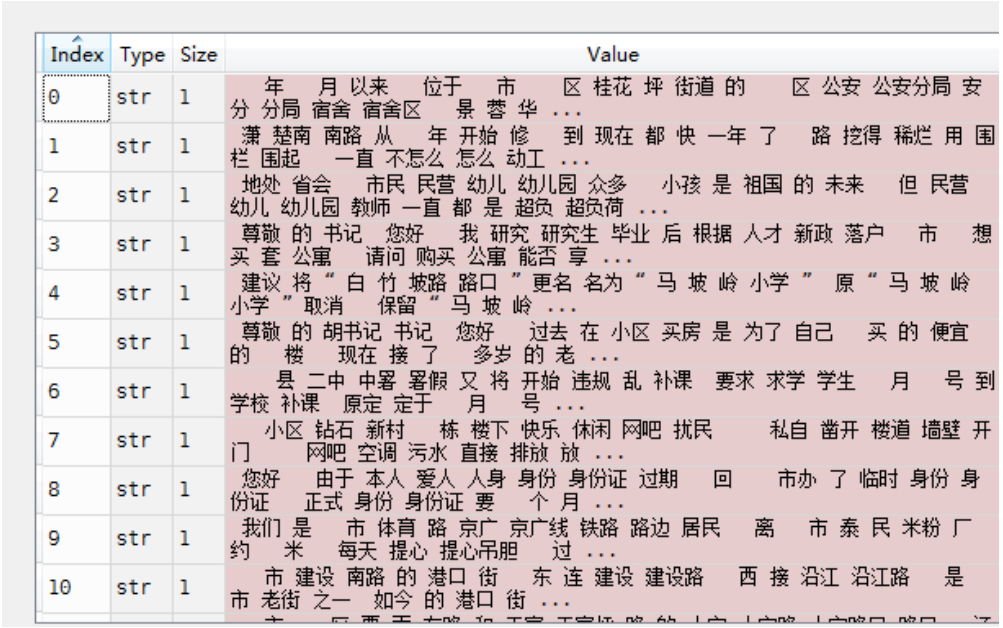
4.3.1 数据预处理

对于附件 4 中数个存在格式不同的数据，设置单元格格式为文本，再填入和其他格式相同的日期见图 17。

留言时间	留言详情	答复意见	答复时间
2019/4/25 9:32:09	物业公司却以交20万保证金，不	意收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53
2019/4/24 16:03:40	店面的生意带来很大影响，里面	，需整体换填，且换填后还有三趟雨污水管道	2019/5/9 9:49:10
2019/4/24 15:40:04	的同时更是加大了教师的工作压	民办幼儿园聘任教职工要依法签订劳动合同，依	2019/5/9 9:49:14
2019/4/24 15:07:30	落户A市，想买套公寓，请问购	年龄35周岁以下（含），首次购房后，可分别	2019/5/9 9:49:42
2019/4/23 17:03:19	“马坡岭小学”，原“马坡岭	保留“马坡岭”的问题。公交站点的设置需要	2019/5/9 9:51:30
2019/4/8 08:37:20	再把泥巴冲到右边，越是上下班	于您问题中没有说明卫生较差的具体路段，也	2019/5/9 10:02:08
2019/3/29 11:53:23	台为老社区惠民装电梯的规范性	A市A3区人民政府办公室下发了《关于A市A3区	2019/5/9 10:18:58
2018/12/31 22:21:59	跑好远，天寒地冻的跑好远，眼	装修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00
2018/12/31 9:55:00	也没得到相关准确开工信息。	任单位落实分户检查后，西地省楚江新区建设工	2019/1/16 15:29:43
2018/12/31 9:45:59	立交桥等地方做立体绿化，取得	全部分也按规划要求完成了建设，其中西边绿化	2019/1/16 15:31:05

图 17 日期数据格式预处理

对留言文本和答复文本利用结巴库进行分词，去停用词，清洗数据，见图 18。



Index	Type	Size	Value
0	str	1	年 月 以来 位于 市 区 桂花 坪 街道 的 区 公安 公安分局 安分 分局 宿舍 宿舍区 景 蓉 华 ...
1	str	1	潇 楚南 南路 从 年 开始 修 到 现在 都 快 一年 了 路 挖得 稀烂 用 围栏 围起 一直 不怎么 怎么 动工 ...
2	str	1	地处 省会 市民 民营 幼儿 幼儿园 众多 小孩 是 祖国 的 未来 但 民营 幼儿 幼儿园 教师 一直 都 是 超负 超负荷 ...
3	str	1	尊敬 的 书记 您好 我 研究 研究生 毕业 后 根据 人才 新政 落户 市 想 买 套 公寓 请问 购买 公寓 能否 享 ...
4	str	1	建议 将 “ 白 竹 坡路 路口 ” 更名 名为 “ 马 坡 岭 小学 ” 原 “ 马 坡 岭 小学 ” 取消 保留 “ 马 坡 岭 ...
5	str	1	尊敬 的 胡书记 书记 您好 过去 在 小区 买房 是 为了 自己 买 的 便宜 的 楼 现在 接 了 多岁 的 老 ...
6	str	1	县 二中 中署 署假 又 将 开始 违规 乱 补课 要求 求学 学生 月 号 到 学校 补课 原定 定于 月 号 ...
7	str	1	小区 钻石 新村 栋 楼下 快乐 休闲 网吧 扰民 私自 凿开 楼道 墙壁 开门 网吧 空调 污水 直接 排放 放 ...
8	str	1	您好 由于 本人 爱人 人身 身份 身份证 过期 回 市办 了 临时 身份 身份证 正式 身份 身份证 要 个 月 ...
9	str	1	我们 是 市 体育 路 京广 京广线 铁路 路边 居民 离 市 泰 民 米粉 厂 约 米 每天 提心 提心吊胆 过 ...
10	str	1	市 建设 南路 的 港口 街 东 连 建设 建设路 西 接 沿江 沿江路 是 市 老街 之一 如今 的 港口 街 ...

图 18 数据分词预处理

4.3.2 文本向量化

4.3.2.1 doc2vec 模型

Doc2Vec 模型有两种模型 ，分别为 Distributed Memory（DM）和 Distributed Bag-of -Words（DBOW）。DM 模型在给定上下文和文档向量的情下预测单词的概率，DBOW 模型在给定 Paragraph 向量的情况下预测文档中一组随机单词的概率。如图 19 所示，DM 模型在训练时，首先将每个文档 ID 和语料库中的所有词初始化一个 One-Hot 编码向量，然后将文档向量和上下文词语向量输入模型，投影层将这些向量累加（或取均值、直接拼接起来）得到中间向量，作为输出层的输入。输出层同样可以采用 Hierarchical Softmax 或 Negative Sampling 降低训练复杂度。在一个文档的训练过程中，文档 ID 保持不变，共享着同一个 Paragraph 向量，相当于在预测单词的概率时，利用了整个句子的语义。

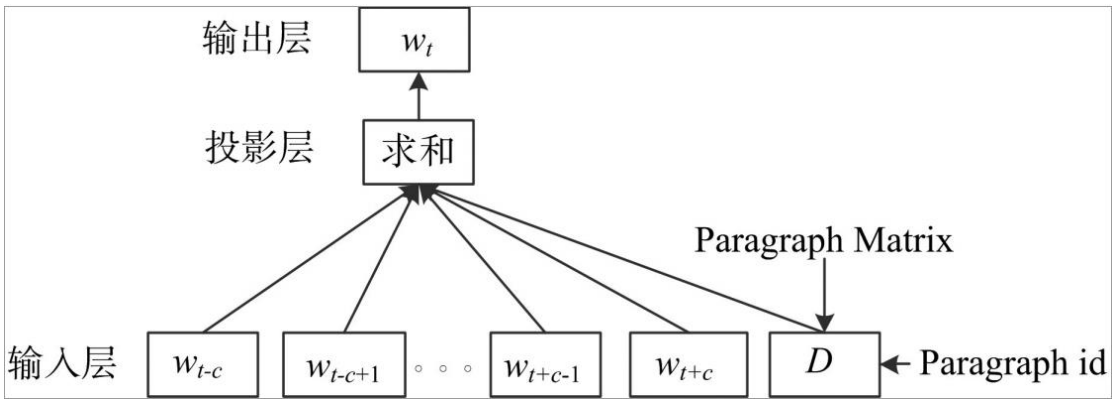


图 19

4.3.2.2 余弦相似度计算

文本语义相似度计算就是比较留言文本的相似程度和答复文本一致性程度. 前面通过 Doc2vec 模型提取特征向量，计算文本相似度就是计算文本特征向量的相似性，本文采用余弦距离来度量向量之间的相似性。余弦相似度利用向量间夹角的余弦值来衡量个体之间差异的大小，注重向量之间方向上的差异。

假设有留言文本 A 和答复文本 B，其对应的特征向量分别为：

$$\begin{aligned} X_A &= (a_1, a_2, ..., a_{n-m}) \\ X_B &= (b_1, b_2, ..., b_{n-m}) \end{aligned} \tag{25}$$

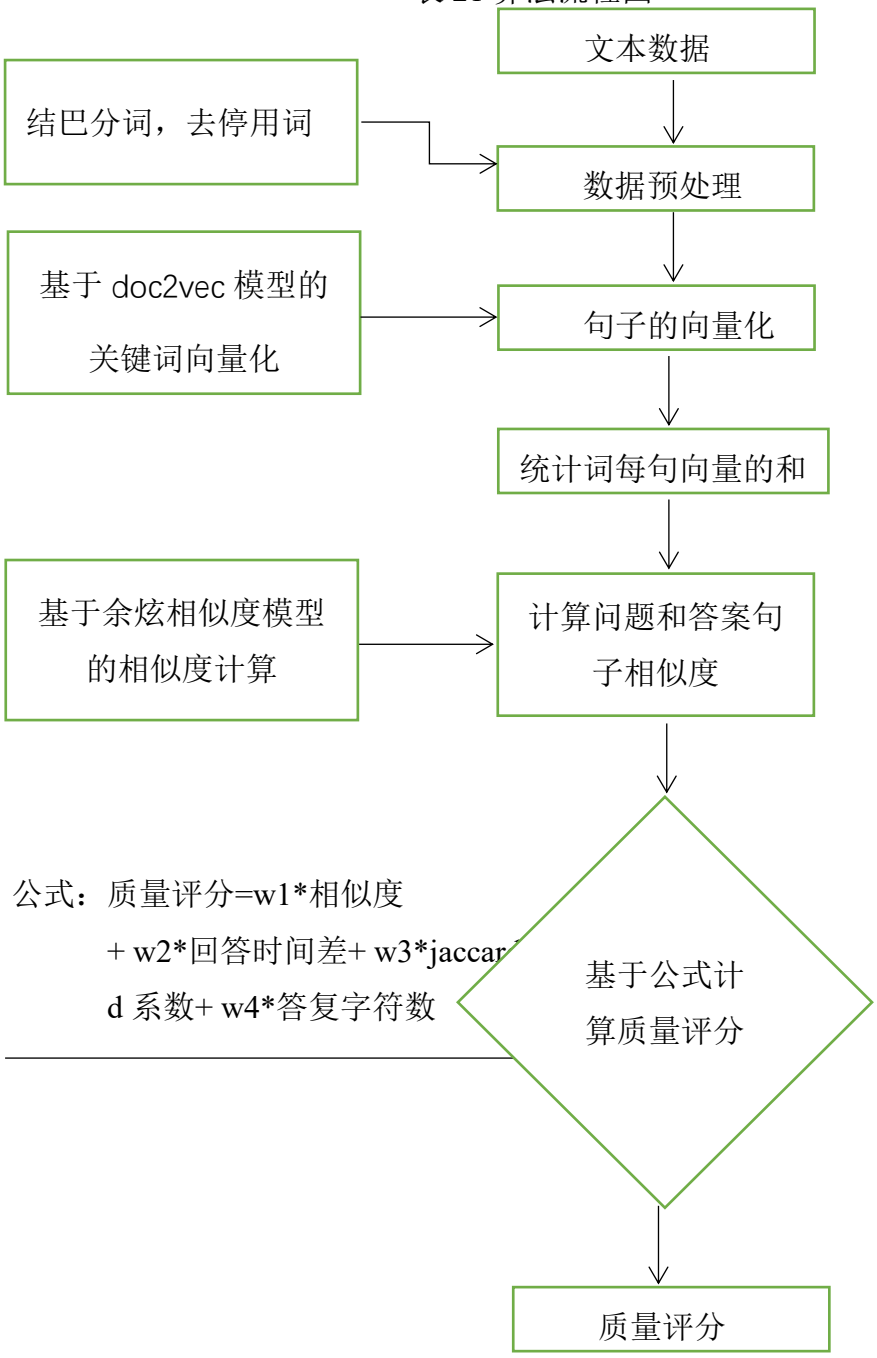
则可以定义留言文本 A 和答复文本 B 的相似度计算方法如公式（26）所示：

$$\begin{aligned} &\mathrm{sim}(A,B) = \cos\left\langle X_A, X_B \right\rangle = \\ &\frac{a_1 b_1 + a_2 b_2 + \cdots + a_{n-m} b_{n-m}}{\sqrt{a_1^2 + a_2^2 + \cdots + a_{n-m}^2} \bullet \sqrt{b_1^2 + b_2^2 + \cdots + b_{n-m}^2}} \end{aligned} \tag{26}$$

余弦相似度的值落在区间[-1，1]内，值越大则相似度越高. 当留言文本 A 和答复文本 B 的相似度值等于 1 时，两个文本内容完全重合，说明他们的相似度最高. 当文本语义的相似度值等于 0 时，则表明他们的相似度最低.。余弦相似度可以衡量留言文本和答复文本语义多大程度的重合。

4.3.3 算法流程

表 21 算法流程图



如上图，数据经过预处理，再通过 4.4.2.1 中的 doc2vec 模型，转化成像[0,0,0,1,0...0]这样的向量，再结合余弦相似度公式，计算出余弦相似度，改相似度代替文本相似度。

4.3.2 数据选择

对于附件 4 中数个存在格式不同的数据，设置单元格格式为文本，再填入和其他格式相同的日期。

4.3.3 模型训练步骤

第一步：将数据放入 4.3.1 中的数据预处理，去停用词等。

第二步：导入 4.3.2.1 中预训练好的 doc2vec 模型，做文本向量化，将全部文本转化为词向量。利用 4.3.2.2 中的文本语义相似度计算余弦相似度

第三步：计算 jaccard 系数，使用 jaccard 系数公式计算 jaccard 系数。

第四步：计算时间差和答复文本长度，使用 jiaba 分词计算答复文本长度，使用 python 中的 datetime 模块，用答复时间减去留言时间算出时间差。

第五步：消除量纲的影响，对文本长度，时间差进行 0~1 标准化。将数据转化到 0 到 1 之间。

第六步；将上述指标加权和，得出答复意见评价评分，其频数分布如下图。

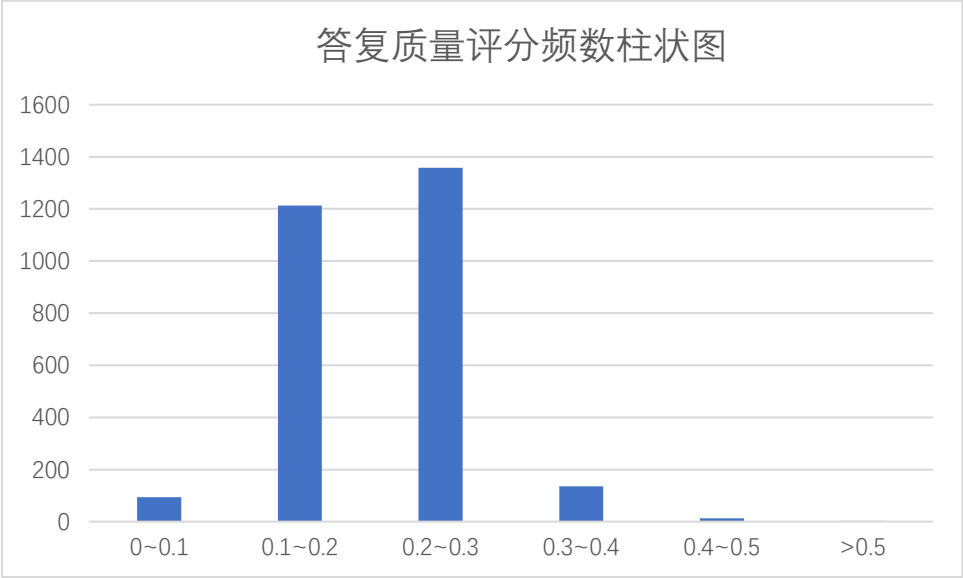


图 20 答复质量评分频数柱状图

最后设置 e 值，将大于 0.2 设置为高质量答复意见，低于 0.1 设置为地址了答复意见，0.1 到 0.2 设置为中等答复意见。

4.4 实验结果

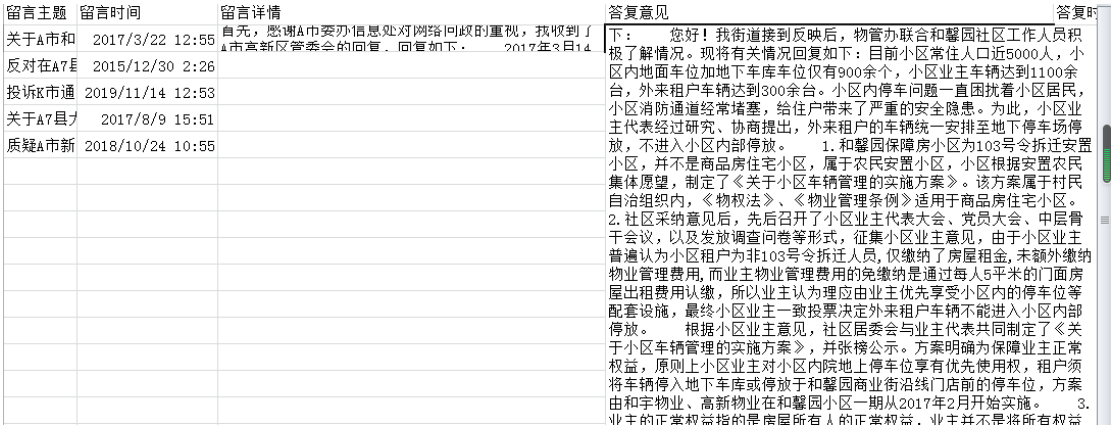
下表 22 是模型求解得到的对各留言答复质量得分部分截图如下：

表 22 留言答复质量得分表

留言编号	留言主题	留言用户	答复质量
2549	A2 区景蓉华苑物业管理有问题	A00045581	0.2460
2554	A3 区潇楚南路洋湖段怎么还没修好?	A00023583	0.1449
2555	请加快提高 A 市民营幼儿园老师的待遇	A00031618	0.2131
2557	在 A 市买公寓能享受人才新政购房补贴吗?	A000110735	0.2275
2574	关于 A 市公交站点名称变更的建议	A0009233	0.2573
2759	A3 区含浦镇马路卫生很差	A00077538	0.1069
2849	A3 区教师村小区盼望早日安装电梯	A000100804	0.2308
3681	反映 A5 区东澜湾社区居民的集体民生诉求	UU00812	0.1620
3683	反映 A 市美麓阳光住宅楼无故停工以及质量问题	UU008792	0.2770
3684	反映 A 市洋湖新城和顺路洋湖壹号小区路段公共绿化带的问题	UU008687	0.2138
3685	反映 A2 区大托街道大托新村违建问题	UU0082204	0.2136
3692	A5 区鄱阳村 D 区安置房人防工程的咨询	UU008829	0.2413
3700	A4 区万国城小区段请求修建一座人行天桥或者地下通道	UU00877	0.1750
3704	举报 A 市芒果金融平台涉嫌诈骗	UU0081480	0.1281
3713	建议增开 A 市 261 路公交车	UU0081227	0.0921
3720	关于 A 市新开铺路与披塘路交叉路口通行安全问题的建议	UU008444	0.2832

4.5 模型评价

为验证模型的实用性，本文从模型运行得到的答复质量得分中取出得分最高的五个如下图 21，从答复的相关性、完整性、可解释性等角度可以看出，答复比较专业，有针对性，完整性，及时性，从而验证本文提出的质量评价模型有效。



留言主题	留言时间	留言详情	答复意见	答复时
关于A市和	2017/3/22 12:55	首先，感谢A市委办信息处对网络问政的重视，我收到了A市高新区管委会的回复，回复如下：2017年3月14	下：　您好！我街道接到反映后，物管办联合和馨园社区工作人员积极了解情况。现将有关情况回复如下：目前小区常住人口近5000人，小区内地面车位加地下车库车位仅有900余个，小区业主车辆达到1100余台，外来租户车辆达到300余台。小区内停车问题一直困扰着小区居民，小区消防通道经常堵塞，给住户带来了严重的安全隐患。为此，小区业主代表经过研究、协商提出，外来租户的车辆统一安排至地下停车场停放，不进入小区内部停放。　1.和馨园保障房小区为103号令拆迁安置小区，并不是商品房住宅小区，属于农民安置小区，小区根据安置农民集体原望，制定了《关于小区车辆管理的实施方案》。该方案属于村民自治组织内，《物权法》、《物业管理条例》适用于商品房住宅小区。2.社区采纳意见后，先后召开了小区业主代表大会、党员大会、中层骨干会议，以及发放调查问卷等形式，征集小区业主意见，由于小区业主普遍认为小区租户为非103号令拆迁人员,仅缴纳了房屋租金,未额外缴纳物业管理费用,而业主物业管理费用的免缴纳是通过每人5平米的门面房屋出租费用认缴，所以业主认为理应由业主优先享受小区内的停车位等配套设施，最终小区业主一致投票决定外来租户车辆不能进入小区内部停放。　根据小区业主意见，社区居委会与业主代表共同制定了《关于小区车辆管理的实施方案》，并张榜公示。方案明确为保障业主正常权益，原则上小区业主对小区内院地上停车位享有优先使用权，租户须将车辆停入地下车库或停放于和馨园商业街沿线门店前的停车位，方案由和宇物业、高新物业在和馨园小区一期从2017年2月开始实施。　3.业主的正常权益指的是房屋所有人的正常权益，业主并不是将所有权益	
反对在A7县	2015/12/30 2:26			
投诉K市通	2019/11/14 12:53			
关于A7县力	2017/8/9 15:51			
质疑A市新	2018/10/24 10:55			

图 21 答复质量得分最高五个

第 5 章 模型的改进与局限性

5.1 模型的改进

本文在解决中文短文本分类的问题中，使用 BERT 模型替换常用的 wod2vec 模型进行短文本的向量表示，提出了一种基于 BERT 模型的中文短文本分类算法，并与传统的 TextCNN 模型进行对比，实验结果表明，Bert 模型在政务留言的多类别分类上具有很好的效果，在劳动和社会保障这一类别上的预测 F1 值最高可达 95.88%，并且随着数据量的增加，BERT 模型在测试集上的分类效果整体 F1 值最高可达 91.94%，高出 TextCNN 模型约 5%，说明 BERT 模型在中文短文本的语义表示上可以达到很好的效果，在一定程度上提升了中文短文本分类的效果。本文方法说明了 BERT 模型在句子层面上的向量表示有很好的表示效果，对其他类似的处理对象为句子级别的自然语言处理下游任务具有一定的参考价值。

5.2 模型的局限性

该算法存在的问题是当短文本中含有较多表情符号或者其他具有重要表征的符号时，不能很好的分类，还有待进一步研究。下一步的研究方向是在该算法

的句子表征上再融入表情符号或其他符号的位置信息用来丰富短文本的句子向量特征表示，以提高短文本的分类效果。

5.3 总结

本文提出利用 BERT 作为语言特征提取与表示方法，既能获取留言文本的丰富的语法、语义特征，又能解决传统基于神经网络结构的语言特征表示方法忽略词语多义性的问题。BERT 模型在测试集上的分类效果整体 F1 值最高可达 91.94%，高出 TextCNN 模型约 5%，说明 BERT 模型在中文短文本的语义表示上可以达到很好的效果，在一定程度上提升了中文短文本分类的效果。本文方法说明了 BERT 模型在句子层面上的向量表示有很好的表示效果，对其他类似的处理对象为句子级别的自然语言处理下游任务具有一定的参考价值。

参考文献

[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[2] Vaswani A, Shazeer N, Parmar N. et al. Attention is all you need[J]. arXiv:1706.03762v5, 2017.

[3] Yang Sen. Application research of credit scoring model for small and micro enterprises based on Softmax regression [D]. Master's degree thesis of Suzhou University, 2017. 杨森.基于 Softmax 回归的小微企业信用评分模 型应用研究[D].苏州大学硕士学位论文,2017.

[4] Li Ran. Research on Short Text Emotional Tendency Based on Deep Learning [D]. Master Degree Thesis of Beijing Institute of Technology, 2015. 李然.基于深度学习的短文本情感倾向性研究 [D].北京理工大学硕士学位论文,2015.

[5] Fu Peng, Yao Jiangang, Gong Lei. Identification of Insulator Pollution Level Using Infrared Features and Softmax Regression [J]. Computer Engineering and Applications, 2015, 51 (13): 181-185. 付鹏,姚建刚,龚磊.利用红外特征和 Softmax 回归识别绝缘子污秽等级[J].计算机工程与应 用,2015,51(13):181-185.

[6] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv:1408.5882v2, 2014.