智慧政务中的文本挖掘应用

摘 要

本文通过自然语言处理和文本挖掘技术，进行了留言文本分类，热点问题挖掘及构建答复意见评价体系，深刻了解了政府部门在‘智慧政务’中的业务需求，具有很强的实际意义。

针对问题一，首先对留言文本进行分词、去停用词等预处理，通过对各类别数据的分布进行可视化操作，我们发现样本数据不均衡的问题，于是采用**回译技术**对所给数据进行**数据增强**，最终各类数据均达到 2000 条左右。考虑到数据集的规模，我们尽可能用更多的数据进行训练，因此将数据划分为训练数据和验证数据，比例为 9：1。接下来使用 **Word2Vec** 对留言文本进行特征表示，训练后得到 300 维的词向量。随后我们构建了基于 **Attention-BiLSTM** 的留言文本分类模型，注意力机制加强了模型对长本文的记忆能力，双向长短时记忆网络可以更好的对特征进行提取和表示。调节参数后，最终在验证数据上 F-score 达 93.75%。

针对问题二，首先同样对留言数据做分词、去停用词等预处理操作。使用 **DF 特征选择**方法去除文档频率过低或过高的词语以降低维度和减少计算量，再用 **TF-IDF** 方法将文本数据进行向量化表示转换为权值矩阵，从而完成了数据的前置准备工作。然后将向量化表示的数据放入基于余弦相似度的 **K 均值聚类模型**中，识别出相似的留言并将其归入不同的类别。根据每类留言的个数、类中留言赞成数和反对数定义合理的热度评价指标，结合数据情况根据热度公式计算出每类留言的热度，深入挖掘出特定地点或特定人群的热点问题。最后选择热度排名前五的留言计算时间范围、识别地点人群，保存热点问题表和对应的热点问题留言明细表。

针对问题三，我们将答复意见的评价分为五个等级，分别为‘非常满意’、‘满意’、‘一般’、‘不满意’、‘非常不满意’。首先观察附件 4 中的答复意见，我们将影响答复意见的因素总结为四个主要方面：**相关性**、**及时性**、**完整性**、**可解释性**。计算留言和答复的**文本相似度**对相关性进行量化，利用答复时间和留言时间的时间差对及时性进行量化，接着通过答复意见是否满足规范量化了完整性，最后根据是否给出法律层面上的依据量化可解释性。随后建立基于**模糊综合评价模型**的留言评价体系，根据四个因素的重要性给出对应权重，利用上述量化方法得到模糊评价判断矩阵，最后求解模糊向量，给出评价等级。

关键字：回译　Word2Vec　余弦相似度　TF-IDF　BiLSTM　模糊综合评价法

一、引言

1.1 挖掘背景

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。政府方面对自然语言处理技术的相关需求逐渐增加，将自然语言处理技术应用到政府工作中也带来了极大的挑战。

群众留言文本数据以往主要依靠人工进行留言划分和热点整理。随着如今文本数据量的攀升，相关政府部门的工作量大大增加，并且大数据、云计算、人工智能等技术不断发展，基于自然语言处理技术建立智慧政务系统符合社会治理创新发展的新趋势。

如何使用自然语言处理技术对群众留言进行留言分类和热点问题挖掘是政府方面急需解决的问题，这对提升政府的管理水平和施政效率具有重要意义。

1.2 挖掘目标

智慧政务文本挖掘需要解决三个问题包括留言分类、热点问题挖掘和答复意见的评价。附件 2 提供了用户留言数据和对应的分类标签，问题 1 的挖掘目标是使用以往历史数据训练模型，建立高度可行的分类器，提高分类泛化能力。当传入新的留言数据时，分类器能够快速准确地将留言划分到正确的类别中。

附件 3 提供了用户的留言数据以及该留言数据获得的赞成数和反对数，问题 2 的挖掘目标是使用聚类算法对一定时间段内能够反映特定地点或特定人群问题的留言归类，并且定义合理的热度指标评价各类别问题的热度。

附件 4 提供了用户留言数据和相关部门对留言的答复意见，问题 3 的挖掘目标是量化留言答复的相关性，完整性，可解释性和即时性，从这四个角度对留言答复意见的质量做出综合评价，建立一套留言答复意见质量的评价方案。

二、基本假设与符号约定

2.1 基本假设

1. 假设题目数据准确，不存在误差

2. 假设所有留言的分类仅在所给类别中

3. 假设留言真实、可靠

2.2 符号约定

关键符号	符号说明	关键符号	符号说明
C_k	第k类	n_k	第k类留言个数
o_i	第i条留言点赞数	s_i	第i条留言反对数
a_i	第i个因素的权重	r_{ij}	被评对象隶属度的
lr	学习率	epoch	迭代次数
L	词最大长度	N	网络层数

三、问题分析

3.1 问题一分析

前述已经给出了问题 1 的挖掘目标，要实现该目标需要具体分析处理相应的数据。分析附件 2 数据中的留言详情和一级分类，使用模型对留言数据做去重、分词等预处理操作，然后使用 Word2Vec 模型将文本数据转化为可供计算机识别的词向量。以上工作完成之后，我们建立基于 Attention-BiLSTM 的留言文本分类模型，解决群众留言的多分类问题。

3.2 问题二分析

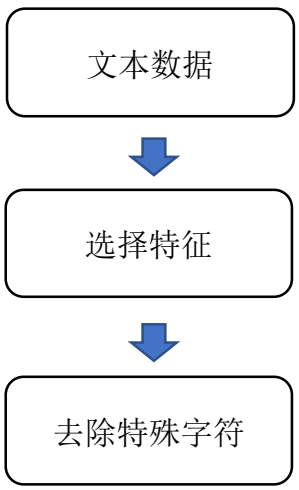
问题 2 要求我们对附件 3 中的相似留言问题聚类，并且定义合理指标评价各类的热度，从中挖掘出热点问题。附件 3 提供的数据与附件 2 类型大致相同，其中多了留言赞成和反对数，这有利于定义合理的热度评价指标。相似留言聚类同样需要对文本数据预处理，然后转换为计算机可以识别的数据类型。我们首先使用 DF 特征选择方法选取文本特征，然后使用 TF-IDF 方法进行文本向量化表示，最后使用基于余弦相似度的 KMeans 算法进行文本聚类，定义合理的热度评价指标从而挖掘出热点问题。

3.3 问题三分析

问题 3 要求我们建立留言答复意见的评价体系，分析数据考虑量化用户留言与答复意见之间的相关性、完整性、可解释性和及时性，这里我们使用文本向量之间的相似度来量化答复相关性，答复意见首尾是否含有敬称和日期来度量完整性，答复意见中是否含有法律条文来度量可解释性，以标准化后的答复时间和留言时间差量化答复的及时性，最后综合四个角度的量化值赋予合适权值计算出量化后的答复意见质量。

四、数据预处理

在进行留言文本特征表示之前，我们首先要对原始文本数据进行文本预处理，通过过滤无关噪声来得到最佳文本特征。下面以处理附件二的数据为例，展示数据预处理流程，如下图所示。



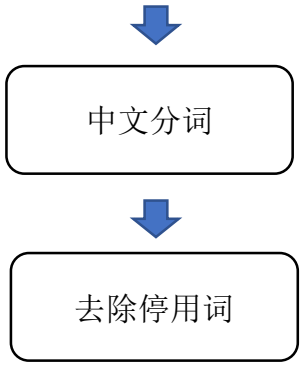


图 4-1 文本数据预处理流程图

a. 选择特征

观察附件 2 中的数据可以发现，可选特征包括：留言编号、留言用户、留言主题、留言主题、留言详情，题目要求建立关于留言内容的一级标签分类模型，所以我们选择留言主题与留言详情作为分类模型的特征。

b. 去除特殊字符

针对所选特征，我们发现文本中存在大量字母、数字及各种符号等，这些对于分类模型的输入特征而言是冗余的，因此我们仅保留汉字，去除数据中非文本部分。

c. 中文分词

中文分词是文本分类模型的基础，将基于连续字序列的文本数据按照一定的规范重新组合成词序列，以便下文通过 Word2Vec 模型将文本数据向量化。

d. 去除停用词

停用词指的是在每条文本数据出现的频率都很高，或对于输入特征的分析没有实际意义的词。比如附件 2 中的“的”、“是”、“但是”等，我们将文本数据匹配停用词表来去除停用词。

经过文本数据预处理后，结果示例如下。

```
0        大道 西行 便道 未管 路口 加油站 路段 人行道 路灯 杆 圈 西湖 建筑 集团 燕子 山...
1        书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 护栏 围着 占用...
2        区苑 火炬 路 物业 市程明 物业管理 有限公司 未经 小区业主 同意 业主 公摊 公共 面...
3        区区 华庭 高层 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 水是 日常生活 必不...
4        区区 华庭 高层 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉味 水是 日常生活 必不...
                                    ...
13395    市县 大船 坳 乡 大船 坳 村 村民 医疗保险 实施 农村 带来 效益 科长 农村 医疗保...
13396    市县 大树 鞍部 乡 大树 鞍部 村 村民 行 医保 农村 带来 绝大 光顾 段 农村 治保...
13397    市县 高 冈村 居民 实施 医疗保险 农村 带来 很大 损失 农村 医疗保险 要交 钱据 文...
13398    鄂市 鄂 县乡镇 卫生院 财政 体制 混乱 院长 公款 吃喝玩乐 搬家 贫困 人民政府 拨出...
13399    市县 乡镇 医院 财政 制上 太 混乱 院长 公款 吃喝玩乐 动不东 千上万 可怜 人民政府...
```

图 4-2 文本数据预处理结果

五、模型的建立与求解

5.1 问题 1 的建模与求解

5.1.1 数据可视化及数据增强

1.数据可视化

分析附件 2，我们得到数据含 9210 条，分为七大类：城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生，通过 Python3.7 对各类一级标签的数目进行可视化如下图所示：

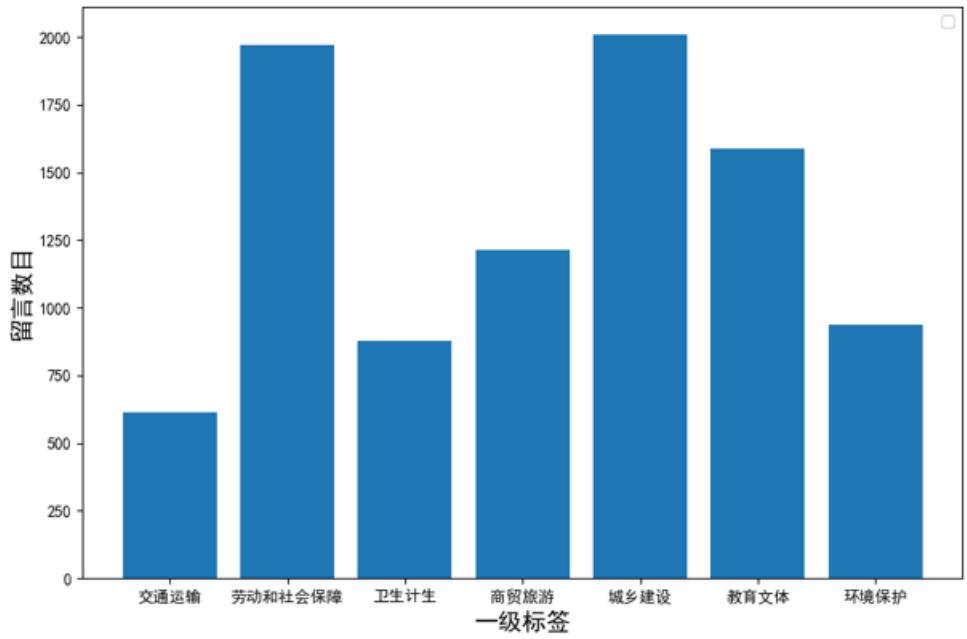


图 5-1 各类一级标签数目分布情况

通过观察上图各类一级标签数目分布情况，我们发现各类数据数目相差较大，存在**样本不均衡**的问题，这样的数据分布并不利于文本分类任务，因此下面我们对现有数据进行处理。

2. 基于回译技术的数据增强

对于数据不平衡问题，处理方法有欠采样、过采样等。本题中数据量不大，如果进行欠采样操作，可能会出现数据量太小，而导致训练不充分、模型欠拟合的问题。因此我们采用过采样，即对小类数据的样本进行采样来增加小类样本数据的数目。

针对文本数据，可使用的数据增强方法如下：

a.同义词替换：从句子中随机选取 n 个不属于停用词集的单词，并随机选择其同义词替换它们；

b.随机插入：随机的找出句中某个不属于停用词集的词，并求出其随机的同

义词，将该同义词插入句子的一个随机位置。重复 n 次；

c.随机删除：以给定概率，随机的移除句中的每个单词；

d.随机交换：随机的选择句中两个单词并交换它们的位置。重复 n 次；

e.回译：将句子译成另一种语言，再将结果译回原始语言。

本文经过实验发现，回译的效果远好于其他四种方法，所以采用回译来增强数据。翻译接口采用百度 API，选择语种英语、粤语、韩语，回译效果对比如下图所示：

当地排污单位、村、居委会代表及自愿者，必要时媒体参加。每月上、中、下旬召开联系会议，把排污单位的自行检测的原始记录放在阳光下，晒晒，与会者在检测原始记录上签字，并当场作出共商结论。这样既能能监督环保是否作为，也能看出设计及施工是否合理、排污单位是否理责和整改，这样能解决越治越污的罪恶现象。

图 5-2 原始未处理文本

当地排污单位村民居委会代表、志愿者，必要时媒体参加。每月一日、中日、最后十日召开联系会，将排污单位自查原始记录放在阳光下，与会人员在自查原始记录上签字，当场达成共识。这样，既可以监督环保工作是否做好，也可以看设计、施工是否合理，排污单位是否负责、整改，从而解决越来越多污染治理的恶果。

图 5-3 回译处理后的文本

附件 2 中最大数目类别城乡建设有 2009 条，故将其他类别数目均扩充至 2000 条左右，数据增强后的效果如下图所示：

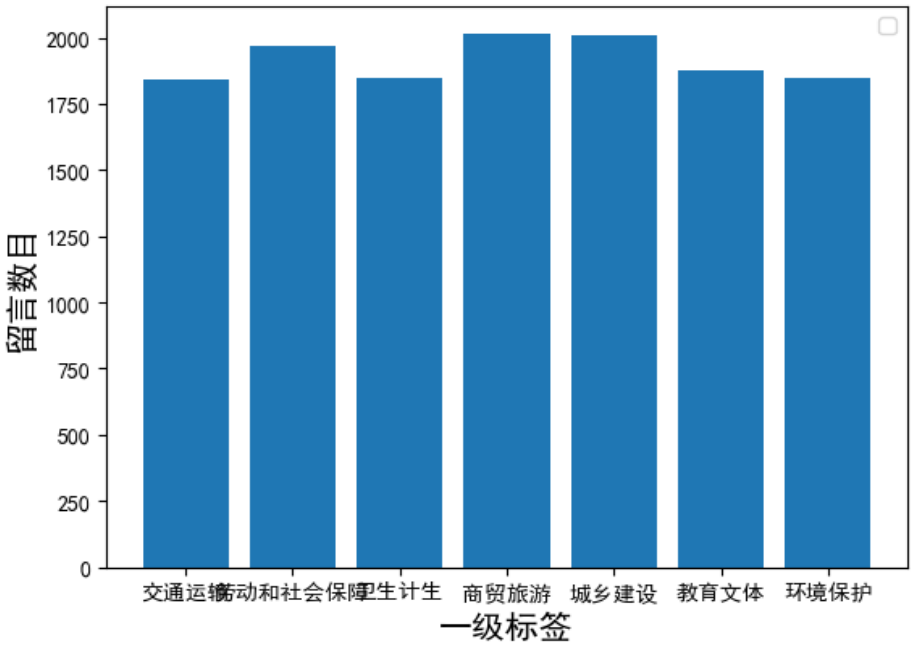


图 5-4 处理后的各类一级标签数目分布情况

5.1.2 基于 Word2Vec 模型的留言文本特征表示

Word2Vec 是一种静态的词嵌入表示，词嵌入是自然语言处理中语言模型与表征学习技术的统称。从概念上讲，它是一种将文本中的词转换成数字向量的方法，指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。Word2Vec 模型用深度学习网络对语料数据的词语及其上下文的语义关系进行建模，以求得到低维度的词向量。该词向量一般在 100–300 维左右，能很好的解决传统向量空间模型高维稀疏的问题。

Word2Vec 包括两个模型，分别为 CBOW 和 Skip-gram 模型。两个模型的目标相同，均为判断某个语句是否为自然语言以及获得每个单词对应的词向量。CBOW 模型是在已知上下文 Context(t)的情况下预测当前词 t，而 Skip-gram 模型则是在已知当前词 t 的情况下预测其上下文词 Context(t)。这两个模型都包括输入层、隐藏层和输出层,结构如下图所示。

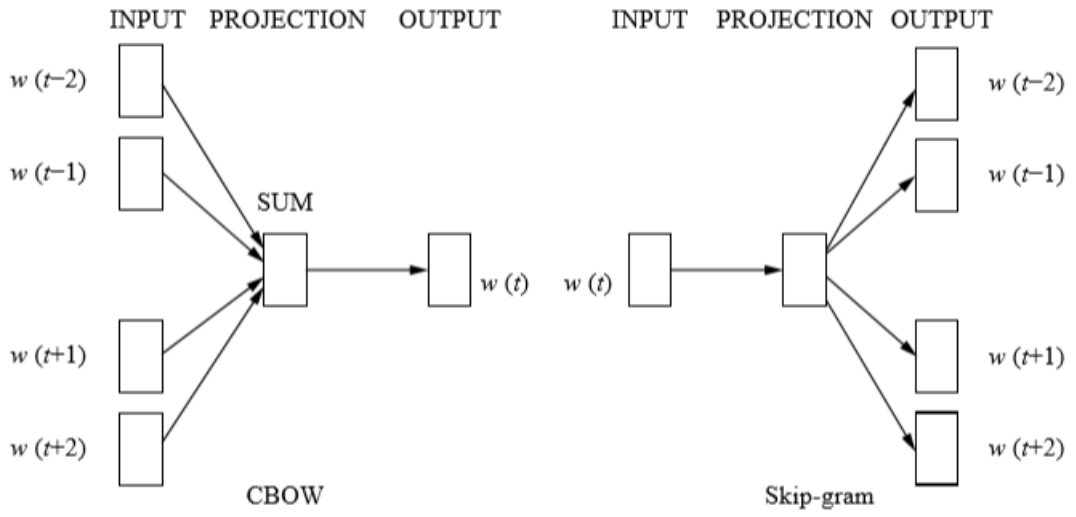


图 5-5 Word2Vec 的 CBOW 和 Skip-gram 模型

本题使用预训练的中文词向量模型，根据预训练模型生成词典，从而得到文本数据的分布式表示如下图所示。

```
In [13]:  1  word_vectors['医生']
executed in 6ms, finished 19:06:15 2020-04-06

Out[13]: array([ 7.381360e-01,  1.761550e-01,  2.110120e-01,  3.750230e-01,
                 2.881380e-01,  7.329220e-01, -5.140290e-01,  6.843080e-01,
                -4.056040e-01, -4.840080e-01, -6.795420e-01,  4.618630e-01,
                -2.259750e-01,  3.422860e-01, -1.414120e-01,  5.381890e-01,
                 5.577420e-01, -3.232490e-01, -8.199300e-02,  5.122570e-01,
                 8.750800e-02,  3.031560e-01,  2.484870e-01, -8.861310e-01,
                -3.346950e-01, -5.240210e-01,  2.650820e-01,  3.419670e-01,
                -4.094690e-01,  1.031600e-02,  2.722370e-01,  1.804770e-01,
                -2.810050e-01,  9.165400e-02, -1.258600e+00,  5.794050e-01,
                 5.346840e-01, -2.980400e-02, -5.730420e-01,  2.420260e-01,
                 3.083350e-01, -2.841290e-01, -2.824650e-01,  1.461150e-01,
                -4.920100e-01, -2.740320e-01,  5.098180e-01,  2.102600e-02,
                -4.085400e-02, -1.167900e-01,  4.343520e-01,  2.735360e-01,
                -1.679880e-01,  8.428070e-01, -1.624070e-01, -3.897800e-02,
                -6.519830e-01, -8.327200e-02,  3.444680e-01,  1.204340e-01,
                -1.023900e-02,  7.358270e-01, -2.816040e-01,  1.603210e-01,
                 4.470930e-01,  5.938380e-01,  7.819760e-01, -5.484740e-01,
                 7.616150e-01,  1.139790e-01, -4.743390e-01, -5.187450e-01,
                -2.150580e-01, -2.989400e-02,  2.082940e-01, -8.029910e-01,
                 3.878030e-01, -3.356420e-01,  1.038700e-02,  3.317250e-01,
                 1.319130e-01,  2.267950e-01,  2.580530e-01,  2.564520e-01,
                -7.312100e-02, -7.604130e-01, -3.855650e-01, -1.218490e-01,
                -4.093900e-02,  7.754300e-02,  7.516170e-01, -5.169900e-02,
                 8.453380e-01,  5.112960e-01,  1.365590e-01, -4.424210e-01,
                -1.416480e-01, -2.211650e-01, -1.742920e-01,  3.196090e-01,
                 4.508490e-01, -2.310500e-01,  1.722220e-01, -7.309800e-02,
                -8.546800e-02,  5.844400e-02, -2.693930e-01, -4.131000e-03,
```

图 5-6 文本分布式表示示例

5.1.3 基于 Attention-BiLSTM 的留言文本分类模型

1. LSTM 网络

长短期记忆网络是一种特殊的循环神经网络，它可以学习长期依赖并解决梯度爆炸或消失的问题。LSTM 的基本神经元是记忆单元，每个记忆单元由输入门、输出门和遗忘门组成，其单元结构如图 1 所示。其中长期状态 c 用于存储长期记忆信息，使得序列的长期状态可以保存下来，并传递到下一层，同时，遗忘门的设计又使得 c 得到更新，丢弃已经过时的信息。

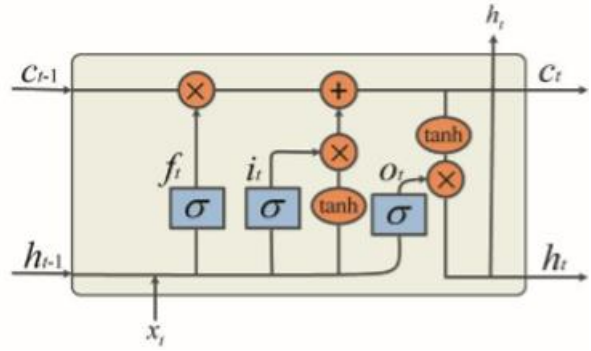


图 5-7 LSTM 神经元

遗忘门决定需要保留或舍弃的信息，实现对历史信息的存储。在时刻t，遗忘门根据上一时刻隐藏层的输出结果h_{t-1}和当前时刻的输入x_t同时作为输入。

$$f_t = sigmoid(W_f \cdot [h_{t-1}, x_t] + b_f)$$

输入门会把上一时刻 LSTM 单元的输出结果h_{t-1}和当前时刻的输入x_t都作为输入。

$$i_t = sigmoid(W_i \cdot [h_{t-1}, x_t] + b_i)$$

对于当前的候选单元记忆值是由当前输入数据x_t和上一时刻 LSTM 单元的输出结果h_{t-1}决定的。

$$\tilde{C} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

当前时刻记忆单元状态值C_i除了由当前的候选单元C_m以及自身状态C_{t-1}外，还需要通过输入门和遗忘门对这两部分因素进行调节。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}$$

输出门O_t用于控制记忆单元状态的输出。

$$O_t = sigmoid(W_o \cdot [h_{t-1}, x_t] + b_o)$$

最后 LSTM 单元的输出为h_t。

$$h_t = o_t * \tanh(C_t)$$

2. 注意力机制

每一个训练文本都由前向和后向的 LSTM 层组成，在第t时刻输入的x_t在 BiLSTM 中通过 BiLSTM 层提取特征后，模型能够学习更长距离的文本关系。BiLSTM 可以看成两个单向的 LSTM，所以 BiLSTM 在t时刻的隐藏层状态通过前向隐藏层状态$\overrightarrow{h_t}$和后向隐藏层状态$\overleftarrow{h_t}$加权求和得到，公式如下。

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h}_{t-1})$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h}_{t-1})$$

$$h_t = w_t\overrightarrow{h_t} + v_t\overleftarrow{h_t} + b_t$$

w_t、v_t分别表示t时刻 BiLSTM 所对应的前向隐藏层状态$\overrightarrow{h_t}$和后向隐层状态$\overleftarrow{h_t}$所对应的权重，b_t表示t时刻隐层状态所对应的偏置。

Attention 层的输入为上一层经过 BiLSTM 层处理过的输出向量h_t，此时的h_t是通过前向隐藏层状态$\overrightarrow{h_t}$和后向隐藏层状态$\overleftarrow{h_t}$加权求和得到的h_t，Attention 机制

层的计算如下：

$$u_t = \tanh(w_w h_t + b_w)$$

$$a_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)}$$

$$s_t = \sum_t a_t h_t$$

h_t是上一层 BiLSTM 层中的输出向量，w_w表示权重系数，b_w表示偏置系数，u_t表示h_t所决定的能量值，a_t为各个隐藏层状态在新的隐藏层状态中所占比重大小的权重系数，u_w为表示随机初始化的注意力矩阵，并在训练过程中不断学习，s_t为经过 Attention 机制的输出向量。输出层的输入是上一层 Attention 机制层的输出，其采用 softmax 函数对输出层的输入进行相应计算，从而进行留言分类并输出结果。

$$y_j = softmax(w_j s_t + b_j)$$

w_j表示 Attention 机制层到输出层的待训练的权重系数矩阵，b_j表示待训练相对应的偏置，y_j为输出的留言类别。基于 Attention 机制的 BiLSTM 模型中注意力机制层和前后向传播层通过h_i计算连接，可以充分发挥 Attention 机制的模型优势。基于 Attention-BiLSTM 的留言文本分类模型构建流程示意图如下图所示。

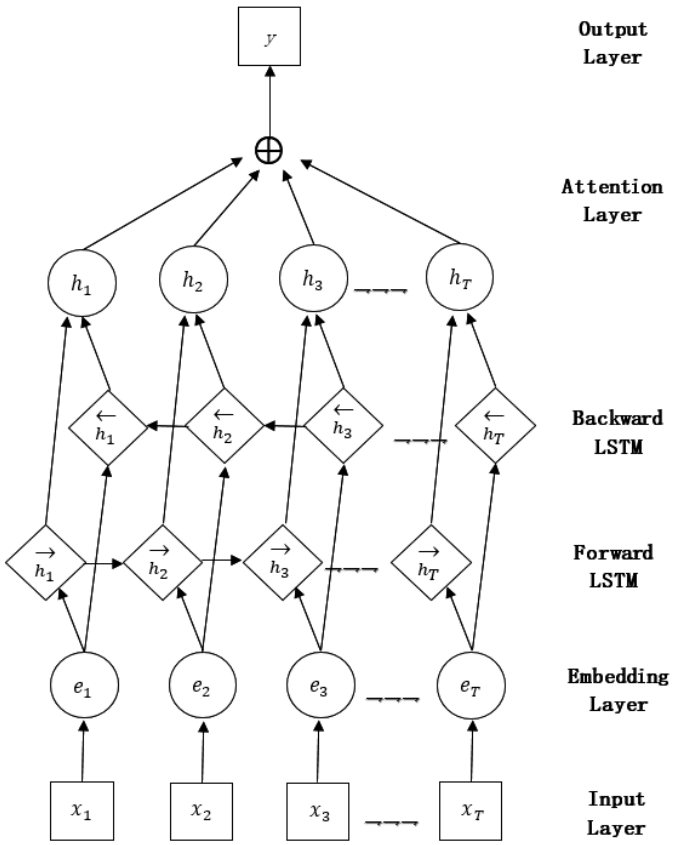


图 5-8 Attention-BiLSTM 模型

5.1.4 模型的求解

1. 训练数据及验证数据介绍

本题以附件 2 中留言详情及通过回译技术得到的文本数据为训练、验证数据，标签分为七类，数据总共有 13400 条，其中用于训练的文档数量共 12060 篇，用于测试的文档数量共 1340 篇，各类别文本数量分布如下表所示。

表 5.1 样本分布

类号	类别	总样本数	训练集	测试集
0	城乡建设	2009	1808	2001
1	环境保护	1846	1662	184
2	交通运输	1839	1655	184
3	教育文体	1876	1688	188
4	劳动和社会保障	1969	1772	197
5	商贸旅游	2016	1814	202
6	卫生计生	1845	1661	184

2. 模型训练及结果分析

本文采用使用 python3.7，基于 TensorFlow1.13 的深度学习库 Keras 进行求解。根据题目要求，评价方法采取 F-socre，公式如下：

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

其中P_i为第i类的查准率，R_i为第i类的查全率。

为了减轻过拟合，我们在训练阶段随机丢弃神经网络中的神经单位，在 BiLSTM 层使用 dropout 函数，并取掩码概率 p 的值为 0.1。输出层则使用 softmax 函数，将 7 维的特征向量$(a_1, a_2, \cdots, a_k)$按概率分布进行归一化处理，同时映射成输出向量$(b_1, b_2, \cdots, b_k)$，其中每个元素的取值范围均在 0 到 1 之间，最后挑选概率值最大的作为预测类别，具体参数设置如下表：

表 5.2 参数设置

参数	值	参数	值
词最大长度	256	网络层数	6
epoch	50	学习率	0.001
batch-size	128	dropout	0.1
优化函数	Adam	BiLSTM 隐藏层大小	256
损失函数	交叉熵	Attention 隐藏层大小	50

最终训练的模型在测试集上 **F-score** 达到 **93.75%**，在训练集上达到 100%。从二者数值来看，仍存在一定过拟合问题，这有可能是因为训练数据不足。另一方面，我们考虑到可能由于模型参数没有调整到最优，或者是网络结构的设计还可以改善，由于比赛时间的限制，我们将在赛后继续完善结果。

5.2 问题 2 的建模与求解

5.2.1 基于余弦相似度的 KMeans 算法

问题二的目的是挖掘热点问题，这就要求我们把相似的留言提取出来进行归类，而文本数据无法被计算机识别，因此我们要选取文本特征，将文本向量化表示并度量文本之间的相似度，然后使用机器学习的方法归类。

1.文本特征选择

在文本的特征选择问题上，我们使用常用的无监督学习特征选择方法文档频率 DF(Document Frequency)处理数据。DF 是简单的一种特征选择算法,它指的是在整个数据集中有多少个文本包含某个单词。

计算文本中每个特征词的文档频次，并且根据预先设定的阈值去除文档频次比较低和比较高的特征词，若该词的 DF 值小于阈值则将其删除,若其 DF 值大于阈值也将其去除。它们分别表示“没有代表性”和“没有区分度”两种极端的情况。DF 特征选取方法去除的特征词要么含有有用信息,要么太少而不足以对分类产生影响。DF 方法的优点在于计算量小并且能去除噪音、降低维度，在实际运用中有很好的效果。

数据分词预处理之后，我们使用 DF 方法特征选择剔除文档频率小于 3 或大于 1000 的词汇，这样在转化 TF-IDF 权值矩阵时其维度降到了一万以下，有效减少了计算量并降低维度。

2.文本向量化表示

选择文本特征词之后，需要将特征词转化为计算机可以识别的信息，使用 TF-IDF 方法表示特征权重，计算文本向量。TF-IDF 方法实际上是 TFIDF，TF(Term frequency) 即为关键词词频，是指一篇文档中关键词出现的频率，IDF(Inverse document frequency) 指的是逆文档频率，是用于衡量关键词权重的指数。

它的主要思想是如果某个词在一篇留言中出现的频率高(即 TF 高)，并且在其他文章中很少出现(IDF 高)，则认为此词或者短语具有很好的类别区分能力，适合用来表示文本的特征信息。TF 刻画了词语 t 对某篇文档的重要性，IDF 刻画了词语 t 对整个文档集的重要性。

假设 N 为某个词语 t 在某篇文档中的频次，M 是该文档的词频数，那么

$$TF = N/M$$

D 为所有的文档个数，D_t 为出现该词的文档个数，那么

$$IDF = log\frac{D}{D_t + 1}$$

$$TF - IDF = TF * IDF$$

由此我们便可以计算出选择的特征词语的 TF-IDF 值，这里就完成了文本数据的向量表达，计算出 TF-IDF 值之后，我们便可使用聚类分析算法对留言数据完成归类。

5.2.2 文本余弦相似性度量

前述已经完成了文本的向量化表达，下一步的目标是识别出相似的留言，那么该如何定义两篇文档的相似性是一个重要的问题，文本最常使用的相似度量算法是向量空间余弦相似度算法。

首先简单计算二维向量的夹角余弦值，平面中二维向量夹角示意如下图所示：

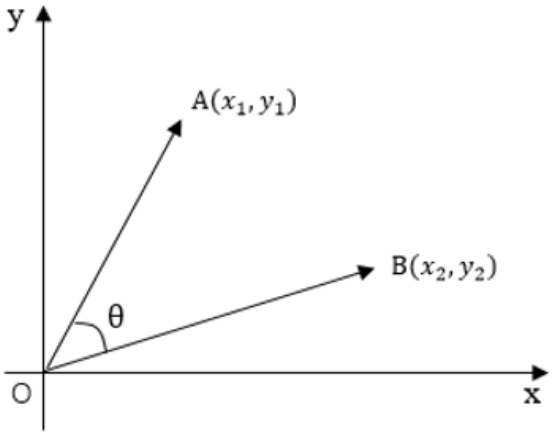


图 5-9 二维平面夹角余弦示意图

二维平面夹角余弦值公式为

$$\cos(\theta) = \frac{\overrightarrow{OA} * \overrightarrow{OB}}{|OA| * |OB|} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

由此我们可以拓展到 n 为向量夹角余弦，对于两个向量 $A_1(x_1, x_2 \cdots x_n)$ ，$A_2(y_1, y_2 \cdots y_n)$ 它们的 n 维夹角余弦公式为

$$\cos(\theta) = \frac{\overrightarrow{OA_1}\ \overrightarrow{OA_2}}{|OA_1||OA_2|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

根据数学知识，我们知道向量的夹角余弦值是有可能为负数的，但是这里算出的每个词语的 TF-IDF 值都是正数，因此我们计算出的两个文本余弦相似度的范围是[0,1]，两个文本的向量余弦值越接近于 1，那么其相似度越高。

5.2.3 KMeans 算法介绍

前述我们使用两条留言的夹角余弦来度量它们之间的相似度，在计算出相似度之后就需要使用机器学习算法对留言进行归类。留言归类使用的是无监督学习算法，最常使用的算法是 KMeans 聚类算法，该算法计算类间相似度一般采用的是欧几里得距离。这里修改其相似度量方法为向量夹角余弦值，使用基于夹角余弦相似度的 KMeans 算法对留言问题进行归类。

KMeans 算法需要首先给出 K 值，也即是聚类的类别数，然后如下五步完成对数据的聚类过程。

Step1: 随机选取 K 样本作为类中心；

Step2: 计算各样本到各类中心的相似度；

Step3: 将各样本归于最近的类中心点；

Step4: 求各类的样本的均值，作为新的类中心；

Step5: 类中心不再发生变动或达到迭代次数，则算法结束，否则回到 Step2。

5.2.4 热度评价指标的定义

完成留言归类工作后，我们还需要定义一个指标来衡量某一类问题的热度以达到挖掘热点问题的目的。留言热度的影响因素包括该类留言的个数、该类留言的点赞数和该类留言的反对数，以这三个因素来定义留言热度评价指标。

定义该指标的主要思想是考虑类中的留言个数在所有留言中的占比、类中留言点赞数之和在所有留言点赞数之和中的占比、类中留言反对数之和在所有留言反对数之和中的占比，这里认为留言赞成或者反对都体现出了留言是值得关注的，因此为每个占比赋予合适的比例加权求和便可得到留言热度。

首先定义各种符号的表示，$C_k, k = 1,2 \cdots K$，表示第k类，$n_k, k = 1,2 \cdots K$ 表示第k类留言个数，s_i表示第i条留言中点赞的个数，o_i表示第i条留言中反对的个数，ω_1，ω_2，ω_3表示分别赋予留言个数、支持度、反对度的权值，表示

$$N = \sum_{k=1}^{K} |n_k|$$

$$N_s = \sum_{i=1}^{N} s_i$$

$$N_o = \sum_{i=1}^{N} o_i$$

定义每类留言的热度指数

$$Heat_k = w_1 \frac{n_k}{N} + w_2 \frac{\sum_{i \in C_k} s_i}{N_s} + w_3 \frac{\sum_{i \in C_k} o_i}{N_o}, \quad k = 1,2 \cdots K$$

通过该公式计算出每类留言的热度便可以挖掘出热点问题。

5.2.5 模型的求解与分析

依据模型对附件 3 的数据做分析处理，主要包括以下过程：

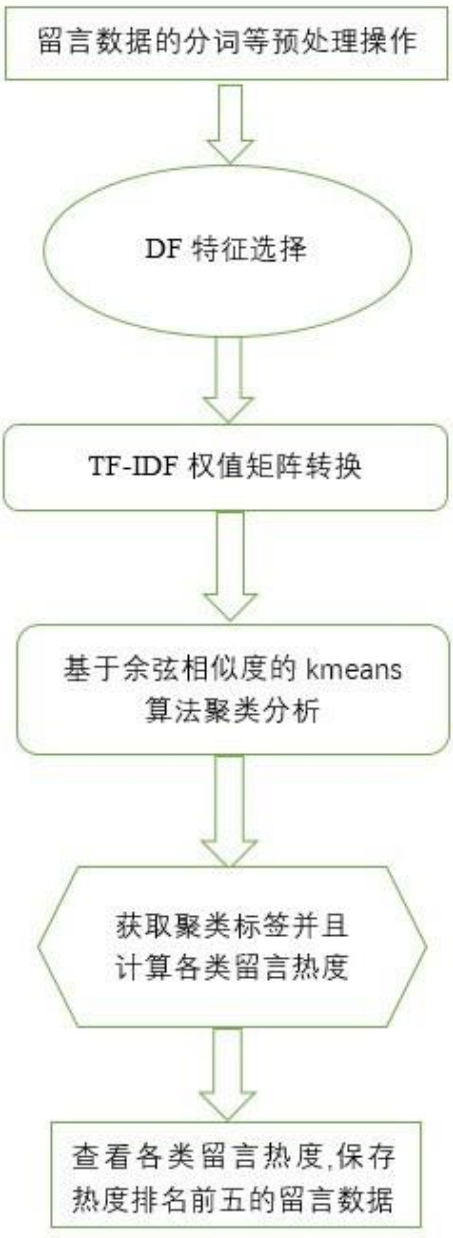


图 5-10 热点问题建模流程图

1. 模型求解

处理数据过程中，DF 特征选择方法保留文档频率大于 5 小于 1000 的词语，降低了 TF-IDF 权值矩阵的维度，减少了聚类算法的计算量。

通过预先的测试并结合轮廓系数的方法，我们最终选择 K=200 进行聚类分析。根据聚类结果计算留言热度，综合分析认为留言个数与留言支持数更能体现热度，赋予各个比例的权值分别是 0.6，0.35，0.05。

查看保存的留言表格，我们发现一些类别数目相对多、热度排名靠前，但是

其问题并不集中。这是聚类过程中可能存在的问题，确定的类别 K 使大量不集中的留言被归为了一类，同时影响了热度的排名。因此我们计算热度时不考虑这些明显异常的类别。根据以上的做法，我们得到了如下留言热点问题表。

表 5.3 热度排名前五的热点问题

热度排名	热度指数	时间范围	地点/人群	问题描述
1	6.88	2019/01/08 至 2019/07/08	A 市 A4 区	58 车贷案件进展
2	0.925	2019/03/26 至 2019/04/15	A2 区丽发新城	搅拌站噪音扰民，环境污染严重
3	0.917	2019/11/02 至 2020/01/26	A 市伊景园滨河苑	强制捆绑销售车位
4	0.717	2019/07/07 至 2019/09/01	A6 区月亮岛路	110kv 高压线规划影响居民生活
5	0.442	2019/01/08 至 2020/01/06	A 市中学	补课收费

2. 结果分析

画出前五类类别个数、支持度、反对度和热度指数的直方图。

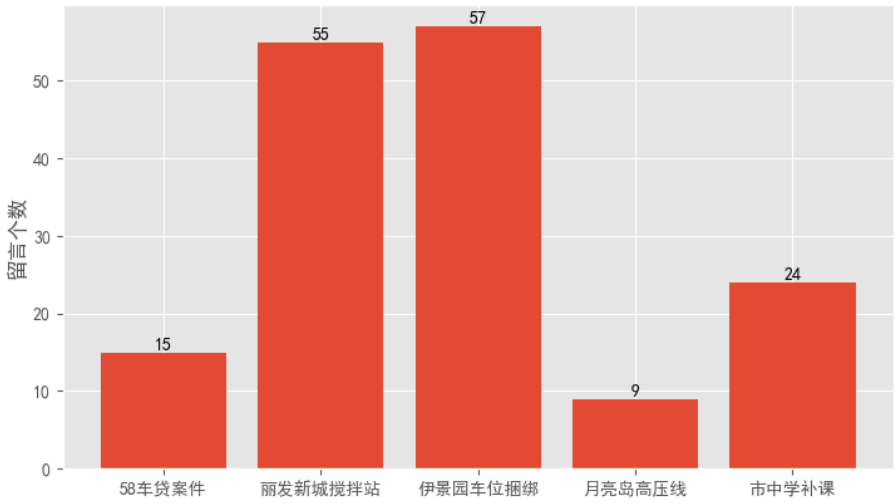


图 5-11 前五类留言个数直方图

留言个数能够最直观地反映出该类留言的热度，它是热度指标的重要部分。

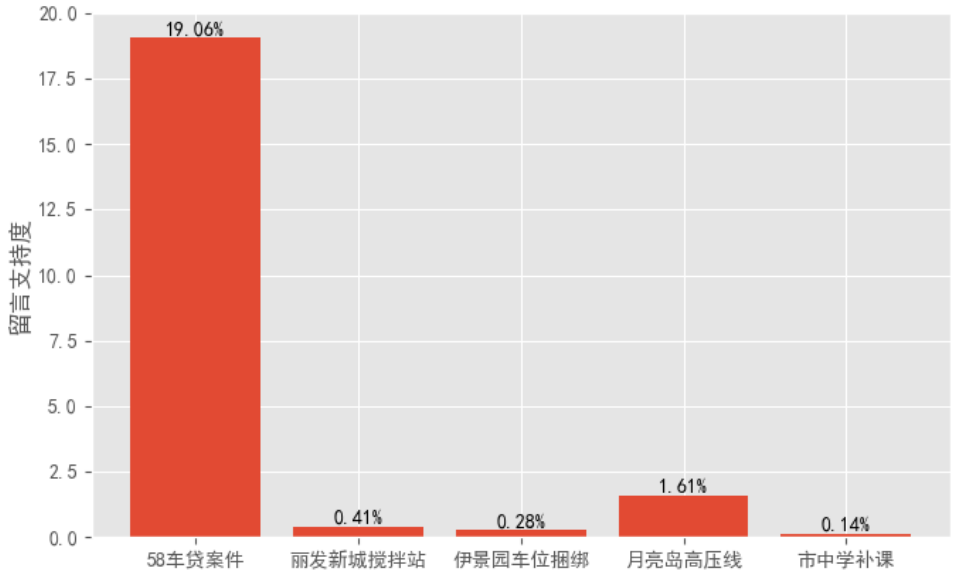


图 5-12 前五类留言支持度直方图

留言的支持度是指该类留言赞成数之和的占比，它反映了群众对该类问题的关注程度，它是热度指标中最重要的部分。直方图可以看出 58 车贷案件受群众关注度极高。

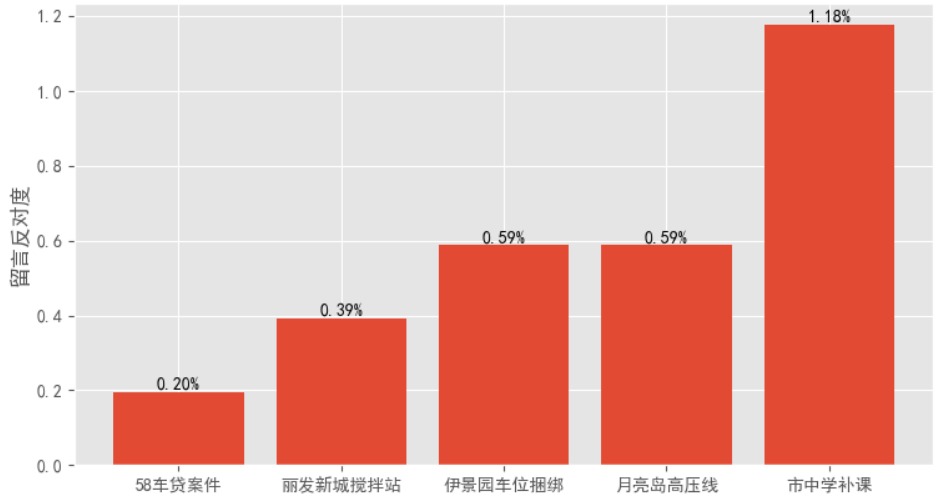


图 5-13 前五类留言反对度直方图

留言的反对度是指该类留言对数之和的占比，它反映了群众对该类问题的不一致意见。全部数据中留言反对数总数较少，因此计算热度时进行了权值调节。

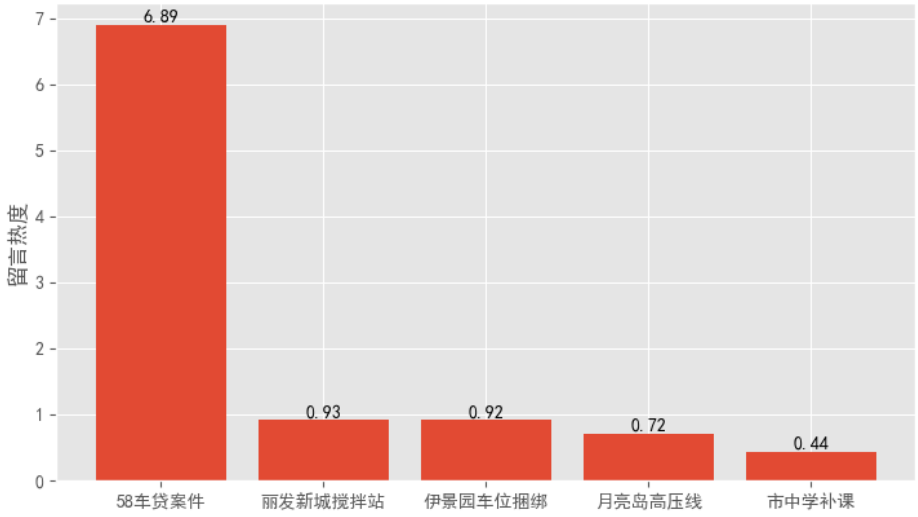


图 5-14 前五类留言热度直方图

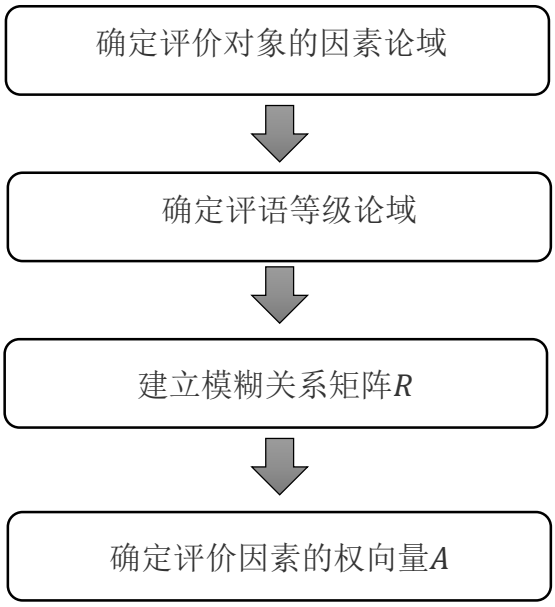
热度直方图直观地显示出了五类留言的热度情况，由此我们便完成了留言热点问题挖掘的目标。

5.3 问题 3 的建模与求解

5.3.1 基于模糊综合评价模型的留言评价体系

1. 模糊综合评价模型

模糊综合评价法是一种基于模糊数学的综合评价方法。该综合评价法根据模糊数学的隶属度理论把定性评价转化为定量评价，即用模糊数学对受到多种因素制约的事物或对象做出一个总体的评价，其建模流程图如下：



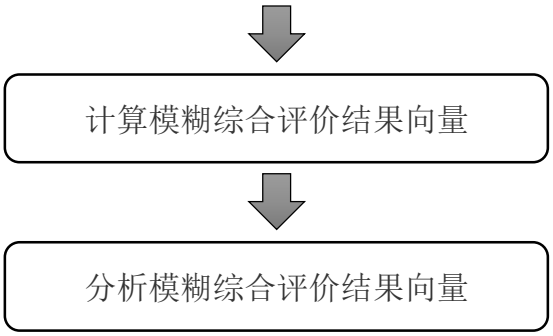


图 5-15 模糊综合评价模型建模流程图

Step1：确定评价对象的因素论域

因素论域是以刻画评价对象的评价指标为元素所组成的一个集合，一般用u表示，$u=\{u_1,u_2,\cdots,u_p\}$，其中u_i表示影响评价对象的第i个因素，其往往存在不同程度的模糊性。

Step2：确定评语等级论域

设$v=\{v_1,v_2,\cdots,v_n\}$，是评价者对被评价对象可能做出的各种总的评价结果组成的评语等级的集合，其中v_i表示第j个评价结果。

Step3：建立模糊关系矩阵R

在构造了等级模糊子集后，就要逐个对被评价对象从每个因素u_i上进行量化，也就是确定从单因素来看被评价对象对各等级模糊子集的隶属度，进而得到模糊关系矩阵，即：

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{bmatrix}$$

其中，r_{ij}表示从因素u_i来看对v_i等级模糊子集的某个被评对象隶属度；

Step4：确定评价因素的权向量A

设$A=(a_1,a_2,\cdots a_m)$为权重分配模糊矢量，其中a_i表示第i个因素的权重，权向量A反映了各因素的重要程度。

Step5：计算模糊综合评价结果向量

将权向量A与模糊关系矩阵R进行矩阵乘法，求得被评对象的模糊综合评价结果向量B，即：

$$A \cdot R = (a_1, a_2, \cdots a_p) \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{bmatrix}$$

$$= (b_1, b_2, \cdots b_m) = B$$

其中，b_j表示从整体上看对v_j等级模糊子集的被评对象的隶属程度。

Step6：分析模糊综合评价结果向量

2. 模型的建立

本题将答复意见的评价分为五个等级，分别为‘非常满意’、‘满意’、‘一般’、‘不满意’、‘非常不满意’，等级划分如下：

表 5.4 评价等级划分

非常满意	满意	一般	不满意	非常不满意
[0.9,1]	[0.6,0.9)	[0.4,0.6)	[0.1,0.4)	[0,0.1)

影响评价答复意见的因素有很多方面，我们综合分析题意及附件 4，总结出以下四个主要方面：

a. 相关性

网络问政平台的主要任务就是倾听老百姓的各种问题，为老百姓排忧解难。而答复意见最重要的一点就是要正确理解对应留言的问题，针对问题给出对应的答复，两者应具有高度相关性，不能答非所问，甚至对问题闭口不提。因此我们考量答复意见的首要指标就是相关性。

利用问题二中的方法求解留言内容和答复意见的**相似度**，并根据以下方案量化：

表 5.5 相关性量化

非常满意	满意	一般	不满意	非常不满意
[0.9,1]	[0.6,0.9)	[0.4,0.6)	[0.1,0.4)	[0,0.1)

b. 及时性

作为百姓和政府沟通的有效途径之一，百姓的留言是否能够得到**及时**的答复是答复意见评价体系中重要的一个评价指标。如果答复的不及时，无论答复意见多么地相关、完整，都是徒劳的，网络问政平台也就失去了它的意义。我们发现附件 4 中有答复时间和留言时间，从而我们可以通过二者的**时间差**来量化答复意见的及时性。

考虑到留言问题分类较多，答复难度参差不齐等实际因素，我们采取以**工作**

周（7 个工作日）为单位的方式量化及时性，方案如下：

表 5.6 及时性量化

非常满意	满意	一般	不满意	非常不满意
(0,1]	(1,2]	(2,3]	(3,4]	(4,+∞]

c. **完整性**

答复意见代表了政府形象，不仅需要在内容上让人民满意，更要在格式上做到完美。通过观察附件 4，我们发现答复意见应满足如下规范：1）**开篇有敬语**，如：你好，您好、尊敬的网友等；2）敬语后应紧接着表达群众的留言内容**收到并知悉**，如函件收悉、留言收悉等；3）文中有对政府部门工作的支持**感谢**，如‘感谢您对我们工作的信任与支持’等；4）**结尾有留言时间**。

根据上述规范对完整性进行量化，方案如下表所示：

表 5.7 完整性量化

非常满意	满意	一般	不满意	非常不满意
满足全部规范	满足三条规范	满足两条规范	满足一条规范	无任何规范

d. **可解释性**

作为官方给出的答复意见，必须要有可解释性才能使百姓信服。在附件 4 中，我们将可解释性归纳如下：1）是否**了解、核实**相关问题；2）是否说明问题**原因**；3）是否给出有效的**解决方案**； 4）是否有**法律层面**上的依据。

根据上述规范对可解释性进行量化，方案如下表所示：

表 5.8 可解释性量化

非常满意	满意	一般	不满意	非常不满意
满足全部规范	满足三条规范	满足两条规范	满足一条规范	无任何规范

5.3.2 模型的求解与分析

本文以留言编号 3910 为例，对模型进行求解与分析，答复意见如下图：

网友“UU0081955”您好！您的留言已收悉。现将有关情况回复如下：根据《西地省物业服务收费管理办法》十七条之规定“房屋交付期限届满之日的次月及以后的物业服务费，由业主交纳”，开发商已于2017年8月24日向您邮寄《岸海保利西交房通知书》，因您自身原因未按交付使用通知书约定的交付时间办理交付手续，开发商履行交房义务的时间应当认定为交付使用通知书约定的交付时间，而并非您所称的实际交付之日。故交付使用通知书约定的交付时间的次月及以后的物业服务费，应由您个人承担。根据《西地省物业服务收费管理办法》的相关规定及合同约定，您反映的问题实属民事合同纠纷，双方可就自身主张提起民事诉讼或仲裁维护自身合法权益。另经与物业公司沟通，物业公司已同意按照合同约定给予空置期间物业费9折优惠。感谢您对我们工作的支持、理解与监督！2018年12月24日

图 5-16 留言编号 3910 的答复意见

1. 计算模糊综合判断矩阵

为了计算模糊综合判断矩阵，我们需要先计算单因素评价矩阵。

1）相似性

利用问题二中的余弦相似度求得留言详情和答复意见的文本相似度 0.9345012，过程不再赘述。对照上文相关性量化表格，$0.9345012 \in (0,1]$，故 $R_1 = (1,0,0,0,0)$。

2）及时性

在附件 4 中，可以看到答复时间为 2018/12/29 15:02:16，而留言时间为 2018/12/7 20:56:34，可以计算留言答复时间差为 22 天，约为 3.14 周，$3.14 \in (3,4]$，故$R_2 = (0,0,0,1,0)$。

3）完整性

观察答复意见，可以发现开篇有敬语‘您好’，随后表达群众的留言内容收到并知悉，文中有对留言群众的感谢，最后有留言时间，满足完整性量化的所有规范，故$R_3 = (1,0,0,0,0)$。

4）可解释性

在答复意见中，工作人员讲清楚了事情的来龙去脉，解释了问题的原因：‘因自身原因未按交付使用通知书约定的交付时间办理交付手续’，同时给出了可行的解决方案：‘给予空置期间物业费 9 折优惠’，并提到了法律层面的依据《西地省物业服务收费管理办法》。综上所述，本条答复意见满足可解释性量化表的全部规范，故$R_4 = (1,0,0,0,0)$。

综上所述，得到模糊评价判断矩阵：

$$R = \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

2. 确定评价因素的权向量矩阵

在各评价因素中，相关性无疑是首先要考虑的，答复务必要和留言问题有对应关系；其次考虑及时性，答复要尽快地给出，否则再好的回复都没有意义；完整性和可解释性作为一种答复规范的要求，也是必不可少的。考虑答复意见的各重要性，我们给出权向量矩阵如下：

$$A = (0.5, 0.3, 0.1, 0.1)$$

3. 求解模糊向量

确定模糊综合判断矩阵 R 和因素权向量 A 之后，对 A 和 R 进行矩阵乘法，求解得到模糊向量：

$$B = A \cdot R = (0.7, 0, 0, 0.3, 0)$$

故对留言编号 3910 的评价为‘非常满意’。

六、模型的评价与进一步讨论

6.1 模型的评价

1. 模型的优点：

①在第问题一中，对非均衡样本进行了采用回译技术的数据增强操作，使各类训练数据达到均衡，效果的提升在 4%左右；

②在第问题三中，利用模糊综合评价法进行建模，对各个评价因素进行量化，构建了完善的政府答复评价体系。

2. 模型的缺点：

①由于时间紧迫，我们对模型参数的调整存在欠缺，同时在网络结构设计上也可以进行更深一步的考虑；

②在构建留言答复评价体系时，考虑的影响因素不是很全面。

6.2 模型的进一步讨论

对于留言答复评价的影响因素，有关部门可以尝试在网上征求群众意见，得到更加贴近百姓实际的业务需要，进而改进评价体系。比如，可以让留言者对得到的答复意见进行评价，通过评价分类和答复意见构建分类模型，实现系统自动审核工作人员的答复意见，若系统评价达到阈值，则可以反馈给留言群众，若未通过，则需要返工修改。

七、参考文献

[1] 方炯焜,陈平华,廖文雄.结合 GloVe 和 GRU 的文本分类模型[J].广东工业大学计算机学院,2020,4.

[2] 关立刚,陈平华.基于注意力机制和残差连接的 BiLSTM-CNN 文本分类[J].广东工业大学计算机学院,2019,6.

[3] 姜启源.数学模型(第四版)[M].北京:高等教育出版社,2011,1.

[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].浙江工商大学管理工程与电子商务学,2020,4.

[5] 但宇豪,黄继风,杨琳,高海.基于 TF-IDF 与 word2vec 的台词文本分类研究[J].上海师范大学信息与机电工程学院,2020,2.

[6] 牛雪莹,赵恩莹,基于 Word2Vec 的微博文本分类研究[J].太原科技大学计算机科学与技术学院,2019,8.

[7] 司守奎,孙玺菁.数学建模算法与应用[M].北京国防工业出版社,2011,5.

[8] 卓金武.MATLAB 在数学建模中的应用[M].北京航空航天大学出版社,2011,4.

[9] 冀相冰,朱艳辉,李飞,徐啸.基于 Attention-BiLSTM 的中文命名实体识别[J].湖南工业大学计算机学院,2019,9.

[10] 李润川,张行进,王旭,陈刚,冀沙沙,王宗敏.基于单心搏活动特征与 BiLSTM-Attention 模型的心律失常检[J].解放军信息工程大学数学工程与先进计算国家重点实验室,2019,10.