第八届“泰迪杯”
全国数据挖掘挑战赛

作品名称：“智慧政务”中的文本挖掘应用

作品单位：**********

作品成员：**********

指导老师：**********

（应组委会要求，匿名提交）

“智慧政务”中的文本挖掘应用

摘要：伴随着互联网的急速发展，互联网+电子政务已经成为简政放权和提升政府服务功能的一个重要手段。而网络问政平台作为互联网电子政务的重要组成部分，是政府了解民意、汇聚民智、凝聚民气的重要渠道。然而，依靠人工进行问政留言类别划分与热点整理已经成为网络问政平台信息处理的一个瓶颈问题。本作品以“智慧政务”中的文本挖掘应用为主要目标，以目前先进的机器学习技术作为手段，在群众留言分类、热点问题挖掘和答复意见评价三个方面进行了探索与实践。

针对问题 1：群众留言分类。由于群众留言通常具有文本长度短，用词差异化大等问题。传统的文本分类特征抽取算法(例如基于 TFIDF 的 One-Hot 表示方法)会遇到特征稀疏问题，因此，本作品尝试基于目前先进的预训练模型 BERT(Bidirectional EncoderRepresentation from Transformers)进行群众留言特征抽取，以获得更好的群众留言的低维稠密特征表示。在真实数据上的实验表明，本作品所用算法相对于传统文本分类算法具有更好的分类性能。

针对问题 2：热点问题挖掘。本作品首先使用经典的聚类算法 K-means 对群众的留言进行聚类的划分。其次，充分利用了 BERT 预训练模型的优越性，尝试将其应用于热点问题元素的抽取，进行地点/人群等热点问题元素划分。进一步，应用了 TextRank 算法对聚类得出的数据进行摘要划分，尝试在高语义层次上对热点问题进行总结。

针对问题 3：答复意见评价。本作品基于问答文本间的相似度，问答文本提及的实体相似程度分别定义了相关性，完整性规则，并设计相关的加权算法给出相应的星级评价，最终将相关性、完整性、星级等字段制成答复评价质量表完成了评价任务。

关键词：BERT; 短文本分类; 热点挖掘; 实体抽取; 文本摘要;

目录

1 背景分析

随着“互联网+政务服务”的不断推进，各级政府部门加快推动政务服务全面转型，同时数字化、精准化、智能化已经成为衡量政府工作能力的“新标尺”。因此，提升政府工作效率的智慧政务，已经成为在互联网新时代市场经济条件下的建设核心。

在政府运行的过程中，政府与民生之间的关系尤为重要。而网络问政平台是政府能够了解民意的一个重要手段之一，快速发现民众最急迫解决的问题和将民众反映的问题精准归类是网络问政平台的关键一环。然而，目前在处理网络问政平台的群众留言时，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低、准确率低等问题。

因此，随着人工智能技术的发展，建立基于自然语言处理技术的智能政务系统就应该成为提升政府的管理水平和施政效率的核心担当，对提升政府的管理水平和施政效率具有极大的推动作用。

2 挖掘目标

本次挖掘目标是根据互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，主要利用 BERT(Bidirectional EncoderRepresentation from Transformers)模型以及 K-Means 聚类实现以下三个目标：

（1） 利用目前先进的预训练模型 BERT 进行群众留言特征抽取，实现对群众留言的一级标签分类。

（2） 利用文本聚类实现相似留言问题的聚集，进行实体识别，基于 BERT 模型抽取地点、人群等热点元素，同时采用 TextRank 算法抽取问题的摘要，最后依据类别的聚集程度定义热点问题的热度指数，完成热点问题的挖掘。

（3） 根据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。

3 实现原理

3.1TF-IDF 算法原理

因为常用的机器学习模型（神经网络等）只能接受数值型输入，而分词处理后的留言信息是符号形式，要将它转换成数值型形式，所以我们就用 TF-IDF 算法将留言信息转换成权重向量[1]。TF-IDF 具体原理如下：

1. 计算词频 TF

$$\text{词频（TF）=某个词在文章中出现次数} \tag{3-1}$$

考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化：

$$\text{词频（TF）}=\frac{\text{某个词在文本中的出现次数}}{\text{文本的词数}} \tag{3-2}$$

或者

$$\text{词频（TF）}=\frac{\text{某个词在文本中的出现次数}}{\text{该文出现次数最多的词的出现次数}} \tag{3-3}$$

2.计算逆文档频率 IDF

需要一个语料库（corpus），用来模拟语言的使用环境。

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档书}+1}\right) \tag{3-4}$$

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近 0。分母之所以要加 1，是为了避免分母为 0（即所有文档都不包含该词）。log 表示对得到的值取对数。

3.计算 TF-IDF

$$\text{IF-IDF=词频（TF）}\times\text{逆文档频率（IDF）} \tag{3-5}$$

TF-IDF 如何表示一个句子：

加入一个句子有 n 个单词，每个单词计算出它的 tfidf 值，即每个单词用一个标量表示，则句子的维度是 1*n。

3.2BERT 模型原理

BERT 的全称是 Bidirectional Encoder Representation from Transformers，即双向 Transformer 的 Encoder，其目标是利用大规模无标注语料训练、获得文本的包

含丰富语义信息的 Representation，即：文本的语义表示，然后将文本的语义表示在特定 NLP 任务中作微调，最终应用于该 NLP 任务[2]。

它是将预训练模型和下游任务模型结合在一起的，也就是说在做下游任务时仍然是用 BERT 模型，而且天然支持文本分类任务，在做文本分类任务时不需要对模型做修改。为此，我们使用 BERT 中文预训练模型实现中文文本分类。

3.3K-Means 原理

K-Means 是典型的聚类算法，K-Means 算法中的 k 表示的是聚类为 k 个簇，Means 代表取每一个聚类中数据值的均值作为该簇的中心，或者称为质心，即用每一个的类的质心对该簇进行描述[3]。

用数学公式表示，若簇划分为$(a_1, a_2, \dots a_k)$，均值向量为u_i则目标是最小化平方误差 Y 值：

$$\mathrm{Y} = \sum_{i=1}^{k} \sum_{x \in a_i} \| x - u_i \|_2^2 \tag{3-6}$$

3.4 命名实体识别原理

命名实体识别（Named Entity Recognition，简称 NER），是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名等。NER 的问题通常被抽象为序列标注(Sequence labeling)问题[4]。

3.4.1BIO 序列标注原理

序列标注问题是 NLP 中最常见的问题，所谓“序列标注”，就是说对于一个一维线性输入序列，每个元素都会对应一个标签集合中的某个标签：

$$X = x_1, x_2, x_3, \dots x_n \tag{3-7}$$

$$Y = y_1, y_2, y_3, \dots\ y_n \tag{3-8}$$

BIO 标注(B-begin，I-inside，O-outside)原理是将每个元素标注为“B-X”、“I-X”或者“O”。其中，“B-X”表示此元素所在的片段属于 X 类型并且此元素在此片段的开头，“I-X”表示此元素所在的片段属于 X 类型并且此元素在此片段的中间位置，“O”表示不属于任何类型。

3.5 TextRank 原理

TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank 算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序[5]。核心算法公式如下：

$$\mathrm{WS(V_i)} = (1-\mathrm{d}) + \mathrm{d} * \sum_{\mathrm{V_i} \in I_n(\mathrm{V_i})} \frac{\omega_{\mathrm{ji}}}{\sum_{\mathrm{V_k} \in O_{ut}(\mathrm{V_j})} \omega_{jk}} \mathrm{WS}(V_j) \qquad (3\text{-}9)$$

等式左边表示一个句子的权重，右侧的求和表示每个相邻句子对该句子的贡献程度。求和的分子ω_{ji}表示两个句子的相似程度，$\mathrm{WS}(V_j)$代表上次迭代j的权重，整个公式是一个迭代的过程。

4 实现方案

4.1 数据预处理

本次建模数据用到附件一，附件二，附件三和附件四一共四张表，附件一是有三个等级标签的体系，附件二主要是用户的留言主题和详情，以及对应的一级标签等，一共有 9210 条数据。附件三主要是用户的留言主题，内容和点赞数等，一共有 4326 条数据，附件四主要是对留言内容相对应的答复，一共有 2816 条数据。

4.1.1 数据清洗

1.去重

在给的留言信息中存在一些重复数据，为了考虑删除了这些数据并要保留其中一个，则使用了 pandas 中的函数 drop_duplicates，因为考虑到同一个留言用户

可以留言多条信息，所以就去除完全相同的行数据，并将其参数 keep 设置成 first 保留了重复行的第一个数据。

2.去除转义字符

查看数据的时候，发现存在转义字符，于是用 replace 将“\n”,“\t”全都替换为空，并处理表中空格。

3.合并

对于表附件二用户的留言主题和详情去重，最终得到 9175 条数据。分别将表附件二、三的留言主题和留言详情合并为一列。

4.1.2 分词

为了让非结构化文本信息转换为计算机能够识别的结构化信息，采取了 Jieba 分词，Jieba 主要是基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法，能更好的实现分词效果。为了确保留言信息中停用词不对模型训练造成干扰，即对表附件二、三合并的留言内容进行了停用词处理。处理结果如下图 4-1 所示。

年 省 通信 公司 下岗 区 注册 一家 通信 科技 分公司 本想 有所作为 市场 评估 导...
前不久 人事部门 办理 人事 续聘 手续 档案 托管 费要 元 感觉 太贵 元 年 人事档案...
落实 精简 退职 职工 救济金 呼吁 荣 年 月 日 出生 现年 岁 居住 西地省 县 野鸡...
市 广播电视 系统 因工 致残 工作人员 市 人事局 评定 伤残 等级 政策 享受 伤残 抚...
今日 接到 县 居委会 电话 通知 身份证 退休证 居委会 报 登记 年 元月 停发 退休金...
希望 加 退休 工资 时能 养老金 低于 平均水平 倾斜 退休工人 加工资
原 市县 机械 灭螺队 亦工亦农 八二 年底 放假 未回 单位 八三年 解散 去年 查到 解...
尊敬 领导 楚税 社保 上交 农村 合作 医疗保险 写 天 审核 天 审核 客服 电话 客服...
妻子 居住 县泉 塘泉星 社区 自购 商品房 妻子 户主 第一个 孩子 户口 迁入 泉塘泉星..
您好 公司 工作 年 半左右 公司 交 公积金 少缴 五险 老板 说 没用 不肯 交 公积金...
生于 年 同 年 同 参加 工作 大专 学历 函授 中共党员 1995 年 年 政工 工作 ...
请问 办理 大中专 学生 就业 培训 资质 条件 请问 市 办理 大中专 学生 就业 培训 ...
尊敬 西地省 市 县 市委 市 人民政府 领导 县市 牧工商 公司 系原 县市 畜牧 水产局...
我原 国有企业 正式 职工 岁 离退休 差 年 交了 社保 金 生存 保障 情况 交上 年 ...
希望 厅长 在位 时多为 老百姓 创造 福利 条件 老百姓 创造 福利 条件
制卡 进度 已到 卡 配送 阶段 市 银行 但经 咨询 区 社保卡 服务 窗口 市 银行 制...
您好 市 公积金 存缴 用户 更换 工作 新老 公司 交接 导致 月份 公积金 续交 月份 ...
张 奶奶 年 八十 社会 劳动 三十多年 世纪 年代 几个 工友 申请 组建 废品收购 站 ...
想 咨询 中级 会计师 职称 考试 工作 年限 算法 大学 本科学历 会计工作 四年 是从 ...
全国 讲 基础 畜牧 工作人员 工资 最低 工作 辛苦 讲 请问 厅长 十二五 会涨 县 动...

图 4-1 数据处理结果

4.2 问题 1：群众留言分类

4.2.1 流程图

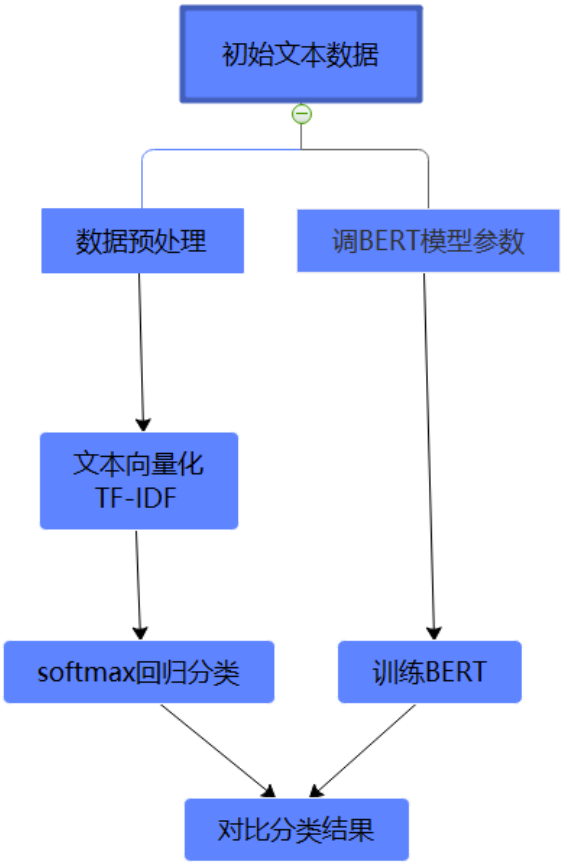

图 4-2 群众留言分类流程图

4.2.2 具体方案

分析群众留言，我们发现留言信息具有文本长度短、不同用词的差异化大的问题，所以本文采用了两种方案进行对比。

4.2.2.1TF-IDF 算法实现

方案一：首先使用 TF-IDF 算法进行文本特征表示，然后使用 softmax 回归训练分类模型。

实现 TF-IDF 向量的具体步骤如下：

（1）计算留言内容中每个词的词频 TF。

（2）计算留言内容中每个词出现的文档数 DF。

（3）用总留言数和 DF，得到逆文档频率，这里的公式如下。

$$\mathrm{IDF} = \log(\frac{\text{留言文档总数}}{\mathrm{DF}+1}) \tag{4-1}$$

（4）生成每个留言内容描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF） (4-2)

softmax 回归其实是逻辑回归的一种变形，逻辑回归模型输出的是两种类别的概率，本作品使用 softmax 回归实现 7 种类别的留言分类问题。

4.2.2.2BERT 参数指标说明

由于传统的文本分类特征抽取算法(TF-IDF)会遇到特征稀疏问题，所以我们尝试了第二种方案，使用 BERT 预训练模型对留言进行特征抽取。

在模型参数选择上，我们使用了 BERT 的 BERT-Base，Chinese 模型作为基础模型。其中总计有 12 层，110M 个参数，参数如下：句子的最大长度为 128。训练集，验证集合测试集每批次大小均为 32。学习率为 2e-5，训练的 epoch 次数为 3。

4.2.3 实验结果

4.2.3.1 评价方案

对于评价标准，我们都希望精准率和召回率都很高，故我们使用 F1 Score 作为用户输入意图分类的标准，以兼顾两种评价，F1 Score 可以看成模型精确率和召回率的一种加权平均。其定义如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{4-3}$$

P，R 分别为查准率和查全率，计算出的F_1值可以很好的评价模型。

4.2.3.2 BERT 与 TF-IDF 实验结果对比

我们先抽取了百分之十的验证集，再将剩下的数据按训练集 10%-90%进行切分，从而对比训练集在不同百分比下，BERT 和 TF-IDF 的准确率以及 F1 Score，其准确率如图 4-3,其 F1 Score 如图 4-4。

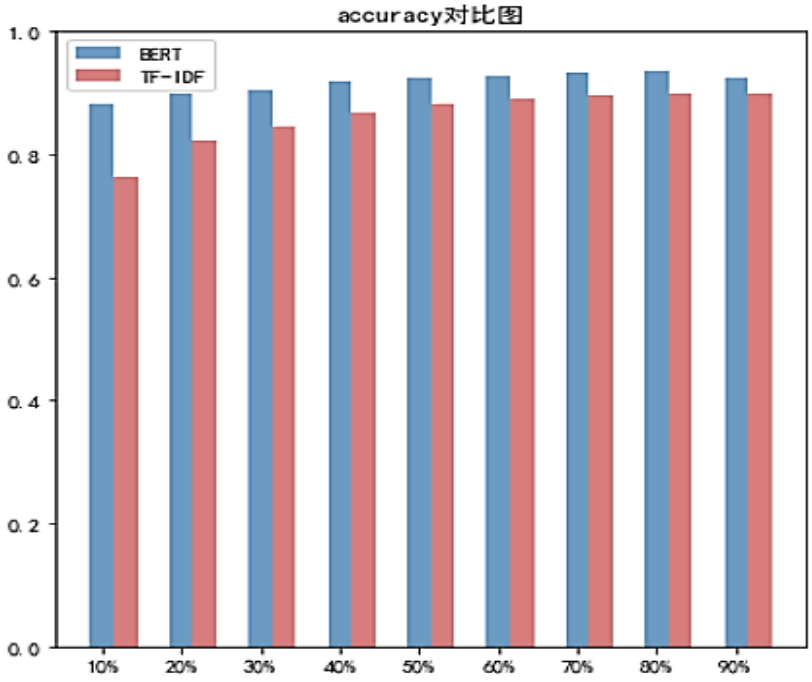

图 4-3BERT 和 TF-IDF 准确率对比图

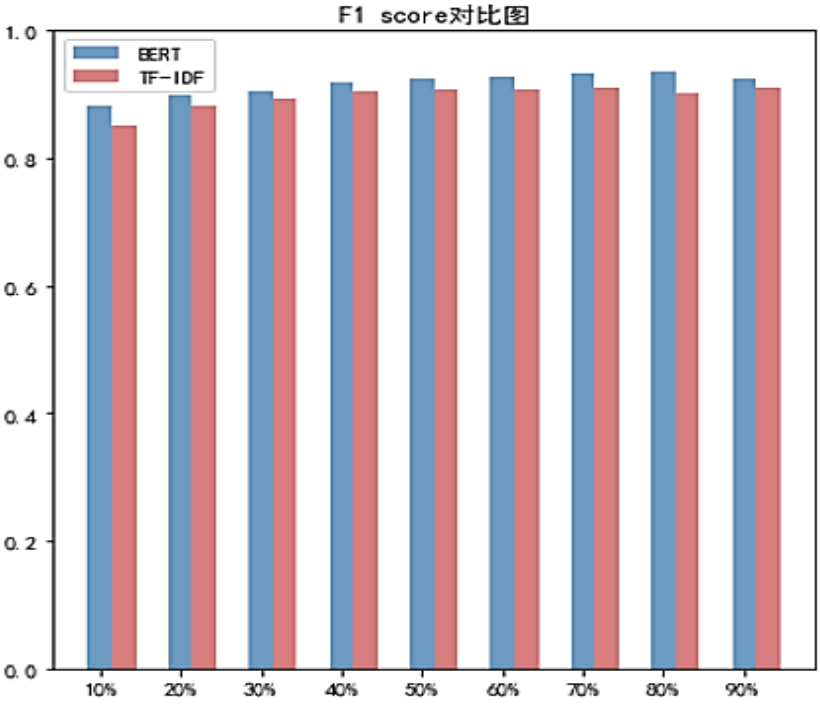

图 4-4BERT 和 TF-IDFF1 Score 对比图

从图 4-3 和图 4-4 可知，TF-IDF 对于训练集数量要求比较高，并且其准确率和 F1 Score 都没有 BERT 高，可能是 TF-IDF 训练向量时因为留言内容文字太短，得到的向量太稀疏。相对于 TF-IDF，就算训练数据只有百分之十的情况下，BERT 模型也达到了很好的准确率和 F1 Score，说明就算我们拥有留言数据很少的情况下，对其留言详情分类的准确率依旧会很好，体现了 BERT 模型的可行性。

4.3 问题 2：热点问题挖掘

为了挖掘热点问题，本作品对留言信息进行聚类，以类别的聚集程度衡量问题的热度指数。在热点问题聚类的基础上再将每个类的所有留言详情和留言主题进行摘要提取得到问题描述。最后针对问题中的地址和人群的提取，我们先使用 BIO 序列标注规则进行数据标注，再通过第三方库 Kashgari 搭建基于 BERT 的中文 NER 模型进行地址和人群的识别。最后针对问题要求的热度指数，我们将每一类问题的样本量除以总数据样本量并保留小数点后三位得到热度指数。

4.3.1 流程图

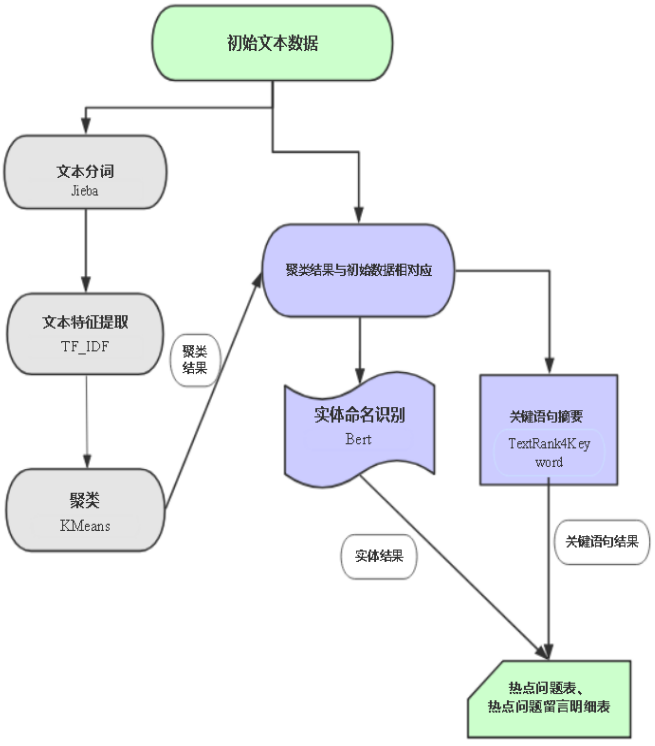

图 4-5 热点问题挖掘流程图

4.3.2 具体方案

4.3.2.1 基于 k-means 算法实现留言问题聚类

在聚类算法的选择方面，因为 k-means 原理简单，聚类效果较优，我们使用 k-means 对群众留言进行聚类。

在使用 K-Means 算法时，首先要注意的是 k 值的选择，可利用轮廓系数来选择最优的 K 值，以达到最优的分类效果。下图是在 4327 条数据中，将 k 的范围设置为 10-150 时的轮廓系数如下图 4-6 所示。

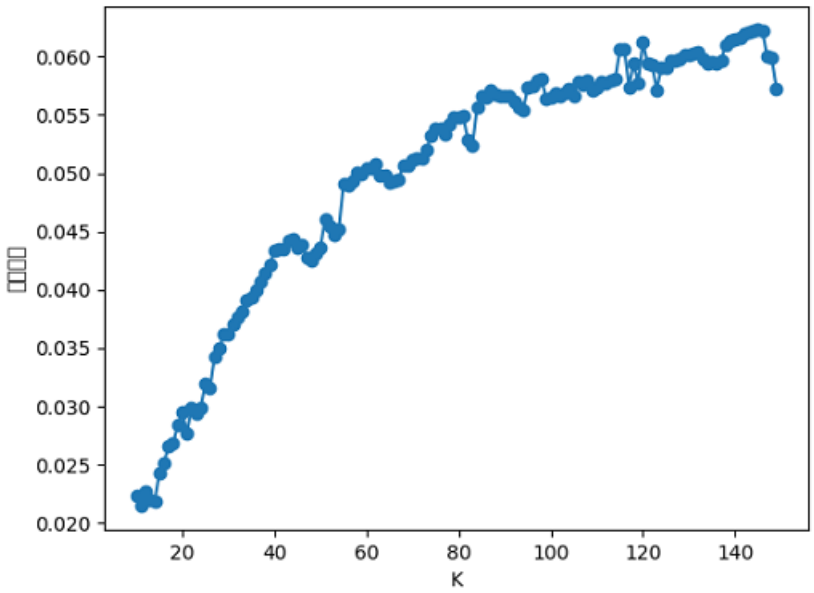

图 4-6 轮廓系数图

由此可见，当 k 等于 140 时聚类效果较优。将 tf-idf 值传入 K-Means 中开始聚类。参数说明如下：

model = KMeans(n_clusters=140,n_jobs=4, max_iter=500)

4.3.2.2 基于 TextRank 算法抽取留言摘要

本作品基于 python 的第三方库 textrank4zh 中的类 TextRank4Sentence 进行摘要提取，返回在文档中权重最大一个句子作为最终摘要。参数说明如下：

tr4s.analyze(text=text, lower=True, source = ‘all_filters’)，text 为文本信息，lower 开启大小写，使用 all_filters 生成句子之间的相似度。

4.3.2.3 基于 BERT 抽取热点元素

1.定义标签

本作品用到的标注标签有（B-LOC,I-LOC,B-PER,I-PER,B-ORG,I-ORG），B-LOC 和 I-LOC 标签对地址实体进行标注，B-PER 和 I-PER 标签对人群实体进行标注，B-ORG 和 I-ORG 标签对机构名称进行标注。

2.参数说明

在模型参数选择上，我们使用了 BERT 的 BERT-Base，Chinese 模型作为基础模型。通过创建 BERT embedding 并导入 BiGRU_Model 模型进行训练。其中参数如下：句子的最大长度为 100。训练集，验证集合测试集每批次大小均为 32，训练的 epoch 次数为 3。

4.3.3 实验结果

4.3.3.1 K-Means 聚类结果

使用 TF-IDF 将文本转为向量使用 K-Means 进行聚类，选取类别样本数排名前 5 的热点问题，具体结果如表 4-1 所示。

表 4-1 排名前五的类别样本

类别编号	类别样本量	热度指数
1	171	0.040
2	114	0.026
3	93	0.021
4	89	0.021
5	80	0.018

4.3.3.2 TextRank 结果

针对热度排名前五问题使用 TextRank 方法提取摘要。根据 TextRank 的原理，权重越大，在文本中就更具有代表性，因此在聚类的基础上进行摘要提取，选定权重最大的句子为该类的问题描述。提取结果如下表 4-2 所示。

表 4-2 排名前五的摘要

热度排名	留言详情
1	建议由政府牵头组织征收指挥部和该楼栋产权证托管部门...
2	我们于今年 9 年月底开业，到了十月我们所有商户找找老民百姓，让我们不再受这种商业陷阱的侵害...
3	业委会责令长城物业公 A 市相关扫黑部门进行举报目前相关部门正在调查当中...
4	40.50 的老人没有收入来源，60.70 年的没有养老...
5	尊敬的市领导，您好我是 A6 区山与墅小区的一名普通业主...

4.3.3.3 BERT+NER 实验结果

通过 BERT+NER 模型将 K-Means 聚出来的每一个类进行实体识别，这里我们选取每一个类中出现频率最高的 3 个实体作为结果，排名前五的热点问题的地址实体和人群实体如下表 4-3 所示。

表 4-3 排名前五的热点问题实体

类别编号	地址实体	人群实体
1	开发区，西地省, A5 区, 湖北, E 市	越卓,朱某某,司机, 公务员,余成
2	西地省,双凫铺镇, A5 区,A3,南托街道	朱华裕,楚雅,公务员,农民
3	银盆岭街道,长兴社区,西地省, A5 区	王明晓,公务员，董事长, 郭军建
4	合心村，龙街道，六路, 西地省，路口镇	黄明刚,蒋清，楚雅, 公务员，司机
5	C5 市，龙街道，西地省	万家,魏家,胡书记,工程师

综上所述，本文针对留言信息进行聚类，以各类样本的聚集程度作为热度指

数的评价指标，选取排名前 5 的热点问题，统计留言时间范围，抽取各类的地址和人群实体及摘要信息。最终结果如下表 4-4 所示。针对问题 ID 为 1 的第一类热点问题，选取前 5 条留言，具体明细如表 4-5 所示。

表 4-4 热点问题表

问题ID	热度排名	热度指数	留言时间	地址或人群	留言详情
1	1	0.040	2019-01-03至2020-01-06	西地省开发区	建议由政府牵头组织征收指挥部和该楼栋产权证托管部门，在充分了解以上房屋历史由来和目前该楼栋内居住户和使用人实际情况后，认真负责的通过协商出台《征收补偿方案》，合理、合法、合规的解决该楼栋所有征收补偿问题，切实保障产权人、集资单位、居住户、经营户的合法权益
2	2	0.026	2019-01-01至2020-01-05	西地省双凫铺镇	我们于今年 9 年月底开业，到了十月我们所有商户找找老板协商过，他说因为消防问题，材料运输问题政府报批时间差等等理由为由推掉了责任，还说会设计另一套装修方案，11 月底会全部整改完工，运营也会跟上，斑马线也会开通，但是到了 11 月底，装修，开通斑马线没有了音信，运营也只是单纯打折，根本没有投入资金，这个商场共有 30 几家商户，老板收取近千万资金，但根本没有投资商场，我们所有商户也都是被虚假的招商承诺诱骗才进来投资的，现在找老板协商却回复按合同办事，装修协议你们签了，就是默认了装修，合同里隐藏了一条说签定本合同后本合同将取代招商之前所有招商计划包括承诺，我们商户现在开业两个月都发生了严重亏损，但是缺少法律武器保护自己的权益，真心希望有领导能帮助我们这些弱势的良民百姓，让我们不再受这种商业陷阱的侵害
3	3	0.021	2019-01-01至2020-01-06	银盆岭街道长兴社区公务员	业委会责令长城物业公司限期离场后，长城物业公司对此置之不理，仍强行霸占本小区的物业管理权，并涉嫌威胁、恐吓业主，多家电视台对此进行了报道，该情况在 A 市内造成了较为严重的影响，针对长城物业此行为融圣国际业委会及广大业主已经向国家扫黑巡视组及西地省 A 市相关扫黑部门进行举报目前相关部门正在调查当中
4	4	0.021	2019-01-01至2020-01-07	合心村	40.50 的老人没有收入来源，60.70 年的没有养老保险跟保障，80.90 的找不到工作，这些都是 A 市现在政府扶持的好呀，全国 500 强闲城连这点事情都解决不好吗，一只拖拖拖，办事效率也不怎么样，我一个 90 后对 A7 县都失望了，哎

5	5	0.018	2019-01-15 至 2019-12-30	C5 市龙街道	尊敬的市领导，您好我是 A6 区山与墅小区的一名普通业主，但是我下面的诉求却是小区 848 户业主的共同诉求整个小区当前只有一个出入口，之前我们很多邻居也向政府反映过这个问题，小区东面规划为学校，与小区仅一墙之隔，希望政府能够想民之所想，急民之所急，开设东门，为我们至少预留一条人行通道出入，不管是对于小区的便利性还是以防发生火灾，居民能够快速疏散的作用都是十分必要的，之前政府也回复说会统筹考虑小区居民出入，小区 848 户业主也一直在殷切的期盼着，幻想着政府会如何考虑小区的出入，如何考虑民之所急，这种对于政府，对于 12345 市长信箱的信任日复一日，直到到今天听闻小区东面学校的围墙规划为直接顶着小区的围墙，完全不给小区居民丝毫活路，堵死了小区居民小孩的咫尺之隔上学距离，将小区区民置于绕道数公里、穿行车流间的不便与险境之中，暂时不知道政府是处于何种考虑，难道这个公立的学校是“个人”的学校

表 4-5 热点问题明细表

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	点赞数
1	188972	A00041922	A 市内道路坑洼多且深	2020/1/6 10:58:57	您好！第一，我是一名老司机，多年来，我跑遍了 A 市的主要马路，总的印像是：无论是主要干线，如 C5 市路、A1 区路，还是其它支线，下水道的水井盖不是低于路面较多，就是高于路面较多，车子从上面通过振动很大，既很心痛车子，又不安全。现代的科学技术发达无比，月球都可以上去了，难道还修不平一条马路吗。第二，A 市的公交车在上下班高峰非常拥挤，这段时间集中的人群主要是：上班族、学生、免费乘车的老人。如果象上海一样，不给老人免费乘车，但给老人发交通补贴，很多老人就不会“花钱找罪受”了，交通拥挤的状况一定会有所好转。愿 A 市越来越美丽。	0	0
1	189733	A0009754	A 市	2019/10/10	A 市的限购政策有效的控制了房	1	0

			限卖房产政策一刀切	17:09:47	价疯涨的局面，也得到了大多数老百姓的支持，但政策出台这么久也没有完善细节。政策规定房产需在产权证满 4 年后才能出售，这种方式太粗暴，侵犯了公民处置财产的正当权益。我是一个 80 后，我去年购置了一套房产，但今年我准备创业，急需大量资金，而我的所有财产都在这套房子上面，现在却不能出卖套现，导致我无法启动我的创业计划。我不是一个炒房者，却被政府一刀切的政策所害，我作为公民的合法权益谁来保障，政府这不是知法犯法吗？相信有我同样问题的普通老百姓也很多，急需资金（如创业、还债、投资、治病、留学等）却无法处置自己的合法财产。希望市委市政府市住建局能够替我们老百姓考虑，完善限购政策细节，还给我们一个处置财产的正当合法权益。		
1	190087	A00052076	A 市新奥燃气服务态度差	2019/7/22 0:01:47	我是新奥用户，燃气换表，却被告知需补缴七千元燃气费，奇怪的是我已十几年未住，房子（A 市农业银行小区一栋七单元 606）一直都是租户在住，后来终于搞清。燃气表具十几年来，可以不购买燃气，可以无限次透支使用，透支使用同时，燃气表也会正常转动，燃气表显示屏也会累加，也就是说，十几年来，租户不用买气，也不用插卡，可以无限使用，表具气量也会正常计数。明明燃气公司存在重大问题，却反而以不买气给我相要挟，蛮不讲理地必须先缴费，再能买气。我强烈恳请政府关注民生。1，这么热的天，不买气给老百姓，怎么办？2，新奥燃气如此大的漏洞，在明知表具有重大问题的情况下，十几年来，从未以任何形式通知或告知用户，哪怕短信、发信件都可以啊，当初安装时都	0	0

					登记用户信息的。3，多次找新奥协商，其大言不惭对用户说:不管你反映至哪里，都不怕。这到底哪里来的底气，就因为是垄断吗？4，既然不用买气，表具也可正常转动，质疑是不是冒用气时，也在计量呢？		
1	190957	A909140	西地省晨鹭互联网科技有限公司涉嫌电话窃听，窃取公民信息	2019/11/12 14:58:30	西地省晨鹭互联网科技有限公司涉嫌长期专业从事电话窃听，网络黑客非法窃取公民信息和盗卖公民信息，有时候甚至可以干扰和切断正常通话！有时候甚至可以病毒攻击导致电脑瘫痪！此恶劣行径严重涉嫌窃取和倒卖公民信息罪！窃取商业机密罪！希望领导能够严肃处理这个长期驻扎在 A 市 A5 区中意一路红星现代商务中心 c 栋 4048 的黑恶势力犯罪团队！还人民群众一个安居乐业，保障公民的信息安全，营造良好的营商环境，还社会一个风清气正！恳请领导严惩，通过高科技涉嫌犯罪的犯罪团伙!谢谢！	0	0

4.4 问题 3：答复意见评价

针对答复意见评价问题，本文从“答复意见”的完整性、相关性两个方面进行评价，最终给出三星、两星、一星三个评定等级。

4.4.1 流程图

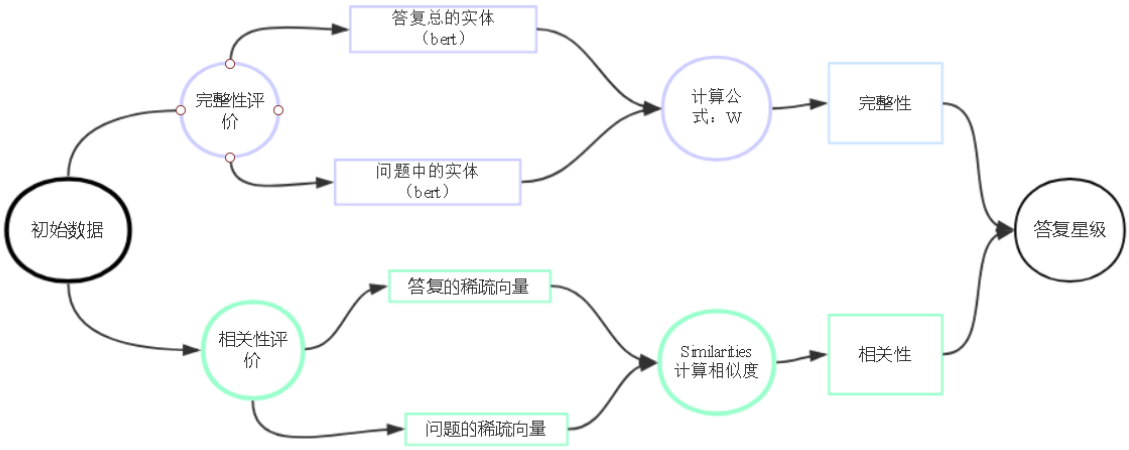

图 4-7 评价指标流程图

4.4.2 答复评价方案

关于答复意见的完整性、相关性、星级三个评价指标，具体评价规则如下：

相关性规则：基于留言详情与答复意见进行文本的相似度计算，相似度即为本文的相关性。

完整性规则：基于留言详情与复意见进行实体识别，两者均有的实体数量与留言详情中的实体数量比例即为本文的完整性。

星级：三个评定等级分别为优：3 星、良：2 星、差：1 星。初始星级为 1 星。若该答复与问题的相似度大于平均值，加一星。否则不加星。若实体比例大于平均值，加一星。否则不加星。

4.4.3 答复评价方案实现

1.相关性计算

利用 Python+gensim 进行文本相似度分析。流程如下：

①.生成分词列表。

②.基于文本集建立词典，获取特征数。

③.使用 dictionary.doc2bow 将匹配的文本转成稀疏向量，建立语料库。

④.将语料库传入 TfidfModel 训练 TF-IDF 模型。

⑤.使用 SparseMatrixSimilarity 进行相似度计算

相似度示例计算如下图 4-8 所示：

留言详情的稀疏向量集：
[(0, 0.15713484026367722), (1, 0.07856742013183861), (2, 0.15713484026367722), (3, 0.07856742013183861), (4, 0.07856742013183861), (5, 0.07856742013183861), (6, 0.07856742013183861), (7, 0.07856742013183861), (8, 0.07856742013183861), (9, 0.07856742013183861), (10, 0.07856742013183861), (11, 0.07856742013183861), (12, 0.07856742013183861), (13, 0.07856742013183861), (14, 0.07856742013183861), (15, 0.07856742013183861), (16, 0.23570226039551584), (17, 0.07856742013183861), (18, 0.07856742013183861), (19, 0.07856742013183861), (20, 0.23570226039551584), (21, 0.07856742013183861), (22, 0.07856742013183861), (23, 0.07856742013183861), (24, 0.07856742013183861), (25, 0.07856742013183861), (26, 0.07856742013183861), (27, 0.07856742013183861), (28, 0.07856742013183861), (29, 0.07856742013183861), (30, 0.07856742013183861), (31, 0.07856742013183861), (32, 0.07856742013183861), (33, 0.07856742013183861), (34, 0.07856742013183861), (35, 0.15713484026367722), (36, 0.078567420131838861), (37, 0.15713484026367722), (38, 0.07856742013183861), (39, 0.07856742013183861), (40, 0.07856742013183861), (41, 0.07856742013183861), (42, 0.3142696805273544 3), (43, 0.07856742013183861), (44, 0.07856742013183861), (45, 0.15713484026367722), (46, 0.07856742013183861), (47, 0.07856742013183861), (48, 0.07856742013183861), (49, 0.15713484026367722), (50, 0.07856742013183861), (51, 0.07856742013183861), (52, 0.15713484026367722), (53, 0.07856742013183861), (54, 0.07856742013183861), (55, 0.07856742013183861), (56, 0.15713484026367722), (57, 0.07856742013183861), (58, 0.549971940922870 3)]
[]

答复意见的稀疏向量：
[(2, 0.4181210050035454), (20, 0.08362420100070908), (35, 0.16724840200141816), (40, 0.16724840200141816), (41, 0.08362420100070908), (42, 0.08362420100070908), (45, 0.4181210050035454), (48, 0.08362420100070908), (58, 0.7526178090063816)]

相似度计算：
该答复与问题的相似度为： 0.6438736

图 4-8 相似度示例

2.完整性计算

将“答复意见”和“问题详情”进行人名、地点名、机构名实体抽取。

完整性计算公式:$w = (\frac{C}{A})$ (4-4)

其中，A:问题中的实体；B:答复中的实体；C:两者均出现的实体，如图 4-9。

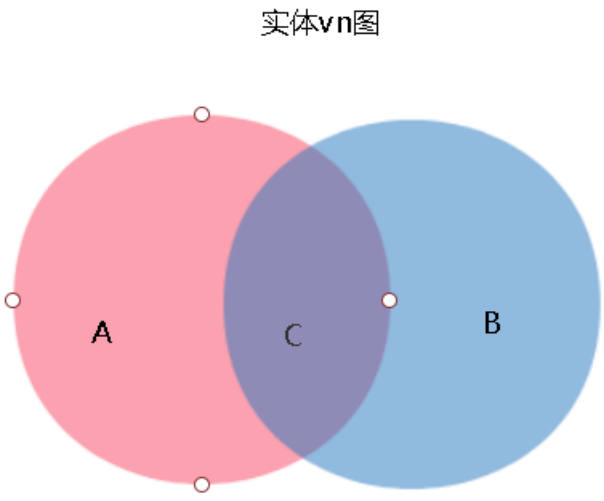

图 4-9 完整性评价示意图

4.4.4 答复意见评价表

本文答复意见表使用了 2816 条问答数据进行评价，依据三个评价指标所制的答复意见评价表的前 15 条结果如下表 4-8 所示。

表 4-8 答复意见表

留言用户	留言主题	留言详情	答复意见	相关性	完整性	星级
A00045581	A2区景蓉华苑物业管理有问题	2019 年 4 月以来，位于 A 市 A2 区桂花坪街道的 A2 区公安分局宿舍区（景蓉华苑)出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电改造造成物业公司的高昂水电费收取不了（原水电在小区买，水 4.23 一吨，电 0.64 一度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪？	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”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”,景蓉华苑业委会于 2019 年 4 月 10 日至 4 月 27 日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5 月 5 日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规	0.54460263	0.333333333	3

			召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019 年 5 月 9 日			
A00023583	A3区潇楚南路洋湖段怎么还没修好？	潇楚南路从 2018 年开始修，到现在都快一年了，路挖得稀烂用围栏围起，一直不怎么动工，有时候今天来台挖机挖两几下，过几天又来挖几下，对当地的交通和店面的生意带来很大影响，里面的车出去和外面的车进来要绕很大一个圈，很不方便，请有关部门对此监管一下，这路修的时间也太长了，至少可以一段一段的修好，方便街上的老百姓出行。	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 区洋湖街道高度重视，立即组织精干力量调查处理，现回复如下：您反映的为潇楚大道西线道路工程项目，该项目位处于坪塘老集镇，目前正在进行土方及排水施工。因该项目为城市次大道，设计标准高，该段原路基土质较差，需整体换填，且换填后还有三趟雨污水管道施工，施工难度较大，周期较长。加之坪塘集镇原有管线、排水渠道较多，需先处理管线和渠道才能进行道路施工，且因近期持续雨天，为保证道路施工质量，需在晴好天气才能正常施工。目前该项目已完成 75 土方及 50 排水，预计今年 8 月底将完工通车。感谢您对我们工作的关心、监督与支持。2019 年 4 月 29 日	0.4452881	0.5	3
A00031618	请加快提高A市民营幼儿园老师	地处省会 A 市民营幼儿园众多，小孩是祖国的未来，但民营幼儿园教师一直都是超负荷工作且收入又是所有行业最低，甚至连养老和医疗金都没交，在国家大力倡导普惠型幼儿园的同时更是加大了教师的工作压力，在降低成本的同时还增加了学生数量，让本来就喘不过气的教师更是	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：为了改善和提高民办幼儿园教师待遇，根据 2019 年 1 月 8 日出台的《中共 A 市委 A 市人民政府关于学前教育深化改革规范发展的实施意见》长发〔2019〕2 号文件精神，对于学前教育教师的培养和待遇问题做出了明确要求。一是在提高教师待遇方面，依法保障民办幼儿园教职工待	0.4083297	0.666666667	3

	的待遇	雪上加霜，希望市委市政府加快提高民办幼儿园教师工资待遇水平和降低工作压力有何具体政策和行动？	遇，民办幼儿园聘任教职工要依法签订劳动合同，依法缴纳城镇企业职工养老保险、医疗保险、生育保险、工伤保险、失业保险和住房公积金，民办园要参照当地公办园教师工资收入水平，合理确定相应教师的工资收入。二是加强监管协同推进，加强对民办幼儿园的日常监管和质量管理，保障民办幼儿园教师待遇，在完善人事（劳动）、工资待遇、社会保障和职称评聘等方面继续推进。谢谢您对我市学前教育的关注和支持！			
A00011 0735	在A市买公寓能享受人才新政购房补贴吗？	尊敬的书记：您好！我研究生毕业后根据人才新政落户A市，想买套公寓，请问购买公寓能否享受研究生3万元的购房补贴？谢谢。	网友“A000110735”：您好！您在平台《问政西地省》上的留言已悉悉，市住建局及时将您反映的问题交由市房屋交易管理中心办理。现将相关情况回复如下：按照《A市人才购房及购房补贴实施办法（试行）》第七条规定：新落户并在A市域内工作的全日制博士、硕士毕业生（不含机关事业单位在编人员），年龄35周岁以下（含），首次购房后，可分别申请6万元、3万元的购房补贴。“首次购房”是指在A市限购区域内首次购买商品住房（含住宅类公寓）。因此，如购买商业性质公寓（非商品住房），则不可申领购房补贴。以上情况，望您知晓和理解。如您还有疑问，建议可拨打市房屋交易管理中心咨询电话0000-00000000详询。特此回复！2019年4月30日	0.21041211	0.333333333	2
A00092 33	关于A市	建议将“白竹坡路口”更名为“马坡岭小	网友“A0009233”，您好，您的留言已悉悉，现将具体内容答复如下：关于来信人建议“白竹坡路口”更名为	0.26043034	0	1

	公交站点名称变更的建议	学”，原“马坡岭小学”取消，保留“马坡岭”	“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”的问题。公交站点的设置需要方便周边的市民出行，现有公交线路均使用该三处公交站站名，市民均已熟知，因此不宜变更。感谢来信人对我市公共交通的支持与关心。2019 年 5 月 5 日			

4.4.5 答复意见评价任务总结

本任务较为灵活，从完整性评价到相关性评价都是为了从不同角度来评判问答之间的关系。完整性评价利用了前文 BERT 实体识别的成果。

从结果来看，相关性和完整性的具体数值具有一定的依据，最后给出的星级则是更直观的展示了答复意见的质量优劣。

5 总结

总结本次比赛，本文最开始采用传统的文本分类特征抽取算法，但其效果并不理想。于是采用了 BERT 中文预训练模型进行群众留言特征抽取，最后发现其效果远比传统的文本分类特征抽取算法好，于是通过 BERT 训练出的模型对留言详情进行分类。

在做实体识别这个任务上，模型对数据的标注的要求上是非常高的，直接影响到了实体识别的整体效果，对此我们除了使用程序标注外，还进行了人工的标注，以此提升实体识别的效果。

在答复评价的任务上，从完整性和相关性等方面对答复意见进行评价，再由完整性和相关性综合评判得到星级，从而更加直观的表达出答复意见的优劣程度。

6 参考文献

[1]石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[2].Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[3]最小方差优化初始聚类中心的 K-means 算法[J].谢娟英,王艳娥.计算机工程. 2014(08).

[4]王子牛,姜猛,高建瓴,陈娅先.基于 BERT 的中文命名实体识别方法[J].计算机科学,2019,46(S2):138-142.

[5]李娜娜.基于 TextRank 的文本自动摘要研究[D].山东师范大学,2019.