“智慧政务”中的文本挖掘应用

摘 要

随着互联网的不断发展和应用，人们的生产生活方式发生了深刻地改变。在“智慧政务”问题中，许许多多需要文本挖掘的领域逐步展现出来，如何利用文本挖掘与“智慧政务”结合，成为了亟待解决的问题。本文主要研究的是针对“智慧政务”中的文本挖掘应用问题。首先对数据进行预处理，得到适合模型使用格式的数据。然后针对给出分类类别进行有监督训练，得到一个精度较高的分类模型。再根据留言对提取留言中的主体，得到留言中的热点问题。最后利用双语评估替补模型对答复意见进行评价，给出一套合适的答复意见评价方案。

针对问题一，本文对附件 2 中的留言数据进行预处理，将留言详情中存在的停用词进行过滤，然后使用 jieba 分词对每一句话进行分词处理并使用 Doc2vec 对分词内容向量化，方便做分类处理。最后利用 Doc2Vec 获得的句向量对 LightGBM 模型进行分类训练，得到能够对留言精确分类的模型，利用 F-Score 方法计算训练集与测试集的得分，得到训练集的得分为 0.462，测试集的得分为 0.368，能够在一定程度上为留言数据进行分类。

针对问题二，首先对附件 3 中的留言数据进行预处理，同样过滤其中的停用词并使用 jieba 分词对其进行分词处理。然后利用潜在狄利克雷分布提取所有数据中体现出的主题，根据每个主题的中留言的数量确定留言中的热点问题，最终得出排名前五的热点问题并统计主题的内容。

针对问题三，与上文类似，对附件 4 中的留言详情与答复意见分别进行去除停用词和分词处理，然后利用双语评估替补模型对留言详情和答复意见的相似度进行计算，从而确定答复意见的质量。

本文利用“智慧政务”中的留言数据和答复数据，对问题的多个角度进行分析，得到了能够运用到实际问题中的模型和方法。

关键词：jieba 分词；Doc2vec；LightGBM；F-Score；潜在狄利克雷分布；双语评估替补模型

Text Mining Application in "Intelligent government Affairs"

Abstract

With the development and application of the Internet, people's production and life style has changed profoundly. In the problem of "intelligent government affairs", many fields that need text mining gradually show up. Use the combination of text mining and "intelligent government affairs" has become an urgent problem. This paper focuses on the application of text mining in "intelligent government affairs". Firstly, the data was preprocessed to get the data suitable for the model. Then the supervised training was carried out for the given classification, and a classification model with high accuracy was obtained. Next, according to the message pairs, the main body of the message was extracted to get the hot issues in the message. At last, this paper used the alternative model of bilingual evaluation understudy to evaluate the response, and given a set of appropriate evaluation scheme.

In order to solve the problem one, this paper preprocessed the message data in annex 2, filtered the stop words in the message details. Then Jieba segmentation was used to segment each sentence and doc2vec was used to vectorize the word segmentation content, which was convenient for classification processing. Finally, the LightGBM model was trained by the sentence vector obtained by doc2vec, and the model can accurately classify messages was obtained. The scores of training set and test set were calculated by F-score method, and the scores of training set and test set were 0.462 and 0.368, which can classify message data to a certain extent.

So as to solve the problem two, this paper preprocessed the message data in annex 3, the stop words was filterd and the Jieba segmentation was used to segment them. Then, this paper used the Latent Dirichlet Allocation to extract the theme reflected in all the data. The hot issues in the message was determined according to the number of messages in each topic, and finally the top five hot issues and statistical content of the theme was got.

To solve the problem three, similarily, the details of the message and the reply in annex 4 were processed by removing the stop words and word segmentation respectively. Then the similarity between the details of the message and the reply was calculated by Bilingual Evaluation Understudy model, and the quality of the reply was determined.

In this paper, the message data and reply data in the "intelligent government affairs" was used to analyze the problems from multiple perspectives, and the models and methods that can be applied to the actual problems was got.

Keywords: Jieba segmentation; Foc2vec; LightGBM; F-score; Linear Discriminant Analysis; Bilingual Evaluation Understudy model

目 录

一、挖掘目标

随着信息技术的不断发展，互联网成为生活中必不可少的一个部分。在当今社会中，各类社会民意的投诉留言成为了向相关部门反映社会现象的重要渠道。虽然这个便捷的渠道为社会带来了极大的便利，但是也使得相关部门的工作量急剧提升[1]。因此，利用数据挖掘替代人工解决部分工作成为了现在亟待解决的一个问题。

本文利用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等数据，利用 jieba 分词、Doc2Vec 方法、LightGBM 算法，潜在狄利克雷分布（Latent Dirichlet Allocation,LDA）算法以及双语评估替补（Bilingual Evaluation Understudy,BLEU）方法达到以下三个目标：

(1) 利用 jieba 分词对文本数据进行分词，并使用 Doc2Vec 将文本数据转换为句向量，最后使用 LightGBM 对数据进行分类，之后结合留言内容可以为留言自动进行分类，能够减少工作量并且提高工作效率。

(2) 利用 jieba 分词对文本数据进行分词，再使用 LDA 提取文本中的主题，从而实现对热点问题的挖掘，得到群众留言的热点问题。

(3) 同样使用 jieba 分词对文本数据进行分词，再利用 BLEU 对留言和答复意见进行相似度比较，得到答复意见的评级方案。

二、分析方法与过程

本用例主要包括如下步骤（如图一所示）：

步骤一：数据预处理，对附件二、附件三、附件四中留言详情部分以及附件四中答复意见部分过滤停用词，再利用 jieba 分词进行处理。

步骤二：文本向量化，在对附件二的留言详情进行分词后，利用 Doc2Vec 将文本数据转换为句向量。

步骤三：文本数据分类，使用步骤二中得到的句向量对 LightGBM 模型进行

训练，并利用 F-Score 对模型的分类结果进行评价。

步骤四：热点问题挖掘，利用 LDA 提取主题词，针对附件三中留言详情部分分别得出每个主题的留言数量、主题内容以及每条留言归属的主题，按照每个主题的留言数量提取热点问题，对于因分类个数接近而无法提取热点的内容用点赞数与反对数作为附加内容提取。

步骤五：答复意见评价，利用 BLEU 算法对对附件四中留言详情和答复意见部分的进行相似度比较，从而得到对答复意见的质量的评价方案。

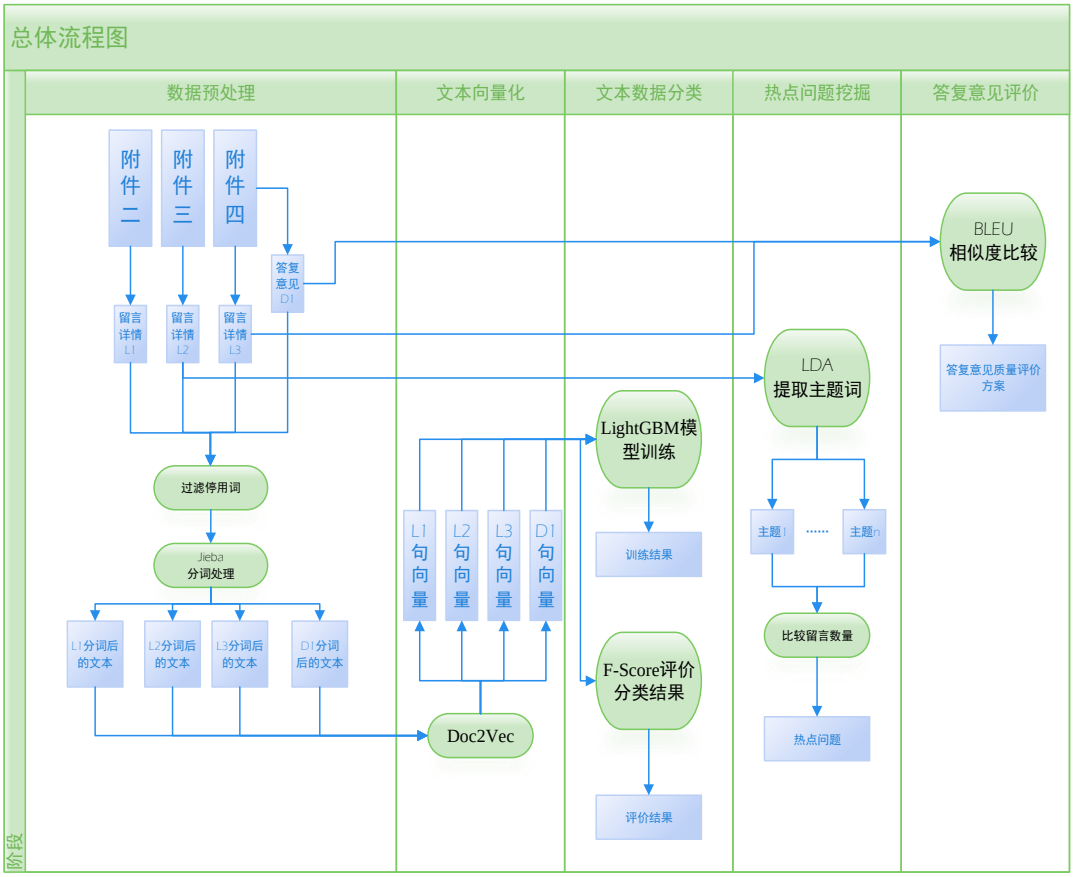

图 1 主要步骤

2.1 问题一的分析及方法

2.1.1 问题分析

对于群众留言分类问题，首先需要考虑数据的预处理问题。在自然语言处理（Natural Language Processing, NLP）中，文本的预处理操作，是在所有任务中都必不可少的一项重要操作，预处理操作完成的结果优劣将直接影响数据分析的准

确性。预处理操作一般会在输入模型前，对数据进行停用词去除、句子分词、向量化以及去除噪声词等操作。对于本文而言，首先需要对群众留言数据进行分词操作，对于已经分词过后的句子，可以对其中多余的停用词进行去除，能够有效的减少模型的计算量以及提高预测的准确度。在完成对数据的处理后，可以对文本进行向量化，本文使用 Doc2Vec 得到了每一条留言的向量表示形式。由于本文中的使用的句向量较为复杂，且其中包含了丰富的信息，为了得到较为准确的分类结果，本文使用留言数据对 LightGBM 模型进行训练。

2.1.2 数据预处理

(1) 停用词处理

停用词是指当执行关于完成信息检索任务的程序任务中，为减少程序运行的时间复杂度即空间复杂度，节约运行资源，提高搜索效率，程序会在处理自然语言数据或自然语言文本之前，有时也在处理自然语言数据或自然语言文本之后，完成自动过滤掉某些字或词语的操作，这些字或词即被称为停用词（Stop Words）。大多数情况下，停用词是通过人工输入的方式生成而非自动生成的，随着停用词的不断生成，会同时形成一个停用词表。但从实际情况来看，至今还没又出现一个停用词表能够做到对所有的工具及算法都适用。不仅如此，有一些工具明确提出避免使用停用词来支持短语搜索。停用词是衔接各个词汇的功能词，在评论的文本中也有很多，停用词很少表达具体的信息，本身含义并不重要，停用词在文本分类中也不会造成明显影响，甚至可能会带来计算负担，去除停用词也是数据预处理非常重要的一步。停用词的过滤可以有效地降低待处理文本的维度，剔除无用的词汇，保留可以表示文本内容的词语，从而提高词语搜索的效率[2]。

(2) jieba 分词

根据分词算法的核心思想，可将分词算法大体分为两种：基于词典的分词以及基于字的分词。基于词典的分词，即先按照字典将句子切分成词，再通过其他方式得到一种词的最佳组合方式；而基于字的分词，即是由字构词，首先将句子划分为多个字的排列，再将所划分出来的每一个字组合成词，其核心思想是寻找最优的句子切分策略，也可转化为序列标注问题。Jieba 分词即是一种基于词典的分词算法。jieba 库的分词是利用一个中文词库，将待分词的内容与分词词库

进行比对，通过图结构和动态规划方法找到最大概率的词组[3]。jieba 分词首先通过对照典生成句子的有向无环图，再根据词典利用动态规划寻找最短路径后，对句子进行截取并获取分词结果，对于未登录词，jieba 分词采用基于汉字成词能力的隐马尔可夫模型（HMM）对新词进行处理，具体流程如图 2 所示。

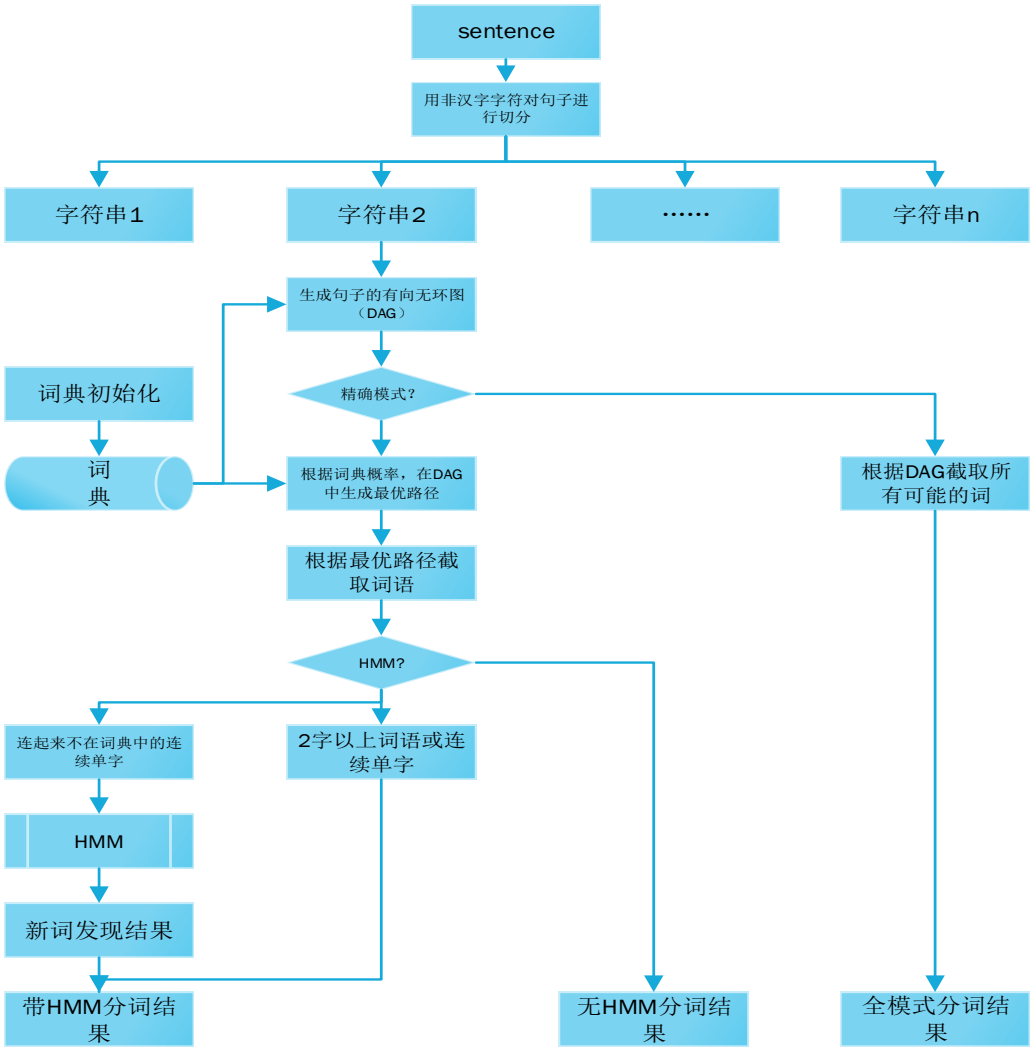

图 2 jieba 分词的操作流程

2.1.3 Doc2Vec 模型

对数据集进行分词后，需要用 Doc2Vec 模型将文本数据向量化，Doc2Vec 是基于 Word2Vec 模型的思想基础提出的，Word2Vec 的基本思想是：根据上下文的单词来预测当前词的概率。在这个框架中，每个单词都被映射成一个向量，映射后的单词向量被用作单词矩阵 W 的输入矩阵，列由短语中的单词索引指定。这些词向量用于预测句子中的下个单词[4]。假设给出上下文单词用于训练，词向

量模型的目标是根据已知单词对数概率平均值的最大值预测未知单词的概率。

给定一个用于训练的词序列 $w_1, w_2, \cdots, w_T$ ，定义词向量模型的目标函数：

$$\frac{1}{T}\sum_{t=k}^{T-K}\log p\left(w_t \middle| w_{t-k}, \cdots, w_{t+k}\right)$$

再利用多分类器完成预测任务。

而 Doc2Vec 算法是从大型原始数据中以完全无监督的方式进行训练，而无需任何针对于特定任务的标记数据。Doc2vec 在此基础上增加了一个与词向量长度相等的段落向量，该向量具有固定长度，它不仅加入了文本语义信息，同时具有更好的泛化能力[5]。Doc2vec 有两种模型：Distributed Bag Of Words（DBOW）和 Distributed Memory（DM）。

本文使用 DBOW 模型，该模型会忽略输入的上下文，使得模型预测段落中随机一个单词。在每次迭代的时候，从文本中采样得到一个窗口，再从这个窗口中随机采样一个单词作为预测任务，并将该单词作为输出值，而输入值为段落向量。因此，在训练单词向量 W 的同时，模型还训练了文档向量 D 。该模型的结构如图 3 所示。

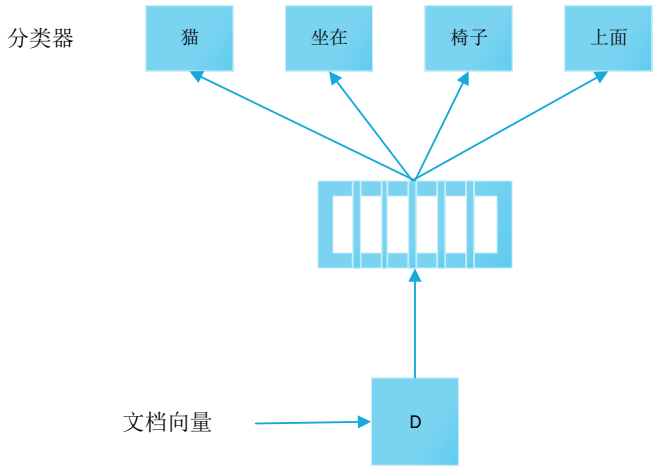

图 3 DBOW 结构示意图

2.1.4 LightGBM 模型

与传统的梯度提升迭代决策树(Gradient Boosting Decision Tree, GBDT)相比，LightGBM(LGB)在 GBDT 的基础上引入了两个新技术：梯度单边采样技术(Gradient-based One-Side Sampling，GOSS)和独立特征合并技术(Exclusive Feature Bundling, EFB)[6]。GOSS 可以剔除很大一部分梯度很小的数据，只使用剩余的数

据来估计信息增益，从而避免低梯度长尾部分的影响。由于梯度大的数据对信息增益更加重要，所以 GOSS 技术在较传统 GBDT 少很多的数据前提下仍然可以取得相当准确的估计值。EFB 实现互斥特征的捆绑，以减少特征的数量。

LGB 通常使用按层分裂的学习方法(Level-wise Learning)，按层分裂的学习方法每次在所有的叶子节点中，经过计算选择信息增益最大的一个叶子节点进行分裂，一直循环直到符合一定条件（如图 4 所示）[7]。这样的操作方式能够在多线程的条件下，加速精确贪心算法，能够很快地训练出符合数据的模型。

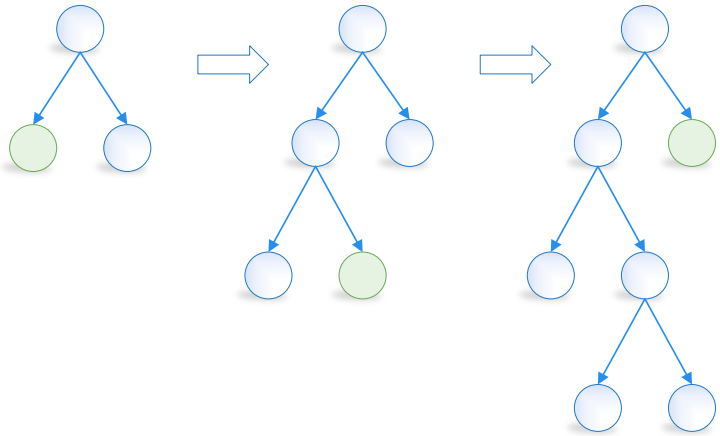

图 4 Level-wise 生长策略

2.1.5 F-Score

F-Score 是一种评价模型分类效果的指标，它不仅考虑到了模型分类的准确性，还有效的考虑到了准确率、精确率、召回率等指数的影响。本文使用 F-Score 评价指标来评价模型的分类效果，首先定义分类问题中的四种情况[8]（如表 1 所示）。

表 1 分类列联表

		真实值	
		p	n
预测值	p	TP	FP
	n	FN	TN

F-Score 的计算公式为：

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

其中，P_i为第i类的查准率，其计算公式为：

$$P_i = \frac{TP}{TP + FP}$$

R_i为第i类的查全率，其计算公式为：

$$R_i = \frac{TP}{TP + FN}$$

查准率主要表示了真正为p的样本占所有预测为p样本的比例，而查准率则表示了真正为p的样本占所有实际为p样本的比例。

2.2 问题二的分析及方法

2.2.1 问题分析

对于问题二而言，需要分析留言内容中的热点问题。热点本身是在一系列的内容中，表现较多的一个群体，而在统计学中，可以被理解为数量较多的一类。在对数据进行大规模的分类过后，得到类中数量较多的一类问题的集合，即为热点问题。本文使用 LDA 模型对留言数据进行主题提取，将所有数据区分到对应于不同主题的类别中，得到每一条留言的主题，根据每个主题中所包含的留言条数以及留言的点赞数和反对数，可以确定出在该文本下热点内容以及热点程度。

2.2.2 LDA 模型

LDA 是一种采用词袋模型的主题生成模型，LDA 同样是一种三层贝叶斯概率模型，这三层分别为词、主题以及文档，该模型主要利用先验分布对数据进行似然估计，然后得到后验分布的一种方式[9]。它的基本思想是假设所有的文档存在K个隐藏主题，每一篇文档的每个词都是以一定概率确定了某个主题[10]，并从这个主题中按照一定的概率选择某个词语，不断抽取隐含主题及其特征词，直到将所有词语都归到确定的主题中。

LDA 假设文档中主题的先验分布是 Dirichlet 分布，即对于任一文档d，其主题分布θ_d为：

$$\theta_d = Dirichlet(\vec{\alpha})$$

其中，α 为一个 K 维向量，代表分布的超参数。

同样，LDA 假设主题中词的分布是 Dirichlet 分布，即对于任一主题 k，其词分布 β_k 为：

$$\beta_k = Dirichlet(\vec{\eta})$$

其中，η 为一个 V 维向量，代表分布的超参数。

对于数据中任意一篇文章 d 中的第 n 个词，则它的主题编号 z_{dn} 的分布为：

$$z_{dn} = multi(\theta_d)$$

则词 w_{dn} 的分布为：

$$w_{dn} = multi(\beta_{z_{dn}})$$

最后根据后验概率，即可得到每个文档所归属的主题以及每个主题所包含的词语和概率。

2.3 问题三的分析及方法

2.3.1 问题分析

针对问题三，需要对答复意见的质量给出一套评价。在 NLP 中，对于答复意见的评价大多是与人工回复进行比较，在本文中，无法确定最优的回复，因此，可以对留言内容和答复内容进行比较，从而获得答复意见的评价。在本文中，使用 BLEU 方法对留言内容和答复内容进行比较，能够得到两个内容的相似程度，内容越相似，说明答复越贴近留言内容，即回复评价较高。

2.3.2 BLEU 方法

BLEU 意思是代替人进行翻译结果的评估。尽管这项指标是为翻译而发明的，但它可以用于评估一组自然语言处理任务生成的文本[11]。在本文中我们用这个方法来计算留言详情与答复意见之间的相似度，进而对答复意见进行评价。

它的取值范围在 0 到 1 之间, 如果两个句子完美匹配, 那么 BLEU 是 1, 反之, 如果两个句子完全不匹配, 那么 BLEU 为 0。

在对单个单词之间的相似度进行计算时，可以看作时 1-gram，1-gram 描述的是逐字对照的能力，但同时我们也要关注对应的流畅性，因此引入了 n-gram[12]，在这里一般 n 不大于 4。BLEU 公式的推导过程如下:

假设参考文档的数量为 M，长度为 n 的 $gram$ 中的一个组合为 W_{n_i}，将当前组合 W_{n_i} 在句子中出现的次数记做 $Count_{W_{n_i}}$，同时计算出这个单词在第 j 个参考文档中出现的次数, 用符号 $\mathrm{Re}\, f_{W_{n_i}}^{j}$ 表示其中 $j=0,1,2,\cdots,M$，由于参考文档有 M 个, 选择这 M 个参考文档取值中的最大值进行下一步计算[13]。记这个最大值为 $\mathrm{Re}\, f_{W_{n_i}}^{\max}$，即：

$$\mathrm{Re}\, f_{W_{n_i}}^{\max}=\max\left(\mathrm{Re}\, f_{W_{n_i}}^{j}\right)\quad j=0,1,2,\cdots,M-1$$

利用已经获取到了某一个长度中一种 $gram$ 的 $Count_{W_{n_i}}$ 和 $\mathrm{Re}\, f_{W_{n_i}}^{\max}$ 中选择计算值中的最小值进行下一步计算, 记为 $Count_{W_{n_i}}^{res}$，即：

$$Count_{W_{n_i}}^{res}=\min\left(\mathrm{Count}_{\mathrm{W_{n_i}}},\mathrm{Re}\, f_{\mathrm{W_{n_i}}}^{\max}\right)$$

定义长度为 n 的 $gram$ 类型数为 K，则长度为 n 的 $gram$ 的最终精度计算公式为:

$$P_n=\frac{\sum_{i=0}^{K-1}Count_{W_{n_i}}^{res}}{\sum_{i=0}^{K-1}Count_{W_{n_i}}}$$

最后，利用计算出的精度，计算得到几何平均精度：

$$P_{avg}=e^{\frac{\sum_{n=0}^{N-1}\ln P_n}{N}}$$

三、结果分析

3.1 问题一的结果分析

在对附件 1 进行基本的统计工作后发现，以及标签共有 15 种，同时也是在做分类时所需确定的分类个数。

在附件 2 中，共包含了 9210 条数据，由于训练过程中需要避免过拟合的问题，本文将从数据集中随机抽取了 6907 条数据作为模型的训练集，其余 2303 条数据作为模型的测试集。利用上述数据对 LightGBM 进行训练，得到了一个分类个数为 15，叶子具有最小记录数为 10，切分最小收益为 0.05，节点分裂的最小 Hession 值之和为 5，学习率为 0.05，决策树的深度为 50 且复杂程度为 120 的 LightGBM 模型，同时控制模型在每 5 次迭代之后执行 1 次 bagging。

在利用 F-Score 对该模型进行评价时，本文在确定随机种子之后分别计算了训练集和测试集条件下的得分。由于训练集更贴近于模型，因此在将训练集投入模型后，得到 F-Score 为 0.462，而将测试集投入模型后，得到 F-Score 为 0.368。

3.2 问题二的结果分析

在附件 3 中，共有 4326 条留言数据，去除停用词后利用 jieba 分词对数据进行预处理，再结合 LDA 算法对这些数据提取文本的主题，能够有效的为每个留言划分主题且提取出关注较高的热点问题。本文从留言数据中提取 150 个主题，其中每个主题中包括 10 个关键词，得出每个主题的内容和分类个数统计并将热度指数定义为：

$$\text{热度指数}=\frac{\text{主体数量}}{\text{数据总量}}$$

根据主题的热度，提取热度前 5 的主题内容作为所有留言的热点问题，其中前 5 类主题的内容如表 2 所示：

表 2 热点问题主题词

类别	关键词
1	"小区"，"解决"，"影响"，"部门"，"居民"，"晚上"，"生活"，"投诉"，"相关"，"休息"
2	"出行"，"公交车"，"马路"，"上学"，"线路"，"孩子"，"时间"，"路"，"上班"，"行人"
3	"开发商"，"业主"，"交房"，"政府"，"楼盘"，"栋"，"收房"，"未"，"开发"，"月"
4	"路"，"道路"，"大道"，"路口"，"车辆"，"车道"，"道"，"一条"，"通行"，"直行"
5	"业主"，"小区"，"物业"，"物业公司"，"业委会"，"开发商"，"小区业主"，"公共"，"情况"，"政府"

根据这些关键词可以确定群众反映的热点问题，可以发现第一个主题是关于晚上噪音影响居民休息，第二类主题是关于交通问题，第三类主题是关于开发商和业主矛盾问题，第四类主题是关于道路交通问题，第五类主题是关于社区管理问题。根据热点问题留言明细表（表 3）就可以得出问题的实际情况，可以更高效地解决群众反映的问题。

表 3 热点问题描述

热度排名	热度指数	时间范围	地点/人群	问题描述
1	0.055	2019/09/04—2020/1/5	A7 区	A7 区晚上施工噪音扰民问题
2	0.030	2019/1/11—2020/1/4	A7 区、A2 区、A5 区	A7 区、A2 区、A5 区交通问题严重，影响居民及师生出行
3	0.027	2019/07/28—2020/1/5	A 市	A 市多地开发商欺骗业主
4	0.026	2019/1/1—2020/1/2	A7 区	A7 区道路需要整改
5	0.025	2019/1/2—2020/1/6	青园社区	青园社区管理问题

由上述表格可以发现排名前五的热点问题的内容，即噪音扰民，交通问题，

业主维权，道路整改，物业管理这些问题。噪音扰民现象与业主维权现象属于阶段性的现象，群众反应长达 4-5 个月，就以上例子而言，时间范围较大且热度较高，说明有关部门并没有及时的解决生活中的问题。而交通问题，道路整改，物业管理问题是持续性的现象，群众反映长达一年，这些现象并不能在短时间内彻底根除，表示此类问题会经常发生在生活中，需要相关部门进一步关注。

将不同的热点的数量展现在图 5，可以发现，群众对于噪音扰民，交通问题，道路整改的现象反映的比较大，说明这种现象已经扰乱了居民的日常生活，给群众的生活造成了一定影响。而业主维权、物业管理等涉及到法律方面的问题，有关部门同时需要及时解决，避免造成更大的损失和影响。

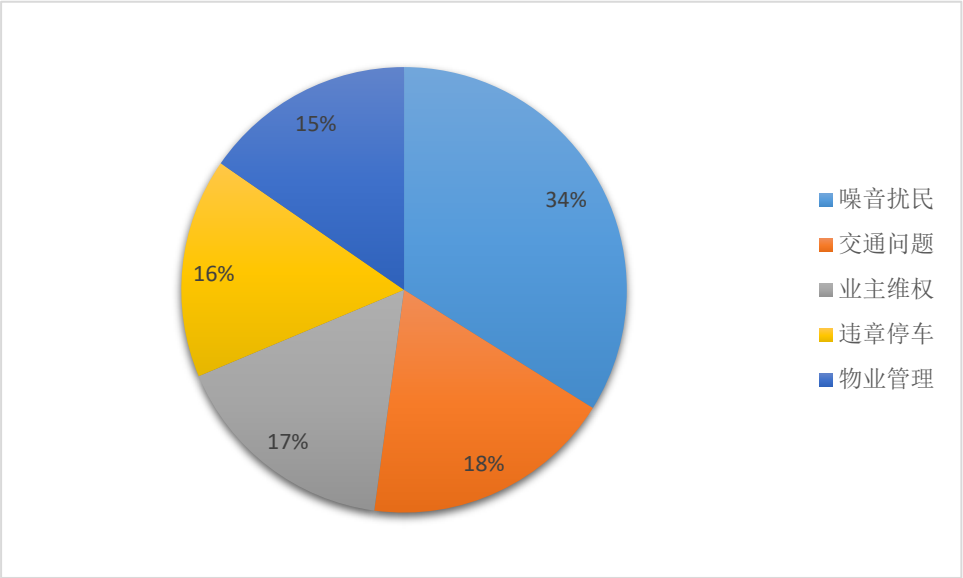

图 5 热点问题饼状图

3. 3 问题三的结果分析

在附件 4 中，共有 2816 条留言与答复数据。对数据去除停用词后利用 jieba 分词对数据进行预处理，再利用 BLEU 方法计算留言与答复意见之间的相似度，得到每一条答复数据的评价得分，由于篇幅原因，抽取部分评价结果如表 4 所示：

表 4　BLEU 计算结果

留言编号	相似度
2549	9.40e-79
2554	3.75e-232
2555	0.03
2557	1.31e-79
2574	0.10
2759	7.63e-232
2849	2.80e-155
3681	3.24e-79
3683	1.35e-155
3684	2.86e-155

其中，相似度越高，代表答复意见与留言越相近，即答复意见评价较好，反之，则答复意见评价较差。

可以发现，根据相似度计算得到评分较高的答复意见更能够贴近留言的内容，而得分较低的答复意见，多数为没有抓住留言的重点或者较为笼统的回答。

四、结论

本文在对数据进行预处理后，首先使用 Doc2vec 将文本数据转化为句向量。为了将留言数据有效的分为一级分类，本文使用上文计算所得的句向量对 LightGBM 模型进行训练，并使用 F-Score 对分类结果进行评价，从而得到高精度的分类模型。除此之外，本文使用 LDA 对大量文本进行主题分析，将每一条留言划分到所属主题，并对每个主题进行留言数量的统计，得到留言数量较多的主题作为所有文本的热点问题。最后通过 BLEU 算法计算留言与答复意见之间的相似度，从而对答复意见进行评价，得到答复意见是否符合留言内容。

本文利用 NLP 中的相关算法，结合机器学习中表现较好的算法，成功的实现了对留言数据的分类、提取热点以及对答复意见的评价，并得到了较高的精确度。但是 NLP 和机器学习的发展远不止于此，本文在计算过程中，虽然得到了非常满意的效果，但是仍然无法做到保证所有结果完全正确，这代表着机器学习领域还有较大的发展空间，也预示着未来会有更多更优秀的算法来解决我们生活中的实际问题，相信在未来会有更加优秀的模型能够为社会带来更大的贡献。

参考文献

[1]饶守艳.智慧政务提升政务效能的理论与实践[J].技术经济与管理研究,2016(05):89-93.

[2]高巍,孙盼盼,李大舟.Twitter 情感分析中停用词处理[J].计算机工程与设计,2019,40(11):3180-3185+3191.

[3]徐博龙.应用 Jieba 和 Wordcloud 库的词云设计与优化[J].福建电脑,2019,35(06):25-28.

[4]申远,黄志良,胡彪,王适之.基于 Doc2Vec 和深度神经网络的战场态势智能推送研究[J].智能计算机与应用,2020,10(01):50-55.

[5]李峰,柯伟扬,盛磊,陈雯,陈丙赛,罗韵晴.Doc2vec 在政策文本分类中的应用研究[J].软件,2019,40(08):76-78.

[6]Ke G, Meng Q, Finley T W, et al. LightGBM: a highly efficient gradient boosting decision tree[C]. neural information processing systems, 2017: 3149-3157.

[7]王建成,蔡延光.LightGBM 算法和 ARIMA 算法在人口流动预测应用的性能的比较[J].东莞理工学院学报,2019,26(05):27-32+44.

[8]薛博. 基于 FOA-SVM 的中文文本分类的研究[D].河北工业大学,2014.

[9]Blei D M, Ng A Y, Jordan M I, et al. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003: 993-1022.

[10]张卫卫,胡亚琦,翟广宇,刘志鹏.基于 LDA 模型和 Doc2vec 的学术摘要聚类方法[J].计算机工程与应用,2020,56(06):180-185.

[11]Papineni K, Roukos S, Ward T, et al. Bleu: a Method for Automatic Evaluation of Machine Translation[C]. meeting of the association for computational linguistics, 2002: 311-318.

[12]叶绍林,郭武.基于句子级 BLEU 指标挑选数据的半监督神经机器翻译[J].模式识别与人工智能,2017,30(10):937-942.

[13]刘世兴.基于多尺度的 n-grams 特征选择加权及匹配算法[J].智能计算机与应用,2020,10(01):61-66.