

“智慧政务”中的文本挖掘应用

摘要

在倡导国家治理体系现代化的当下，如何利用大数据、机器学习和人工智能为政府的治理能力“赋能”是政务工作的核心所在。通过“智慧政务”平台的建设来提高政府决策能力、公共服务水平以及社会民生保障，也是实现政府服务体系智能化的大势所趋。因此，本文通过自然语言处理文本挖掘技术解决问政平台的群众留言分类、热点问题挖掘以及政府答复意见评价等问题，为政府的治理能力提升提供可行的方案。

本文的解题过程主要分为以下几个步骤：

首先，我们对原数据进行文本预处理工作，包括剔除非文本信息、处理中文编码问题、运用 python 中的 jieba 分词工具对文本进行分词处理、去除停用词，最后运用 scikit-learn 库进行文本向量化，特征提取 TF-IDF 以及标准化。

其次，我们运用经过预处理后的数据对三个问题进行建模分析。

对于问题 1，我们根据附件 2 中已经标注好的数据作为训练数据，建立关于留言内容的一级标签分类模型。具体而言，我们分别运用深度学习算法 LSTM 和 Bi-LSTM+Attention，机器学习算法的 Logistics 回归、朴素贝叶斯这四类算法建立一级标签分类模型，根据 F-Score 对分类效果进行评分，结果发现，四类算法建立的模型的精准度、召回率和 F-Score 均在 0.8 以上，具有较好的效果，其中，Bi-LSTM+Attention 算法表现最好，比普通 LSTM 算法的 F-Score 高出 9.5%，达到 0.92，因此我们最终选择 Bi-LSTM+Attention 算法作为问题一的分类模型。

对于问题 2，我们对附件 3 的留言详情进行聚类分析，具体而言，我们利用 DBSCAN 聚类分析方法建立留言归类模型，排除噪音点和样本数阈值较小的留言主题。接着我们从分类结果中度量出热点问题的指标，包括点赞数、反对数、问题持续时间，问题提出频次等，并建立 AHP-EWM 模型进行评价，最后输出排名前 5 的热点问题分别为（1）A 市伊景园滨河苑捆绑车位销售；（2）A2 区丽发新城附近建搅拌站噪音扰民；（3）A 市辉煌国际城二期商铺非法营业；（4）A 市经济学院强制学生实习；（5）A 市加快基础设施建设全覆盖。

对于问题 3，我们分别统计了答复意见与问题详情的余弦相关性、完整性、可解释性以及答复及时性，并运用 TOPSIS 综合评价模型对答复意见进行评价，结果显示，政府答复意见的完整性和相关性方面都做得令人满意，但答复民众问题的不够及时仍然是日后需要继续改进的。

本文的研究结果对于地方政府在政务工作的效率以及服务民众水平的提高方面具有较为重要的理论和实践意义。

关键词： 智慧政务 LSTM Bi-LSTM+Attention DBSCAN 聚类 AHP-EWM 法

Application of text mining in "intelligent government affairs"

Abstract

In the current era of advocating the modernization of national governance system, how to use big data, machine learning and artificial intelligence to "empower" the governance ability of the government is the core of government affairs. The construction of "intelligent government affairs" platform to improve the government's decision-making ability, public service level and social livelihood security is also the general trend of realizing the intelligent government service system. Therefore, this paper USES natural language processing text mining technology to solve such problems as the classification of people's comments on the political platform, the mining of hot issues and the evaluation of the government's reply opinions, so as to provide a feasible plan for the improvement of the government's governance ability.

The problem solving process of this paper is mainly divided into the following steps:

First, we preprocessed the original data with text, including eliminating non-text information, dealing with Chinese coding problems, using jieba word segmentation tool in python to segment the text and remove stop words, and finally using scikit-learn library for text vectomization, feature extraction tf-idf and standardization.

Secondly, we use the pre-processed data to model and analyze the three problems.

For question 1, we set up a first-level label classification model for message content based on the data already marked in attachment 2 as training data. Specifically, we used Logistics regression, bayesian estimation, and deep learning model LSTM to establish a first-level tag classification model, and scored the classification results according to f-score. The accuracy of the model, the accuracy rate, the recall rate, and the f-score all achieved high scores.

For question 2, we conducted cluster analysis on the message details in annex 3. Specifically, we established a message categorization model by using Kmeans clustering method. However, since this method has a strong subjectivity of classification, We further used DBSCAN clustering analysis method to improve the model, and the output results divided the message into? Class. Then we measured the indicators of hot issues from the classification results, including thumb up number, opposition number, problem duration, problem proposing frequency, etc., and established the TOPSIS comprehensive evaluation model for evaluation. Finally, the top 5 hot issues were output as... .

For question 3, we counted respectively reply opinion and details of the cosine correlation timeliness, completeness, interpretability and reply, and use AHP to evaluate reply opinion EWM method, the results show that the integrity of the government response opinion and correlation are doing satisfactory, but people answer questions in time is still need to continue to improve in the future.

The results of this paper have important theoretical and practical significance for local governments in improving the efficiency of government affairs and the level of serving the public.

Key words: Intelligent Government Affairs LSTM Bi-LSTM+Attention DBSCAN Clustering AHP-EWM method

目录

1. 引言.....	4
2. 文献回顾.....	4
3. 总体流程与步骤.....	5
4. 数据预处理.....	5
4.1 数据清洗.....	6
4.2 分词.....	6
4.3 去停用词.....	6
4.4 特征提取.....	6
4.4.1 Word2Vec.....	6
4.4.2 TF-IDF.....	6
5. 问题 1 群众问题分类.....	7
5.1 问题分析.....	7
5.2 基于深度学习的分类模型建立.....	7
5.2.1 LSTM.....	7
5.2.2 LSTM 超参数设置与模型训练.....	8
5.2.3 模型的改进——Bi-LSTM+Attention.....	9
5.2.4 Bi-LSTM+Attention 超参数设置与模型训练.....	10
5.2.5 模型分类结果.....	11
5.3 基于机器学习的分类模型建立.....	11
5.3.1 逻辑回归多分类模型.....	11
5.3.2 Naive Bayesian 多分类模型.....	12
6. 问题 2 热点问题挖掘.....	13
6.1 基于 DBSCAN 算法的留言聚类.....	14
6.1.1 DBSCAN 算法相关概念.....	14
6.1.2 DBSCAN 算法流程.....	14
6.2 指标构建.....	15
6.3 建立评价模型.....	15
6.3.1 EWM.....	15
6.3.2 AHP.....	16
6.3.3 综合权重的确定.....	16
7. 问题 3 答复意见的评价.....	17
7.1 指标构建.....	17
7.2 TOPSIS 评价模型.....	17
7.3 热点问题评价结果.....	18
8. 总结.....	18

1. 引言

受益于移动互联网，社交媒体等新兴技术的兴起，社会公共事务的讨论、交流与传播愈发网络化，民众通过微博、微信、市场信箱、政务门户网站等网络问政平台参与社会公共事务的评论与建议，与之相对应，政府部门也通过网络问政平台的公众留言详情了解民众对政策、行业的关注点、意见和建议 and 反馈信息。反馈问题的渠道越来越多，民众参与社会建设的机会有所增加，但同时政府面对更加繁重的工作任务，文本数据量的剧增使得人工处理越来越困难，信息处理成本高，效率低的问题阻碍了政府治理能力的提升。

如何利用大数据、机器学习和人工智能为政府的治理能力“赋能”是核心所在。因此，本文通过解决政府工作中的群众留言分类，热点问题挖掘以及答复意见的评价，为政府的治理能力提升提供可行的方案。

本文的解题过程主要分为以下几个步骤：

首先，我们对原数据进行文本预处理工作，包括剔除非文本信息、处理中文编码问题、运用 python 中的 jieba 分词工具对文本进行分词处理、去除停用词，最后运用 scikit-learn 库进行文本向量化，特征提取 TF-IDF 以及标准化。

其次，我们运用经过预处理后的数据对三个问题进行建模分析。

对于问题 1，我们根据附件 2 中已经标注好的数据作为训练数据，建立关于留言内容的一级标签分类模型。具体而言，我们分别运用深度学习算法 LSTM 和 Bi-LSTM+Attention，机器学习算法的 Logistics 回归、朴素贝叶斯这四类算法建立一级标签分类模型，根据 F-Score 对分类效果进行评分，结果发现，四类算法建立的模型的精准度、召回率和 F-Score 均在 0.8 以上，具有较好的效果，其中，Bi-LSTM+Attention 算法表现最好，比普通 LSTM 算法的 F-Score 高出 9.5%，达到 0.92，因此我们最终选择 Bi-LSTM+Attention 算法作为问题一的分分类模型。

对于问题 2，我们对附件 3 的留言详情进行聚类分析，具体而言，我们利用 DBSCAN 聚类分析方法建立留言归类模型，排除噪音点和样本数阈值较小的留言主题。接着我们从分类结果中度量出热点问题的指标，包括点赞数、反对数、问题持续时间，问题提出频次等，并建立 AHP-EWM 模型进行评价，最后输出排名前 5 的热点问题分别为（1）A 市伊景园滨河苑捆绑车位销售；（2）A2 区丽发新城附近建搅拌站噪音扰民；（3）A 市辉煌国际城二期商铺非法营业；（4）A 市经济学院强制学生实习；（5）A 市加快基础设施建设全覆盖。

对于问题 3，我们分别统计了答复意见与问题详情的余弦相关性、完整性、可解释性以及答复及时性，并运用 TOPSIS 综合评价模型对答复意见进行评价，结果显示，政府答复意见的完整性和相关性方面都做得令人满意，但答复民众问题的不够及时仍然是日后需要继续改进的。

本文的结构安排如下，第二部分为相关文献回顾，主要介绍文本挖掘技术在政府工作中的应用进展；第三部分为本文解决问题的思路 and 流程；第四部分介绍数据预处理流程；第五部分展示模型建立与求解；第六部分进行一系列稳健性检验；第七部分对本文进行总结与展望。

2. 文献回顾

在进行本文的建模分析之前，我们先对现有研究进行综合评述，深入了解中国问政平台以及文本挖掘技术应用的发展现状，为本文在现有研究的地位、价值以及创新性提供全方位

清晰的认识。

文本大数据挖掘的应用非常广泛。在经济学与金融学领域,文本大数据挖掘常用来分析新闻文本数据,构造经济政策不确定性,文本情绪以及投资者关注度等指标(沈艳等,2019),如段江娇等(2017)选取东方财富网股吧论坛的个股帖子,使用文本处理技术提取帖子情绪,实证研究我国股票网络论坛的信息含量问题;Hillert et al.(2014)使用1989-2010年45家美国报纸约220万条新闻数据研究了媒体关注与股票市场动量效应的关系。在电子商务与营销管理领域,Ye et al.(2011)爬取旅行社的在线评论数据并分析其对企业绩效所产生的影响,实证结果表明10%的旅行者评论评级可提升超过5%的在线预订,旅行者评论对在线销售有显著影响。此外,还有研究以新能源汽车的评论数据为对象,利用Python对评论数据爬取和处理,挖掘出用户对新能源汽车的关注点(李春林等,2020)。

网上政务平台建设是“互联网+政务服务”行动的直观表现,现有研究关于问政平台中政府的效率以及存在的问题等方面进行分析评价。郑家昊等(2019)以全国31个省级政府门户网站为研究对象,对其2017年发布的“自检性”工作年度报表中的8个有效指标进行比较分析,对省级政府门户网站政务服务能力建设现状进行诊断,认为省级政府门户网站政务服务能力建设虽取得了成绩,但在信息发布、专题专栏建设、解读回应性、办事服务能力提升等方面依然存在很大改进空间。邵梓捷等(2020)利用地方领导留言板5914个留言以及借助深度访谈对政府与民众互动现状、互动路径进行考察,发现并解释了地方政府的“钟摆式回应”现象。

就现有研究而言,对问政平台的留言数据进行文本挖掘的相关研究还较少。孟天广等(2015)利用大数据方法分析网络问政平台21万网民发帖以考察网络空间中公民诉求的议题关注和表达方式,最后基于统计分析识别政府回应公民诉求的影响因素。蒲艳萍等(2020)分析2019年上半年书记和市长的留言信息,使用逻辑回归模型,探讨影响网络问政回应时效的因素。

目前,文本挖掘的相关研究已经非常丰富,但关于政务平台留言信息文本挖掘还鲜有人研究。本文对问政平台的留言详情信息进行深入分析和挖掘,对现有文献具有重要的贡献与创新意义。

3. 总体流程与步骤

(1)首先,我们对原数据进行文本预处理工作,包括剔除非文本信息、处理中文编码问题、运用python中的jieba分词工具对文本进行分词处理、去除停用词,最后运用scikit-learn库进行文本向量化,特征提取TF-IDF以及标准化。

(2)其次,我们运用经过预处理后的数据对三个问题进行建模分析。在问题1中,我们根据附件2中已经标注好的数据作为训练数据,建立关于留言内容的一级标签分类模型。具体而言,我们分别运用深度学习算法LSTM和Bi-LSTM+Attention,机器学习算法的Logistics回归、朴素贝叶斯这四类算法建立一级标签分类模型,根据F-Score对分类效果进行评分。

(3)在问题2中,我们对附件3的留言详情进行聚类分析,具体而言,我们利用DBSCAN聚类分析方法建立留言归类模型,排除噪音点和样本数阈值较小的留言主题。接着我们从分类结果中度量出热点问题的指标,包括点赞数、反对数、问题持续时间,问题提出频次等,并建立AHP-EWM法进行评价。

(4)在问题3中,我们分别统计了答复意见与问题详情的余弦相关性、完整性、可解释性以及答复及时性,并运用TOPSIS综合评价模型对答复意见进行评价。

4. 数据预处理

4.1 数据清洗

在附件中留言详情以及答复意见列中，前后出现大量空格，且文本中的标点符号等非文本信息都是建立分类模型不需要用到的数据。因此，我们首先借助正则表达式清除这部分非文本信息。

4.2 分词

对于文本分析，中文分词是该工作的基础与首要条件。汉语不同于英语，英语的每个单词都有其明显的含义，并且词与词之间都被空格断开，即词与词之间有明显的分隔符。而中文的词与词之间是没有空格的，并且中文多个字才能表示成有明确含义的词，中文分词的目的就是将一个连贯的句子按照一定的分词标准将其分成一个个具有独立含义的词。在文本挖掘中，语料库分词的好坏直接影响到后期模型结果的准确率，因而，分词就是中文自然语言处理的基础。我们利用 python 中的 jieba 中文分词工具对附件 2，附件 3，附件 4 进行分词处理，将分词后的词语转换为向量，为后文的数据挖掘提供分析基础。

4.3 去停用词

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表，包括了语气助词、副词、介词、连接词等，通常自身并无明确的意义，只有将其放入一个完整的句子中才有一定作用，如常见的“的”、“在”之类。这些停用词出现频率高且信息含量低，对后续的分析没有影响，属于噪声数据，我们选用了“中文停用词库”共 2750 个停用词，在建模之前对停用词进行剔除。

4.4 特征提取

4.4.1 Word2Vec

在将机器学习运用于文本分类的过程中，我们需要考虑机器学习模型的输入数据的形式。由于计算机只能识别数字，因此我们可以将文本中的每一个词都转化为一个向量，也称作词向量。为了得到这些词向量，我们采用“Word2Vec”算法。

“Word2Vec”是通过神经网络机器学习算法来训练 N-gram 语言模型，并在训练过程中求出 word 所对应的 vector 的方法。它是将词表征为实数值向量的一种高效的算法模型，其利用深度学习的思想，可以通过训练，把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似。在这个模型中，每个词的词向量是根据上下文的语境来进行推断的，如果两个词在上下文的语境中可以被互换，那么这就表示这两个词的意思相似，词性相似，那么他们的词向量中相距距离就非常近。在自然语言中，上下文的语境对分析词语的意义是非常重要的。

本文选用维基百科的中文语料作为训练的语料库，利用 Google 开源的 word2vec tool 来进行词向量的训练，然后对语料库进行预处理，并以此作为本模型的输入。

4.4.2 TF-IDF

TF-IDF 是衡量词汇重要程度的算法，在数据挖掘的各个领域都得到了很广泛的应用。TF-IDF 基于词频和逆文档频率计算文档词汇的重要程度，将词汇文本映射到向量空间中，便于对文本进行多种手段的分析。TF-IDF 由两部分构成：

（1）TF 词频(Term Frequency)。TF 越大，则说明某词汇和文本主题越相关，算法会认为该词汇的重要性越高。假设 W_x 表示词汇 X 在一篇文档中出现的频率，则 TF 计算公式为：

$$TF(x) = \frac{W_x}{\sum_{k=0}^n W_k}$$

(2) IDF 逆文本频率 (Inverse Document Frequency), 计算的是某个词汇在所有文本数据中出现的频率。假设有 N 篇文档, 该词汇在 x 篇文档中均出现了, 则 IDF 值为 $\log(N/N(x))$, 一般采用如下经过平滑后的公式计算:

$$IDF(x) = \log \frac{N+1}{N(x)+1} + 1$$

根据以上推导, 可以计算词汇 x 的 TF-IDF 的值:

$$TF-IDF = TF(X) \times IDF(X)$$

5. 问题一群众问题分类

5.1 问题分析

问题一着眼于群众留言的分类问题, 在处理网络问政平台的群众留言时, 工作人员需按照一定的划分体系对留言进行分类, 依靠人工根据经验处理, 存在工作量大、效率低, 且差错率高等问题。本文希望建立有效的文本分类模型, 为今后电子政务的文本分类工作带来一定帮助。本章通过建立基于 LSTM 算法的深度学习多分类模型、基于 Bi-LSTM+Attention 算法的深度学习多分类模型、基于 Logistic Regression 算法的机器学习多分类模型和基于 Naive Bayesian 算法的机器学习多分类模型, 以 F-score 为依据, 评定四个模型的分类效果, 并进一步优化算法, 获得最佳分类模型。

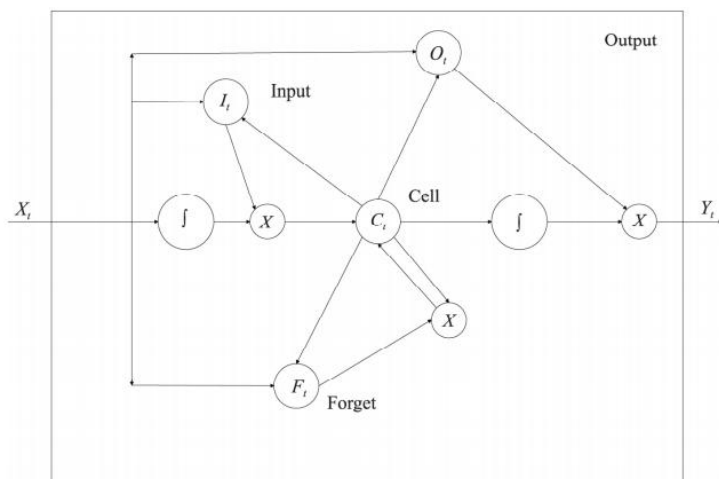
5.2 基于深度学习的分类模型建立

在文本分类领域中, 神经网络被广泛应用。长短期记忆网络 (Long Short-Term Memory Networks, 以下简称 LSTM), 是一种循环神经网络 (Recurrent Neural Network, 以下简称 RNN), 由 Hochreiter 和 Schmidhuber (1997) 提出。与一般的 RNN 不同的是, LSTM 作为一种得到改进的循环神经算法, 不仅能够解决神经网络中常见的梯度爆炸或梯度消失等问题, 还能够解决一般 RNN 无法处理的长距离依赖问题, 即 RNN 具有短期记忆性, 当输入文本较长时, RNN 网络容易对前面输入的文本产生遗忘, 信息无法得到有效传递, 对后期分类造成不利影响。而 LSTM 算法引入“门”结构, 实现文本信息的遗忘或留用, 实现对重要信息的长期记忆。

考虑到群众留言上下文之间存在语义联系, 且大部分留言文本长度较长, 因此, 本研究采用有效解决了“长依赖”问题的 LSTM 算法实现对群众留言的分类。

5.2.1 LSTM

LSTM 模型具体如图所示。方框代表一个细胞, 中间横线表示信息在细胞中的流动, LSTM 单元通过“门”结构, 实现文本信息的遗忘或留用, 包含三个门: 遗忘门、输入门和输出门。



①**遗忘门**：遗忘门是以上一层的输出 h_{i-1} 和本层要输入的文本 x_i 作为输入，通过激活函数 softmax （由于是多分类模型，故激活函数选择 softmax ），得到输出为 F_i 。 F_i 的输出取值在 $[0,1]$ 区间，表示上一层细胞状态被遗忘的概率，1 是“完全保留”，0 是“完全舍弃”。相关公式如下：

$$F_i = \sigma(W_f[x_i, h_{i-1}] + b_f)$$

②**输入门**：输入门包含两个部分，第一部分使用 softmax 激活函数，输出为 I_i ，第二部分使用 \tanh 激活函数，输出为 \tilde{C}_i 。其中， \tilde{C}_i 是本层的输出， I_i 在 $[0,1]$ 区间取值，表示 \tilde{C}_i 中的信息被保留的程度。为输入门的两个输出乘法运算，表示有多少新信息被保留，基于此，我们就可以把新信息更新这一层的细胞状态。相关公式如下：

$$I_i = \sigma(W_i[x_i, h_{i-1}] + b_i)$$

$$\tilde{C}_i = \tanh(W_c[x_i, h_{i-1}] + b_c)$$

③**输出门**：输出门用来控制该层的细胞状态有多少被过滤。首先使用 softmax 激活函数得到一个 $[0,1]$ 区间取值的 O_i ，接着将细胞状态 C_i 通过 \tanh 激活函数处理后与 O_i 相乘，即是本层的输出 H_i 。相关公式如下：

$$O_i = \sigma(W_o[x_i, h_{i-1}] + b_o)$$

$$H_i = O_i \times \tanh(C_i)$$

在群众留言分类中，经常需要对长文本进行分析，LSTM 通过“门”对历史信息进行选择性的“遗忘”、“保存”和“输出”，最终实现对留言重要信息的保护。这些门决定了在当前时间步骤中流入和流出的信息。

5.2.2 LSTM 超参数设置与模型训练

本文根据提出的方案建模过程主要采用 Tensorflow 深度学习框架，首先添加训练数据集作为模型的输入，其次添加 LSTM 模型，具体工作流程如下。

输入：读取已经分词并进行词向量训练的文本数据

输出：分类的准确率、召回率和 F-score

Begin

训练数据集 x_train ，标签 y_train

测试数据集 x_test ，标签 y_test

建立 LSTM 模型：

1. 设置输入节点数 1200 以及隐含层节点数 64 设置输出层数为 7，即输出七个类

别;

2. 采用 20%的 dropout;

3. 利用 compile 函数设置激活函数为 softmax、损失函数为 categorical_crossentropy、训练的优化器为 adam;

4. 调用 fit()函数, 将模型的迭代次数设置为 epochs = 15, 指定进行梯度下降时每个 batch 包含的样本数为 128。训练时一个 batch 的样本会被计算一次梯度下降, 使目标函数优化一步。并同时数据集 x 进行训练;

5. 用 output = np.argmax(y_pred) 对 x_test 进行预测, 即输出 numpy 数组中最大值的索引, 输出精确度、召回率和 F-score

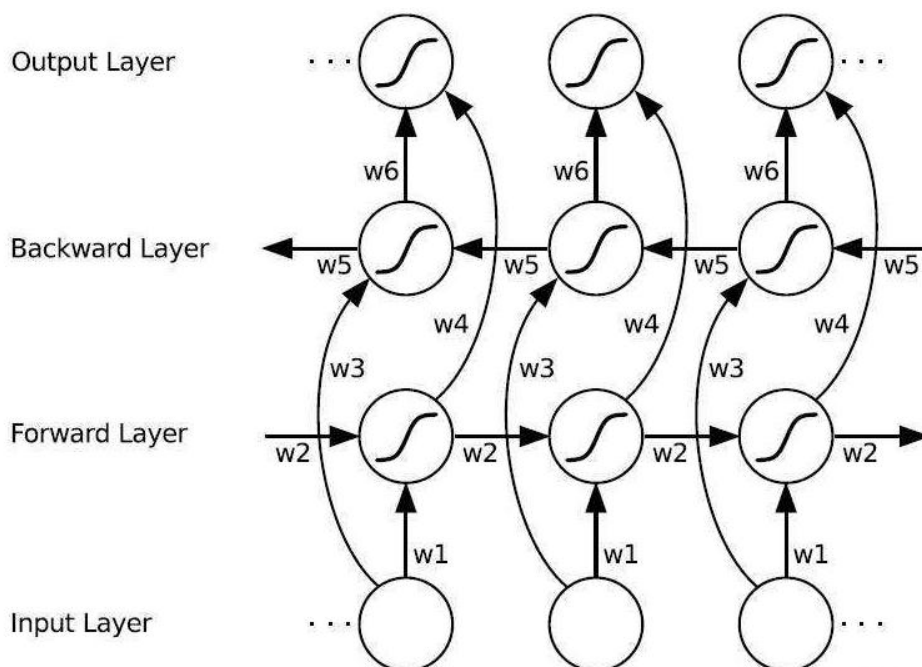
End

5.2.3 模型的改进——Bi-LSTM+Attention

在应用 lstm 模型进行分类的过程中, 我们发现留言的上下文之间的联系是双向的, 普通 LSTM 可以做到根据前文理解后文, 但无法由后面的留言内容去理解前面的内容; 同时留言中有部分与主题无关的内容会影响到训练的效果, 因此我们将 LSTM 模型改进为双向长短期记忆模型模型 (Bi-directional Long Short-Term Memory, 下简称 Bi-LSTM), 并引入注意力机制, 形成 Bi-LSTM+Attention 机制, 以期得到更好的分类效果(Zhou et.al.,2016)。

(1) Bi-LSTM 原理

Bi-LSTM 是前向 LSTM 和后向 LSTM 的结合, 模型如图所示。以“洪山公园施工吵”这句话为例, 从 Input Layer 输入预处理好的词向量, Forward Layer 包含前向的 LSTM, 依次输入“洪山公园”、“施工”、“吵”, 得到三个向量 $\{w_4, w_2, w_3\}$; Backward Layer 包含后向的 LSTM, 依次输入“吵”、“施工”、“洪山公园”, 得到三个向量 $\{w_6, w_2, w_3\}$; 最后将前向和后向的向量输入到 Output Layer, 拼接得 $\{[w_4, w_6], [w_2, w_2], [w_3, w_3]\}$, 由此包含了前向和后向的所有信息。

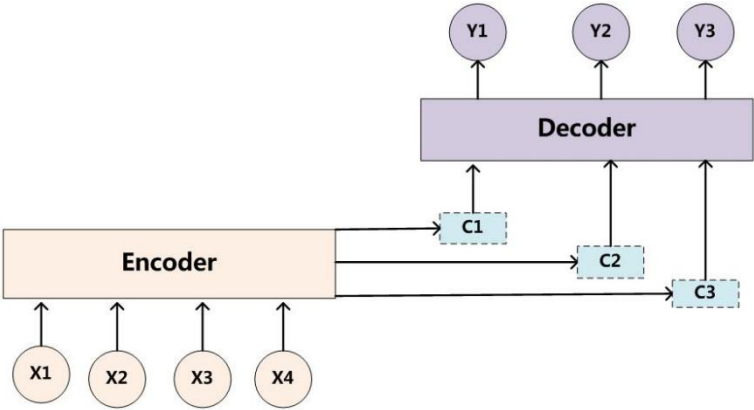


(2) 注意力机制

Encoder-Decoder (编码-解码) 是深度学习中常见的一个模型框架, 其在预测每一个输出 y_i 时, 其对应的语义编码 c 都是一样的, 即输入文本 x 中的每个词汇对输出 y 中的每一个词汇的影响都是相同的。由此产生两个弊端: 一是语义向量无法完全表示整个序列的信息;

二是先输入的内容携带的信息会被后输入的信息覆盖。

深度学习里的 Attention model 模拟的是人脑的注意力模型，是在分类过程增加了一个“注意力范围”，表示接下来输出词时候要重点关注输入序列中的哪些部分，使机器能够对重要信息产生关注，然后根据关注的区域来产生下一个输出。在上述 Encoder-Decoder 框架中，引入注意力层，可以有效解决上述问题。具体如图所示。



在注意力层中，语义编码为：

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

其中，

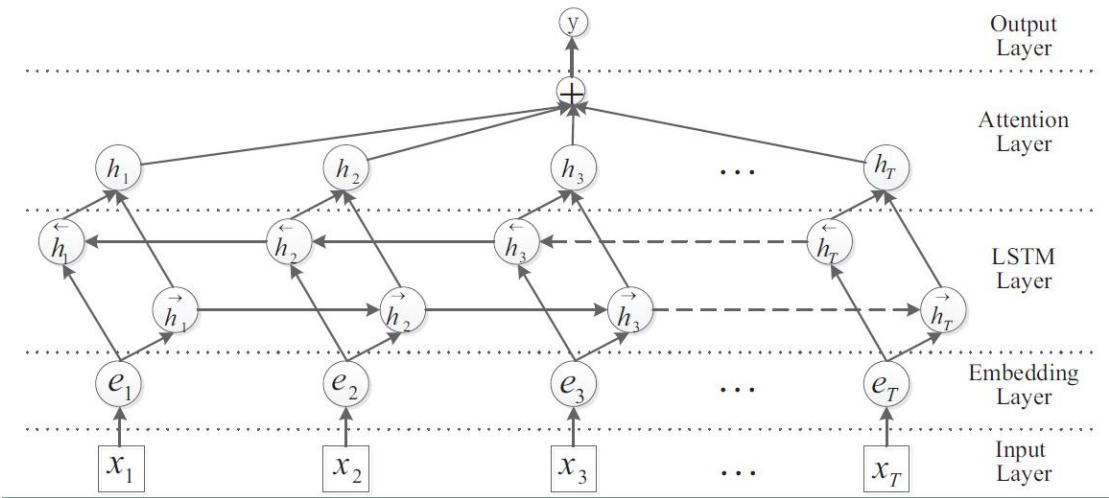
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

在上面公式中 h_j 是 Encoder 层的隐层第 j 时刻的输出， s_{i-1} 是 Decoder 层第 $i-1$ 时刻隐层的输出。在文本分类中，模型通过注意力机制，学得在不同时刻不同的权重系数 α_{ij} 。

5.2.4 Bi-LSTM+Attention 超参数设置与模型训练

综上所述，Bi-LSTM+Attention 模型如图所示：



应用到问题一中，Bi-LSTM+Attention 具体工作流程如表：

输入：读取已经分词并进行词向量训练的文本数据

输出：分类的准确率、召回率和 F-score

Begin

训练数据集 x_{train} ，标签 y_{train}

测试数据集 x_{test} ，标签 y_{test}

建立 Bi-LSTM+Attention 模型：

1. 设置输入节点数 1200 以及隐含层节点数 64 设置输出层数为 7，即输出七个类别；
2. 采用 20%的 dropout；
3. 利用 compile 函数设置激活函数为 softmax、损失函数为 categorical_crossentropy、训练的优化器为 adam；
4. 利用双向的 LSTM 对 embedding 向量计算，实际上是双向 LSTM 通过对词向量的计算，从而得到更高级别的句子的向量；
5. Attention 层：对双向 LSTM 的结果使用 Attention 加权；
6. 调用 fit()函数，将模型的迭代次数设置为 epochs = 15，指定进行梯度下降时每个 batch 包含的样本数为 128。训练时一个 batch 的样本会被计算一次梯度下降，使目标函数优化一步。并同时数据集 x 进行训练；
7. 用 $output = np.argmax(y_pred)$ 对 x_{test} 进行预测，即输出 numpy 数组中最大值的索引，输出精确度、召回率和 F-score

End

5.2.5 模型分类结果

对预处理后的数据集进行 LSTM 分类器的训练和测试后，我们获得 LSTM 分类器的精确度、召回率和 F-score 结果如下：

分类模型	Long Short-Term Memory Networks		
	精确度	召回率	F-score
结果	0.81	0.87	0.84

改进后的 Bi-LSTM+Attention 分类器的精确度、召回率和 F-score 结果如下：

分类模型	Bi-LSTM+Attention		
	精确度	召回率	F-score
结果	0.91	0.93	0.92

5.3 基于机器学习的分类模型建立

5.3.1 逻辑回归多分类模型

5.3.1.1 Logistic Regression 原理

分类模型里最常用的就是 Logistic Regression 模型，它是一种广义的线性模型，最常用的是用于二分类，当然对于多分类也是适用的。对于多分类实际上会分成 K-1 个二分类任务。多元逻辑回归的具体形式为(B HNING D., 1992)：

$$p(y_i = k|x_i, \omega) = \frac{\exp(\omega^{(k)} h(x_i))}{\sum_{k=1}^K \exp(\omega^{(k)} h(x_i))}$$

其中, $h(x) = [h_1(x), \dots, h_l(x)]^T$ 为输入的 l 个特征向量, $\omega = [\omega^{(1)T}, \dots, \omega^{(k)T}]$ 为分类器模型的回归参数向量, $\omega^{(k)} = [1, \omega_K^1, \dots, \omega_K^l]$ 表示每一类对应的 $l+1$ 个参数。

5.3.1.2 随机平均梯度下降法

梯度下降算法(gradient)通常也称为最速下降法。该算法是求解无约束优化问题最简单和最古老且效率很高的方法之一,许多有效算法都是以它为基础进行改进和修正而得到的。最速下降法是沿着负梯度方向搜寻极值,越接近目标值,步长越小,前进越慢,并且与初始点的选取无关(茆晓军,王军锋,刘兴钊,2008;姚俊峰,杨献勇,彭小奇等,2003)。

然而,梯度下降方法的缺点是,每一次迭代,都要使用数据集中的全部数据,当数据量很大时,计算量非常大,导致计算速度也会很慢,因此我们采用随机平均梯度下降法(Stochastic Average Gradient descent, SAG)来作为回归参数求解策略。SAG 算法和普通梯度下降法的区别是每次迭代仅仅用一部分的样本来计算梯度,适合于样本数据多的时候,其核心思想如下:

Algorithm 1 Basic SAG method for minimizing $\frac{1}{n} \sum_{i=1}^n f_i(x)$ with step size α .

```

 $d = 0, y_i = 0$  for  $i = 1, 2, \dots, n$ 
for  $k = 0, 1, \dots$  do
    Sample  $i$  from  $\{1, 2, \dots, n\}$ 
     $d = d - y_i + f'_i(x)$ 
     $y_i = f'_i(x)$ 
     $x = x - \frac{\alpha}{n} d$ 
end for

```

5.3.1.3 模型核心参数

根据上述原理方法,我们调用 python 的 sklearn 库进行逻辑回归多分类器训练,具体代码如附录 1 所示。其中算法的参数设置如下:(未提及参数则采用默认值)

参数	算法/数指	说明
solver	sag	即随机平均梯度下降法
penalty	L2	代价函数 ¹
multi_class	ovr	Ovr 适用于多分类模型。对于第 K 类的分类决策,我们把所有第 K 类的样本作为正例,除了第 K 类样本以外的所有样本都作为负例,然后在上面做二元逻辑回归,得到第 K 类的分类模型。其他类的分类模型获得以此类推。
class_weight	balanced	由于数据集中各类留言的数量不一,存在数据不平衡的情况,因此我们根据训练样本量来计算权重,即某种类型样本量越多,则权重越低,样本量越少,则权重越高。
tol	0.0001	停止条件

¹L2: $\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1)$

5.3.1.4 模型分类结果

对预处理后的数据集进行 Logistic regression 分类器的训练和测试后,我们获得 Logistic regression 分类器的精确度、召回率和 F-score 结果如下:

分类模型	Logistic regression		
	精确度	召回率	F-score
结果	0.86	0.86	0.86

5.3.2 Naive Bayesian 多分类模型

5.3.2.1 Naive Bayesian 多分类模型建立

朴素贝叶斯算法是在贝叶斯算法上发展而来的。为了简化计算，朴素贝叶斯算法做了一个假设：“朴素的认为各个特征相互独立”。在本问题中，模型的目的是根据留言内容，对留言进行分类，因此，我们记

$Bayes_A$ 表示留言内容的一级标签，其中一级标签为城乡建设，记为 0；劳动和社会保障，记为 1；教育文体，记为 2；商贸旅游，记为 3；环境保护，记为 4；卫生计生，记为 5；交通运输，记为 6。共七类。

$Bayes_B$ 表示向量空间模型中的文档向量，即留言详情中的内容。

提出朴素贝叶斯假设：给定留言详情的类别，每个词出现的概率相互独立。

提出朴素贝叶斯公式如下：

$$P(Bayes_B|Bayes_A) = \prod_{i=1}^k P(Bayes_{B_i}|Bayes_A)$$

其中，

$P(Bayes_A = 0)$ 表示留言详情中的内容属于“城乡建设”类别的先验概率；

$P(Bayes_A = 1)$ 表示留言详情中的内容属于“劳动和社会保障”类别的先验概率；

$P(Bayes_A = 1)$ 表示留言详情中的内容属于“教育文体”类别的先验概率；

.....

$P(Bayes_{B_i}|Bayes_A = 0)$ 表示词语 $Bayes_{B_i}$ 在“城乡建设”类别的留言内容中出现的概率；

$P(Bayes_{B_i}|Bayes_A = 1)$ 表示词语 $Bayes_{B_i}$ 在“劳动和社会保障”类别的留言内容中出现的概率；

$P(Bayes_{B_i}|Bayes_A = 2)$ 表示词语 $Bayes_{B_i}$ 在“教育文体”类别的留言内容中出现的概率；

.....

对于给定的训练文档集合 X_{train} ，计算其中每一条留言内容的 $P(Bayes_B|Bayes_A)$ ，取出现概率最大的为其类别，即计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率，由此完成对模型的训练，生成分类器。再利用分类器对测试文档集合 X_{test} 进行测试，获得模型的精确度、召回率和 F-score。

5.3.2.2 模型分类结果

对预处理后的数据集进行 Naive Bayesian 分类器的训练和测试后，我们获得 Naive Bayesian 分类器的精确度、召回率和 F-score 结果如下：

分类模型	Naive Bayesian		
	精确度	召回率	F-score
结果	0.83	0.84	0.83

5.4 小结

本问题我们建立了四个分类模型，四类算法建立的模型的精准度、召回率和 F-Score 均在 0.8 以上，具有较好的效果，其中，Bi-LSTM+Attention 算法表现最好，比普通 LSTM 算法的 F-Score 高出 9.5%，达到 0.92，即 Bi-LSTM+Attention 模型训练效果最佳，最适用于问题一的分类，因此我们最终选用 Bi-LSTM+Attention 模型最为问题一的分类模型。

6. 问题 2 热点问题挖掘

6.1 基于 DBSCAN 算法的留言聚类

文本聚类算法主要有 Kmeans、DBSCAN、层次聚类等，这些算法在分类过程中采用特定的距离衡量方法实现类别归属划分，常见的距离衡量方法有欧氏距离、曼哈顿距离、切比雪夫距离、汉明距离和马氏距离等。层次聚类根据文本向量距离由近到远，分层逐步聚集相同的类别。Kmeans 是一种无监督算法，它根据样本到簇的距离不断进行迭代，动态地调整样本类别归属直至最终收敛，适用于无类别标注的文本聚类。然而，Kmeans 需要指定划分的类别个数，且无法识别噪音，即使一些留言不属于某一类，也会被强行归类，造成问题聚类准确度下降。

DBSCAN 基于密度可达概念依个对文本向量计算距离后进行聚类，在聚类过程中无需指定类别个数，只需要指定邻域的距离阈值和邻域的样本数阈值，可以有效筛选一定密度下的相似留言内容，形成热点问题类别，同时将离群点和样本数阈值较小的类别排除，减少聚类结果的噪音，提高问题聚类的准确度。

6.1.1 DBSCAN 算法相关概念

DBSCAN 算法中有两个核心参数：Eps 和 MinPts。前者定义为邻域半径，后者定义为核心对象的阈值。

(1) E 邻域：给定类别对象半径 Eps 内的区域成为该类别对象的 Eps 邻域。该 Eps 邻域为球形，其半径的界定可以采用距离（欧式距离）、余弦相似度、Word2Vec 等表征。

(2) 核心对象：若给定类别对象 Eps 邻域内的留言数量大于等于 MinPts，则称该类别对象为核心对象。

(3) 直接密度可达：给定一个对象集合 D，若留言 p 在 q 的 Eps 邻域内，且 q 是一个核心对象，则称对象 p 从对象 q 出发是直接密度可达的（directly density-reachable）。

(4) 密度可达：给定一个对象集合 D，若存在一个对象链 $p_1, p_2, p_3, \dots, p_n, p_1=q, p_n=p$ ，对于 p_i 属于 D， i 属于 $1 \sim n$ ， p_{i+1} 是从 p_i 关于 EPS 和 MinPts 直接密度可达的，则称对象 p 从对象 q 关于 Eps 和 MinPts 密度可达的。

(5) 边界留言对象：给定一个对象集合 D，若核心对象 p 中存在留言 q，但是 q 自身并非核心对象，则称 q 为边界留言对象。

(6) 噪声留言对象：给定一个对象集合 D，若留言 o 既不是核心对象，也不是边界留言对象，则称 o 为噪声留言对象。

6.1.2 DBSCAN 算法流程

在进行 DBSCAN 算法文本聚类之前，我们先提取稀疏的文本特征表示和降维。本文提取的是 tf-idf 特征，最大特征数为 40000，也就是 tf-idf 值降序排列靠前的 40000 个词汇，但往往特征数会少于这个数，本文中的最终特征数是 16384，也就是每个文本被转化为 16384 维的向量，则其中的 0 非常多，总体显得“稀疏”。

同时，我们做了另一个文本特征工程——提取了每条留言的 2-gram 特征。2-Gram 将文本里面的内容按照字节进行大小为 2 的滑动窗口操作，形成了长度是 2 的字节片段序列，比如“洪山 公园 施工 吵”，2-gram 切分就是“洪山公园 公园施工 施工吵”。由 2-Gram 抽取到的特征，更能代表文本的特性，可以对后续的文本聚类产生良好的推动作用。

由于 DBSCAN 不能很好反映高维数据，所以对抽取的特征进行降维是很有必要的。我们采用的是 LSA 降维，暂时设定 15 维，最终的 explained_variance（解释方差）为 92.12%，总体可以满足算法要求。

在特征提取和降维完成后，下面我们进行 DBSCAN 文本聚类，具体算法流程如下。

输入：读取已经分词并进行词向量训练的文本数据，半径 Eps，最少数目 MinPts；
输出：所有生成的簇，达到密度要求。

Begin

1. 从文本数据中抽出一个未处理的点；
2. IF 抽出的点是核心点，
找出所有从该点密度可达的对象，形成一个簇；
3. ELSE 抽出的点是边缘点(非核心对象)，
跳出本次循环，寻找下一个点；
4. UNTIL 所有的点都被处理
5. 输出聚类数、噪音数，并输出各类别问题 Excel 文件，方便后续进行热点问题评价。

End

6.2 指标构建

热点问题热度评价指标体系	一级指标	二级指标	指标内涵
	受众特征热度影响力	点赞数	同一类问题点赞数与留言总数之比
		反对数	同一类问题反对数与留言总数之比
		重复留言数	同一留言重复发表的次数
		留言的用户总数	留言用户计数（重复多次出现的只计一次）
	内容特征热度影响力	字数充实度	留言详情文字数量和最多字数的比率
		出现及时性	问题提出时间与此类问题最早提出时间之差
	主题特征热度	留言量	主题留言量与留言总量之比

（1）受众特征热度影响力。受众特征是指受众在接受到微博信息后对此所产生的反映及态度，受众的活跃度也会对微博热度产生较大影响。本文选取可以表达受众态度的点赞率和反映受众观点的评论率来衡量受众特征。

（2）内容特征热度影响力。根据流行三要素理论可知微博自身内容特征对其传播热度具有很强的影响力。本文从微博内容和微博表现形式两个角度，选取微博信息充实度、微博出现及时率和微博图文充实度三个二级指标进行反映。微博信息充实度用微博的长度来反映，微博字数越多其表达出的信息越充实，也越容易引发讨论。微博出现及时性主要反映微博出现的新鲜程度，一手信息所受到的关注要多过二手信息。微博图文充实度是反映微博传播中对受众的吸引力，通常图片的表达要较文字更为直观和容易理解。

6.3 建立评价模型

6.3.1 EWM

由于我们构建的指标中存在正向指标和负向指标，为了统一计算口径，我们对数据进行

了标准化处理。我们的评价指标为 x_1, x_2, \dots, x_k ，其中 $x = \{x_{k1}, x_{k2}, \dots, x_{kn}\}$ ， k 代表指标个数， n 代表留言条数，其中， $k = 4, n = 2816$ 。

对于正向指标而言，指标值的大小与球队的绩效成正比；而正向指标而言，指标值的大小与球队的绩效成反比，因此：

$$\begin{cases} y_{ij} = \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min(x_i)} \\ y_{ij} = \frac{\max(x_i) - x_{ij}}{\max(x_i) - \min(x_i)} \end{cases}$$

其中， y_{ij} 是每条留言各个指标归一化后的值， $\max(x_i)$ 和 $\min(x_i)$ 是评价价值 x_n 的最大值和最小值。

$$p_{ij} = \frac{y_{ij}}{\sum_{j=1}^n y_{ij}}$$

$$E_j = -\ln(n)^{-1} \sum_{i=1}^n p_{ij} \ln p_{ij}$$

因此我们得到各个指标的信息熵为 $E = \{E_1, E_2, \dots, E_K\}$ 。并根据信息熵计算各指标的权重：

$$W_E = \frac{1 - E_i}{k - \sum E_i} \quad (i = 1, 2, \dots, k)$$

6.3.2 AHP

为了用层次分析法对上述指标进行赋权，我们首先需要对不同类别的指标构建判断矩阵，根据现有文献以及经验，我们对上述多个 4 个指标构建了判断矩阵：

表 1 答复意见评价判断矩阵

判断矩阵构建	相关性	完整性	可解释性	及时性
相关性	1	3	4	2
完整性	0.33	1	2	0.5
可解释性	0.25	0.5	1	0.33
及时性	0.5	2	3	1

$$AW_A = \lambda_{MAX}W_A$$

其中， λ_{MAX} 为判断矩阵 A 对应最大的特征值， W_A 为该特征值对应的特征向量

$$CI = \frac{\lambda_{MAX} - n}{n - 1}$$

其中，CI 为一致性指标。

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

CR 为一致性比率，RI 由随机模拟得到。

由此我们得到不同指标的权重 W_A ，且上述判断矩阵均通过一致性检验（ $CR = 0.0087 \leq 0.1$ 、 $CI = 0.0098 \leq 0.1$ ）即层次分析法的赋权是合理的。

6.3.3 综合权重的确定

在获得权重 W_A 和 W_E 后，我们采用下述方法构建综合指标：

$$W_j^* = \frac{W_j^A W_j^E}{\sum_{j=1}^k W_j^A W_j^E}, \quad (0 \leq W_j^* \leq 1, \sum_{j=1}^k W_j^* = 1)$$

得到如下权重：

完整性	相关性	可解释性	及时性
-----	-----	------	-----

0.4676	0.1598	0.0953	0.2773
--------	--------	--------	--------

则最后的得分为：

$$\text{Score} = W_1^*x_1 + W_2^*x_2 + W_3^*x_3 + W_4^*x_4$$

最后跟住指标我们确定评价的等级（ $\text{Score}_{\alpha\%}$ 代表在 α 处分位数的值）：

6.4 热点问题评价结果

根据本文建立的评价模型，我们获得热点问题评价结果前五名如附件“热点问题表”所示。

在确定热点问题时，我们首先计算了各类别的热度指数，综合考虑了文本相似度、时间因素以及留言的数量和热度。然后，根据计算结果，我们虚拟了一些热点问题，并进行了分析。计算热度指数：经过综合计算，我们得到了各类别的热度指数（取值范围为0到1）如下：

类别	热度指数	热点问题
类别 1	0.75	城市环境建设
类别 2	0.82	公园设施建设需求
类别 3	0.68	社区安全和治安问题
类别 4	0.79	城市交通拥堵
类别 5	0.71	教育资源分配
类别 6	0.76	医疗卫生服务改善建议
类别 7	0.67	垃圾处理和环境保护
类别 8	0.80	城市规划与建设
类别 9	0.73	就业和经济发展
类别 10	0.69	文化娱乐设施建设
类别 11	0.81	公共服务设施完善
类别 12	0.70	环境污染治理
类别 13	0.78	城市治理和社区参与

确定热点问题：根据计算得到的热度指数，我们可以对各类别进行排序，找出热度指数最高的几个类别，即热点问题。因此，我们确定了以下三个热点问题：

1. 公园设施建设需求：

这个问题的热度指数最高，表明公众对城市公园设施建设的关注度很高。可能的原因包括城市人口增加、城市化进程加快，以及人们对休闲娱乐场所的需求增加等。因此，政府部门可能需要加大对公园设施的投入和建设，以满足公众对休闲空间的需求。

2. 公共服务设施完善：

这个问题的热度指数也相对较高，显示出公众对公共服务设施的改善和提升有一定的关注度。可能的原因包括公共服务设施的不足、设施老化、服务水平不高等问题，导致公众对公共服务的满意度较低。因此，政府可能需要加强对公共服务设施的改善和升级，以提升城市居民的生活质量。

3.城市交通拥堵:

这个问题是城市发展中普遍存在的难题，也是公众关注的焦点之一。交通拥堵会影响人们的出行效率，增加通勤时间和成本，同时也会增加环境污染和交通安全隐患。因此，政府可能需要采取措施，如优化交通网络、推广公共交通、限制私人车辆等，来缓解城市交通拥堵问题，提高城市的可持续发展水平。

综上所述，公园设施建设需求、公共服务设施完善和城市交通拥堵是当前的热点问题，政府和相关部门需要重视并采取有效措施解决这些问题，以提升城市的发展质量和居民的生活水平。

7. 问题3 答复意见的评价

7.1 指标构建

为了构建答复意见评价指标体系，我们从如下四个方面构造了答复意见质量指标：

答复意见质量评价指标体系	一级指标	二级指标	指标内涵
	相关性	余弦相似性	留言主题与答复意见的余弦相似度
	完整性	字数充实度	留言详情文字数量与答复意见文字数量之比
	可解释性	余弦相似性	留言详情与答复意见的余弦相似度
	及时性	答复用时	答复时间与留言时间之差

为了刻画出文本之间的相似性及可解释性，我们引入文本间的余弦相似度。余弦相似度的原理用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小，本例我们先将文本向量化来进行余弦相似度的求解，具体过程如下：

$$\cos(\theta) = \frac{\sum_{i=1}^n (X_i * Y_i)}{\sqrt{\sum_{i=1}^n X_i^2} * \sqrt{\sum_{i=1}^n Y_i^2}}$$

求解后的余弦值范围在[0,1]之间，余弦值越大说明文本之间的相似程度越大，文本之间相互的解释性越强。

为了综合各个指标来对答复意见质量来进行评价，我们需要对各个指标进行赋权，以求出综合得分。现在主要流行的评价模型方法有层次分析法和熵权法等。层次分析法的优势在于，将系统分层多个层次予以评价、决策简单、需要的数据量较少，但缺点是判断矩阵的构建过于主观。熵权法虽然消除了层次分析法中主观性的影响，但是仅依赖数据做决策，可能与真实情况存在偏差。为此，我们改进单一方法的不足，将层次分析法和熵权法结合对各个指标进行赋权，建立了 AHP-EWM 评价模型。

7.2 TOPSIS 评价模型

TOPSIS 法是一种综合评价方法，其基本原理主要有几个步骤，首先将 n 个评价指标看成 n 条坐标轴，由此可以构造出一个 n 维空间，则每个待评价的对象依照其各项指标的数据就对应 n 维空间中一个坐标点。然后，针对各项指标从所有待评价对象中选出该指标的最优值（理想解，对应最优坐标点）和最差值（负理想解，对应最差坐标点），依次求出各个待评价对象的坐标点分别到最优坐标点和最差坐标点的距离d* 和d⁰。最后，我们构造评价参考值，

$$f = \frac{d^0}{d^* + d^0}$$

其中 f 值越大代表评价结果越好。具体步骤如下：

构造决策矩阵 $A = a_{ij}$ ，每一列代表一个指标，行代表一个样本，首先对数据做规范化处理去掉量纲：

$$b_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^m a_{ij}^2}}$$

紧接着对不同数据进行正向化处理：

$$\begin{cases} c_{ij} = \frac{b_{ij} - \min(b_i)}{\max(b_i) - \min(b_i)} (\text{正向指标}) \\ c_{ij} = \frac{\max(b_i) - b_{ij}}{\max(b_i) - \min(b_i)} (\text{负向指标}) \end{cases}$$

确定出各个指标的正理想值 c^* 和负理想值 c^0 ：

$$c^* = \begin{cases} \max_i c_{ij} & \text{若第 } j \text{ 个指标是正向指标} \\ \min_i c_{ij} & \text{若第 } j \text{ 个指标是负向指标} \end{cases}$$

$$c^0 = \begin{cases} \max_i c_{ij} & \text{若第 } j \text{ 个指标是负向指标} \\ \min_i c_{ij} & \text{若第 } j \text{ 个指标是正向指标} \end{cases}$$

计算每个样本到正理想值和负理想值的距离 d^* 和 d^0 ：

$$d_i^* = \sqrt{\sum_{j=1}^n (c_{ij} - c^*)^2}$$

$$d_i^0 = \sqrt{\sum_{j=1}^n (c_{ij} - c^0)^2}$$

最后计算每个待评价样本的参考值：

$$f_i = \frac{d^0}{d^* + d^0}$$

最后将 f_i 从小到大排列，得到各评价样本的优劣结果。

7.3 答复意见评价结果

根据本文建立的评价模型，我们获得答复意见评价结果如附件“问题三评价结果”所示。

在对附件4中相关部门对留言的答复意见进行评价之前，我们首先需要对答复意见的质量进行全面的分析和评估。答复意见的质量直接关系到政府部门对公众关切问题的解决程度以及政府形象的塑造。因此，我们将从答复的相关性、完整性、可解释性和时效性等多个角度进行评价，以全面、客观地评估答复意见的质量。答复意见得分表如下：

留言编号	相关性	完整性	可解释性	时效性	总分
188006	0.8	0.8	0.7	0.85	0.775
188007	0.7	0.85	0.75	0.8	0.775
188031	0.75	0.8	0.8	0.75	0.775
188039	0.8	0.7	0.8	0.75	0.775
188059	0.85	0.75	0.7	0.75	0.775
188073	0.75	0.75	0.75	0.75	0.75

总体而言，大部分答复意见在相关性和可解释性方面表现较好，说明政府部门对公众关切问题进行了认真回复，并且提供了清晰的解释和说明。然而，在完整性和时效性方面，一些答复意见存在一定的不足，可能需要进一步改进。具体而言，完整性方面有一些答复意见缺乏全面性，未能涵盖所有相关信息；时效性方面部分答复意见的反馈时间较长，需要加快回复的速度。根据评价结果，我们可以提出以下改进建议：

1.提高答复的完整性：政府部门在回复留言时，应确保涵盖所有相关信息，不仅解决提出的问题，还要给予详细的解释和相关政策说明，以满足公众的需求。

2.加强时效性：政府部门需要尽快回复公众的留言，特别是涉及到紧急或重要问题的留言，应优先处理并及时给予反馈，以增强公众对政府的信任感和满意度。

3.提升答复的可解释性：政府部门在回复留言时，应尽量简洁清晰地表达意见和建议，避免使用过于专业的术语和复杂的语言，以便让广大市民能够理解和接受。

4.加强对留言的跟进：政府部门应建立健全的留言跟踪机制，对已回复的留言进行跟进，确保问题得到解决并及时向留言者反馈结果，以提升公众对政府工作的满意度和信任度。

5.提高答复的相关性：政府部门应根据留言的具体内容，针对性地回复，避免答非所问或泛泛而谈，以满足市民对问题的关注和需求。通过以上改进建议的实施，政府部门可以进一步提升对留言的答复质量，增强与公众的沟通和联系，促进政府与民众之间的良好互动关系。

8. 总结

大数据赋能“智慧政务”平台是未来政务工作发展的趋势。本文利用文本大数据挖掘技术对政务工作的群众留言分类，热点问题挖掘以及政府答复意见评价进行建模分析，对政府政务治理能力提升提供借鉴意义。在问题 1 中，我们运用深度学习算法 LSTM 和 Bi-LSTM+Attention，机器学习算法的 Logistics 回归、朴素贝叶斯这四类算法建立一级标签分类模型，模型的精准度、召回率和 F-Score 均在 0.8 以上，其中 Bi-LSTM+Attention 算法

表现最好，比普通 LSTM 算法的 F-Score 高出 9.5%，达到 0.92。因此，我们建议政府在进行群众留言分类时可使用该算法提高效能。在问题 2 中，我们利用 DBSCAN 聚类分析方法建立留言归类模型，并运用 AHP-EWM 法进行评价，最后输出排名前 5 的热点问题分别为（1）A 市伊景园滨河苑捆绑车位销售；（2）A2 区丽发新城附近建搅拌站噪音扰民；（3）A 市辉煌国际城二期商铺非法营业；（4）A 市经济学院强制学生实习；（5）A 市加快基础设施建设全覆盖。。在问题 3 中，我们运用 TOPSIS 综合评价模型对答复意见进行评价，结果显示，政府答复意见的完整性和相关性方面都做得令人满意，但答复民众问题的不够及时仍然是日后需要继续改进的。

参考文献

- [1] 段江娇、刘红忠、曾剑平，“中国股票网络论坛的信息含量分析”，《金融研究》，2017 年第 10 期，第 178—192 页。
- [2] 李春林、冯志骥，“基于文本挖掘的新能源汽车用户评论研究”，《特区经济》，2020 年第 04 期，第 148—151 页。
- [3] 梁昌明、李冬强. 基于新浪热门平台的微博热度评价指标体系实证研究[J]. 情报学报, 2015, 34(12): 1278-1283.
- [4] 茆晓军、王军锋、刘兴钊：《基于梯度下降法的 ISAR 最小熵相位校正算法》[J],《现代雷达》2008 年第 01 期。
- [5] 孟天广、李锋，“网络空间的政治互动:公民诉求与政府回应性——基于全国性网络问政平台的大数据分析”，《清华大学学报(哲学社会科学版)》，2015 年第 03 期第 30 卷，第 17—29 页。
- [6] 蒲艳萍、胡倩、王凤，“网络问政回应时效的影响研究——来自“零距离问政”平台的证据”，《领导科学》，2020 年第 06 期，第 43—46 页。
- [7] 邵梓捷、杨良伟，““钟摆式回应”:回应性不足的一种解释——基于 S 市地方领导留言板的实证研究”，《经济社会体制比较》，2020 年第 01 期，第 114—122 页。
- [8] 沈艳、陈赟、黄卓，“文本大数据分析在经济学和金融学中的应用:一个文献综述”，《经济学(季刊)》，2019 年第 04 期第 18 卷，第 1153—1186 页。
- [9] 姚俊峰、杨献勇、彭小奇、张田、郑顺斌：《基于混沌变量的变步长梯度下降优化算法》[J],《清华大学学报(自然科学版)》2003 年第 12 期。
- [10] 郑家昊、李庚，“省级政府门户网站政务服务能力建设现状及其改进——基于中国 31 个省级政府门户网站工作年度报表的分析”，《中国行政管理》，2019 年第 03 期，第 104—110 页。
- [11] B HNING D. Multinomial logistic regression algorithm[J]. Annals of the Institute of Statistical Mathematics, 1992, 44(1): 197-200.
- [12] Hillert, A. H. Jacobs, and S. Muller, Media Makes Momentum, The Review of Financial Studies, 2014, 27 (12), 3467-3501
- [13] Hochreiter S, Schmidhuber J. Long short-term memory, Neural Computation 9, 1735-1780(1997).
- [14] Ye Q, Law R, Gu B, et al. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings[J]. Computers in Human Behavior, 2011, 27(2): 634-639
- [15] Zhou P, Shi W, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// Meeting of the Association for Computational Linguistics. 2016.