智慧政务中的文本挖掘应用

摘要

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。随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

第一问：本问解决的是一个分类问题，由已知道分类结果的数据构造一个训练模型，然后用于后面一系列数据的预测，最后用 F-Score 这个标准来评价所构建的模型，首先对数据进行了去停用词和词向量嵌入这两步预处理操作，在删除停用词后使得数据集大小减小，训练模型的时间也减少；删除停用词可能有助于提高性能，因为只剩下更少且唯一有意义的词。因此，它可以提高分类准确性。词向量的嵌入提高了文本的相似程度，进而提高训练的效率，在深度模型的选择上，文章选择了常用的文本分类方法 cnn、lstm 以及混合方法，文本分类的 CNN 提取类似于 n-gram 的特征。忽略了词序，所以在词序不敏感的场景效果很好，一般 CNN 是一个很强的 baseline，LSTM 可以捕捉到序列信息，在情感分析这种词序很重要的应用场景中效果更好，但是通过本文的实验得出基于多层的 cnn 可以更好的提取出文本特征，进而实现文本的精确分类。

第二问：本问所要解决的有两个问题：将给出的所有话题文本按照一定的标准实现相同话题的聚类，然后得出一个热度值评价指标对每一大类话题进行排序，首先进行了第一问的文字预处理，减轻分类模型的运算时间，然后用卷积神经网络进行粗分类，再进行人工细分，根据时间得到一个决对点赞数，最后通过热度评价得到每一类话题的具体热度，

第三问：三问需要将具体给出的相关性、可解释性、完整性以及时效性进行数字量化，因此，运用算法对留言数和留言回复两条文本进行相似度的判别，接着运用层次分析法将每一个指标对应的权重进行求解，最后得到每一条留言回复的评分值，然后通过一个分数指标得出政府留言内容的具体情况。

关键词：词向量；F-Score 标准；聚类；卷积神经网络；LSTM；文本相似度；相关性；可解释性；完整性

Text mining application in smart government

abstract

In recent years, Wechat, Weibo, Mayor's mailbox, Sunshine hotline and other online government platforms are gradually becoming important channels for the government to listen to the public. The amount of text data related to various social problems and public opinions is huge and still increasing. With the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend, which now a days plays an important role in upgrading the government management and efficiency.

The first question is to solve a classification problem. A training model is constructed from the data that already knows the classification results, and then it is used for the prediction of the following series of data. Finally, the model is evaluated by the standard of F-score. Firstly, the data

is preprocessed by two steps: deactivation words and word vector embedding. After the deactivation words are deleted, the data set size is reduced and the training is carried out. The time to practice the model is also reduced due to deleting the stop words may help improve performance, as there are fewer and only meaningful words left. Therefore, it can improve the accuracy of classification. The embedding of word vector improves the similarity degree of text, and then improves the efficiency of training. In the selection of depth model, the paper selects the common text classification methods CNN, LSTM and hybrid methods. The CNN of text classification extracts features similar to n-gram. It ignores word order, so the effect is very good in the scene where word order is not sensitive. Generally, CNN is a strong baseline, LSTM can capture sequence information, and the effect is better in the application scenario where word order is very important for emotional analysis. However, through the experiment in this paper, it is concluded that multi-layer CNN can better extract text features, and then achieve accurate text classification.

There are two problems to be solved in the second question: cluster all the topic texts according to certain standards, and then get a calorific value evaluation index to rank each major topic. First, preprocess the first question, reduce the operation time of the classification model, and then use convolutional neural network for rough classification, and then carry out artificial classification Subdivide, get a decision like number according to time, and finally get the specific heat of each topic through heat evaluation.

The third question needs to quantify the specific relevance, interpretability, integrity and timeliness. Therefore, the algorithm is used to distinguish the similarity between the number of messages and the number of message replies. Then the AHP method is used to solve the weight corresponding to each index. Finally, the score value of each message reply is obtained, and then a score is used Index to get the specific situation of the content of the government message.

Keywords: word vector; F-score standard; clustering; CNN; LSTM; text similarity; relevance; interpretability; integrity

1.问题背景

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，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。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

给出了收集自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。请利用自然语言处理和文本挖掘的方法解决所给问题。

2.符号定义

2.1 符号定义

符号	意义	符号	意义
F_1	分类得分率	RI	随机一致性指标
n	样本总数	CR	检验系数
R_i	查全率	q	话题热度值
P_i	查准率	A	句子 A
d	d 维词向量	B	句子 B
b	偏置项	I	因素 I
w	滤波器	j	因素 j
loss	损失函数	i	热度周期
train	训练集	N	文本总数
test	测试集	N(x)	文本中出现的次数
x_0	平均绝对点赞数	CI	完全一致性指标

3.问题一模型求解及算法设计

3.1 问题定义

针对群众留言分类问题，在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，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，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

通常使用 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

3.2 数学模型建立及求解

3.2.1 问题一数学模型建立

微信、微博、市长信箱、阳光热线等文本由于自身的特点和信息量，一段文本会描述一个特定的主题。其中既有短文本长文本类型，但涉及的长文本类型并非数千字的文本，我们也将其归。短文本由于其自身长度的原因是缺少这种特征的。一般对于短文本的处理会借鉴上料或者同义词来扩充短文本的含义。但是由于文本的领，上下文的语料和同义词的分布未必和原始语料。

在深度学习领域，对于短文本分类 CNN 是一种常见的方法，但是这种方法通常需要大规模的语料。文中认为短文本的多标签任务所面临的的问题主要是由多标签带来的数据稀疏的问题。在多标签分类中，我们会遇到标签是平行和结构的情况，例如我们看新闻的时候有生活类，科技类，娱乐类，而娱们采用了 N的 LSTM 来进行对比。为了体现实验的合理性，我们同时设置了用以对照。

3.2.2 问题一卷积神经网络模型原理求解

网络获取的文字数据，需要转换为电脑所能识别的机器数据，所以在对文本进行分类前，需要对文本进行一些预处理工作。文本预处理包括分词、去停用词、词向量嵌入、建立文本表示模型等步骤。经过预处理步骤后，使用训练集训练分类器，使用测试集验证分类器效果，文本讨论了典型的高效处理方式：停用词、词向量嵌入，因此接下来文章进行详细介绍：

1、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。但是，并没有一个明确的停用词表能够适用于所有的工具。甚至有一些工具是明确地避免使用停用词来支持短语搜索的。

人类语言包含很多功能词。与其他词相比，功能词没有什么实际含义。最普遍的功能词有语气助词、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。这些功能词的两个特征促使在搜索引擎的文本处理过程中对其特殊对待。第一，这些功能词极其普遍。记录这些词在每一个文档中的数量需要很大的磁盘空间。第二，由于它们的普遍性和功能，这些词很少单独表达文档相关程度的信息。如果过程中考虑每一个词（stopword）。称它们为停用词是因为在文本处理过程中如果遇到它们，则立即停止处理，将其扔掉，增加了检索效率，并且通常都会提高检索的效果，甚至像 的搜索引擎也会删除停用词，以便从数据库中快速地检索数据

有以下几种方法去停用词：使用 NLTK 删除停用词；NLTK 是文本预处理的自然语言工具包。此方法具有 16 种不同语言的停用词列表；使用 spaCy 删除，spaCy 是 NLP 中功能最多，使用最广泛的库之一。可以使有效地从给定文本中删除停用词；使用 Gensim 删除停用词，Gensim 是一个非常方便的库，可以处理 NLP 任务。在预处理时，

gensim 也提供了去除停用词的方法；使用 TextBlob 进行文本标准化，可以使用 TextBlob 来执行词形还原。但是，TextBlob 中没有用于词干化的模块。

基于获取的文本数据，综合比较以上几类方法，以及参考多篇文献，本文构建了的 STW 方法作为去停用词的方法，此方法在删除停用词后使得数据集大小减小，训练模型的时间也减少；删除停用词可能有助于提高性能，因为只剩下更少且唯一有意义的词。因此，它可以提高分类准确性。

2、词向量嵌入：

传统方法使用索引的方式无法表达词之间的相似性，n 元模型在很多场合难以取得明显的进步和表现。one-hot 存在维度方面的问题以及无法表示词和短语之间的相似性。

WordNet 是一个由普林斯顿大学认识科学实验室在心理学教授乔治 A 米勒的指导下建立和维护的英语字典。由于它包含了语义信息，所以有别于通常意义上的字典。WordNet 根据词条的意义将它们分组。WordNet 为每一个 synset 提供了简短，概要的定义，并记录不同 synset 之间的语义关系。首先它既是一个字典，又是一个辞典，它比单纯的辞典或词典都更加易于使用。它还能支持自动的文本分析以及人工智能应用，因此本文选用 WordNet 进行词向量的嵌入。

网络获取的文字数据，需要转换为电脑所能识别的机器数据，所以在对文本进行分类前，需要对文本进行一些预处理工作。文本预处理包括中文分词，去停用词、词向量嵌入、建立文本表示模型等步骤。经过预处理步骤后，使用训练集训练分类器，使用测试集验证分类器效果，文本讨论了典型的高效处理方式：停用词、词向量嵌入，因此接下来进行详细介绍：

1、停用词：

人类语言包含很多功能词。与其他词相比，功能词没有什么实际含义。最普遍的功能词有语气助词、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。这些功能词的两个特征促使在搜索引擎的文本处理过程中对其特殊对待。第一，这些功能词极其普遍。记录这些词在每一个文档中的数量需要很大的磁盘空间。第二，由于它们的普遍性和功能，这些词很少单独表达文档相关程度的信息。如果在检索过程中考虑每一个词而不是短语，这些功能词基本没有什么帮助。在信息检索中，这些功能词的另一个名称是：停用词（stopword）。称它们为停用词是因为在文本处理过程中如果遇到它们，则立即停止处理，将其扔掉，增加了检索效率，并且通常都会提高检索的效果，甚至像 Google 这样的搜索引擎也会删除停用词，以便从数据库中快速地检索数据。

一般处理文本有以下几种方法去停用词：使用 NLTK 删除停用词；NLTK 是文本预处理的自然语言工具包。此方法具有 16 种不同语言的停用词列表；使用 spaCy 删除停用词，spaCy 是 NLP 中功能最多，使用最广泛的库之一。可以使用 SpaCy 快速有效地从给定文本中删除停用词；使用 Gensim 删除停用词，非常方便的库，可以处理 NLP 任务。在预处理时，gensim 也提供了去除停用词的方法；使用 TextBlob 进行文本标准化，可以使用 TextBlob 来执行词形还原。但是，TextBlob 中没有用于词干化的模块。

基于获取的文本数据比较简洁，综合比较以上几类方法，以及参考多篇文献，本文构建了自己的去停函数用来作为去停用词的方法，停用词库选取微博大数据统计停用词库。此方法在删除停用词后使得文本规模减小，同时最大程度保留了短文本的特征，训练模型的时间也减少；删除停用词可能有助于提高性能，因为只剩下更少且唯一有意义的词。因此，它可以提高分类准确性。

2、词向量嵌入

传统方法使用索引的方式无法表达词之间的相似性，n 元模型在很多场合难以取得明显的进步和表现。one-hot 存在维度方面的问题以及无法表示词和短语之间的相似性。并且使

用 one-hot 方式在生词多的情况下构建的样本属于超稀疏矩阵，严重影响模型收敛性。词向量是基于大数据的统计，统计样本来源于选定数据集。首先对文本进行分词，如果是拉丁语系可以跳过这一步，如果是东方语系，则分词必不可少。接下来统计每个词附近，其他词出现的频率，作为这个词的词向量。词向量的维度可由人工设定，假如词向量维度为 100，则表示统计所选词附近一百个常用词出现的频率。

国内外实验室研究了不同的词向量嵌入方法，当前最常用的为 Gensim 库的 Word2vec 工具包，这也是本文选用的词向量训练方法，具体过程即代码见附件。

1.输入层

首先我们输入一个一维的由 7 个词构成的句子，为了使其可以进行卷积，首先需要将其转化为二维矩阵表示。d=5 表示每个词转化为 5 维的向量。

2.卷积层

在处理图像数据时，CNN 使用的卷度和高度的一样的，但是在 text-CNN 中，卷积核的宽度是与的维度一致。因为我们输入的每一行向量代表一个词，在抽取特征的过程中，词做为文本的最小粒度，如果我们使用卷积度小于词向量的维度就已经不是以词作为最小粒度了。而高行设置（通常取值 2,3,4,5）。由于我们的输入是一个句子，句子中相邻的词之间关联性很高，因此，当我们用卷积核进行卷积时，不仅考虑了词义而且考虑了词序及其上下文。（类似于 skip-gram 和 CBOW 模型的思想）。

详细讲解卷积的过程：卷积层输入的是一个表示句子的矩阵，维度为 n*d，即每句话共有 n 个词，每个词有一个 d 维的词向量表示。假设 $X_{i:i+j}$ 表示 X_i 到 X_{i+j} 个词，使用一个宽度为 d，高度为 h 的卷积核 W 与 X_{i+j} (h 个词)进行卷积操作后再使用激活函数激活得到相应的特征 c_i，则卷积操作可以表示为：（使用点乘来表示卷积操作）

$$c_i = f(w \cdot x_{i:\ i+h-1} + b) \tag{2}$$

因此经过卷积操作之后，可以得到一个 n-h+1 维的向量 c 形如：

$$c = [c_1, c_2, ...c_{n-h+1}] \tag{3}$$

以上是一个卷积核与输入句子的卷积操作，同样的，我们可以使用更多高度不同的卷积核，且每个高度的卷积核多个，得到更多不同特征。 如上图，这里有（2，3，4）三种 size，每种 size 有两个 filter，一共有 6 个 filter。然后开始卷积，从图中可以看出，stride 是 1，对于高是 4 的 filter，最后生成 4 维的向量。对于高是 3 的 filter，最后生成 5 维的向量。

每个卷积块是两个卷积层，每个卷积层后面有一个批处理范数和一个 ReLU 非线性层 Padding 以保留(或在本地池化时减半)维度。

我们在处理第一问分类问题的时候，根据以上步骤采用深的卷积网络用于文本分类即多层卷积网络来进行处理。

多次卷积池化操作得到最终特征向量： $z = [\hat{c}_1, \hat{c}_2, ...\hat{c}_m]$（假设 m 个 filters w）。

最后通过的一个简单的 softmax 层：

$$y = soft\max(W^{(s)} z + b) \tag{4}$$

3.2.3 问题一 LSTM 模型求解

长短期记忆（Long short-term memory, LSTM）是一种特殊的循环神经网络，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。简单来说，就是相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现，

LSTM 拥有如图这种链状结构，但是重复的模块则拥有不同的结构。在图 1 中，每一行都带有一个向量，该向量从一个节点输出到其他节点的输入。粉红色圆圈表示点向运算，如向量加法、点乘，而黄色框是学习神经网络层。线的合并表示连接，而线的交叉表示其内容正在复制，副本将转到不同的位置。LSTM 的关键是细胞状态，表示细胞状态的这条线水平的穿过图的顶部。细胞的状态类似于输送带，细胞的状态在整个链上运行，只有一些小的线性操作作用其上，信息很容易保持不变的流过整个链。LSTM 确实具有删除或添加信息到细胞状态的能力，这个能力是由被称为门(Gate)的结构所赋予的。

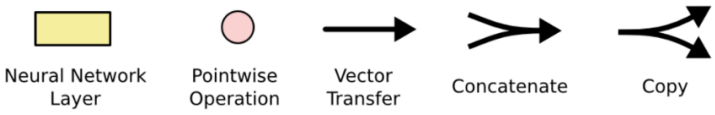


图 1. 图表符号结构

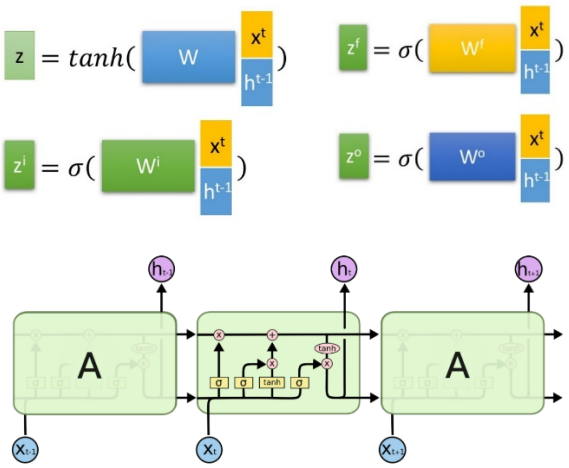


图 2.门控结构图

门(Gate)是一种可选地让信息通过的方式。它由一个 Sigmoid 神经网络层和一个点乘法运算组成。

Sigmoid 神经网络层输出 0 和 1 之间的数字，这个数字描述每个组件有多少信息可以通过，0 表示不通过任何信息，1 表示全部通过，LSTM 有三个门，用于保护和控制细胞的状态。

输入门：处理当前序列位置的输入，确定需要更新的信息，去更新细胞状态。此过程分为两部分，一部分是使用包含 sigmoid 层的输入门决定哪些新信息该被加入到细胞状态；确定了哪些新信息要加入后，需要将新信息转换成能够加入到细胞状态的形式。所以另一部分是使用 tanh 函数产生一个新的候选向量。(可以这么理解，LSTM 的做法是对信息都转为能加入细胞状态的形式，然后再通过第一部分得到的结果确定其中哪些新信息加入到细胞状

态。)

输出门：最后要基于细胞状态保存的内容来确定输出什么内容。即选择性的输出细胞状态保存的内容。类似于输入门两部分实现更新一样，输出门也是需要使用 sigmoid 激活函数确定哪个部分的内容需要输出，然后再使用 tanh 激活函数对细胞状态的内容进行处理

双向 LSTM，有些时候预测可能需要由前面若干输入和后面若干输入共同决定，这样会更加准确。因此提出了双向循环神经网络，网络结构如下图。可以看到 Forward 层和 Backward 层共同连接着输出层，其中包含了 6 个共享权值 w1-w6。

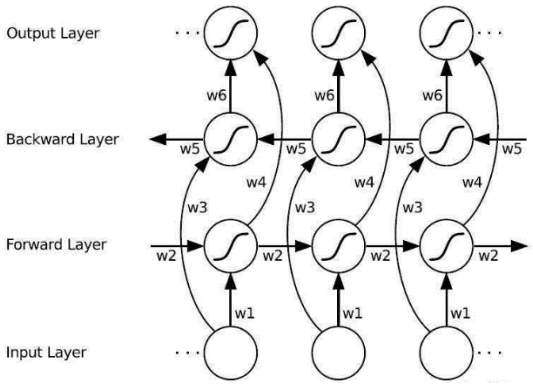


图 3. 双向 LSTM 结构示意图

在 Forward 层从 1 时刻到 t 时刻正向计算一遍，得到并保存每个时刻向前隐含层的输出。在 Backward 层沿着时刻 t 到时刻 1 反向计算一遍，得到并保存每个时刻向后隐含层的输出。最后在每个时刻结合 Forward 层和 Backward 层的相应时刻输出的结果得到最终的输出，用数学表达式如下：

$$\begin{aligned} h_t &= f(w_1 x_t + w_2 h_{t-1}) \\ h_t' &= f(w_3 x_t + w_5 h_{t+1}') \\ o_t &= g(w_4 x_t + w_6 h_t') \end{aligned} \tag{5}$$

3.3 问题一算法设计

针对问题一，一级标签留言分类，整个模型算法流程图，如下图 4 所示。大体分为数据预处理、构建组合模型、训练组合模型和测试组合模型。

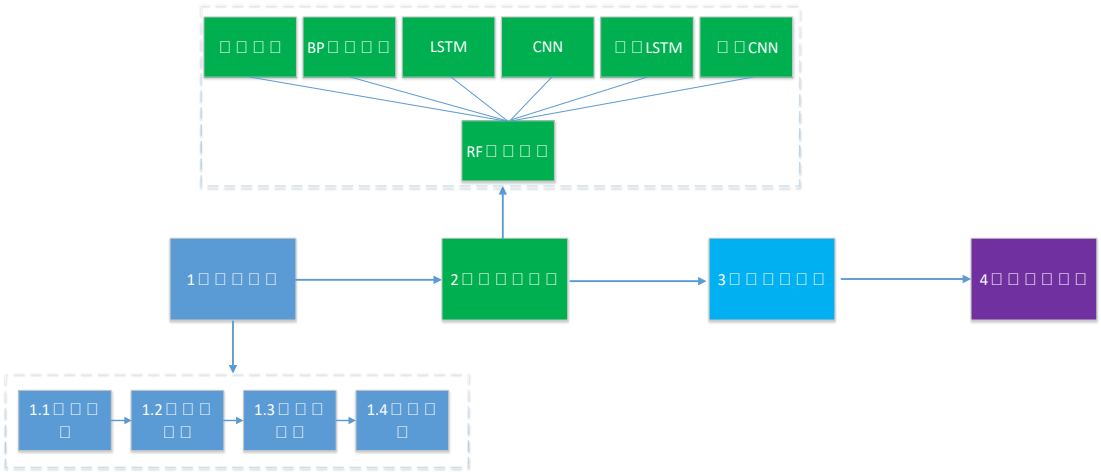


图 4 一级标签分类算法流程图

为了对比实验我们同时使用了 3 种基础网络模型及其变种形式，以及结合随机森林算法的混合网络模型，共计 13 个网络。神经网络模型以生物神经系统为蓝本，以神经元作为构造基础，以激活函数限制输出的阈值，以优化算法更新神经元的权重，最终得到最接近真实输出的模型。BP 神经网络只有最简单的神经元，常用来特征与降低处理维度，常用于图像分割和识别问题。而 LSTM 是基于循环神经网络的改良模型，在 rnn 的基础上保留了 rnn 的序列处理能力，增加了对久远数据的记忆。图 5、图 6 和图 7 分别表示 BP 神经网络结构图、CNN 的拓扑结构图和 LSTM 结构示意图

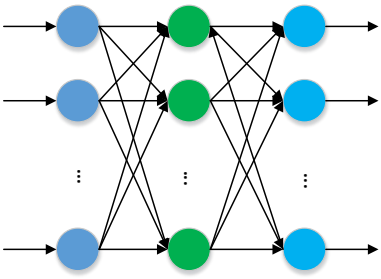


图 5.BP 神经网络结构图

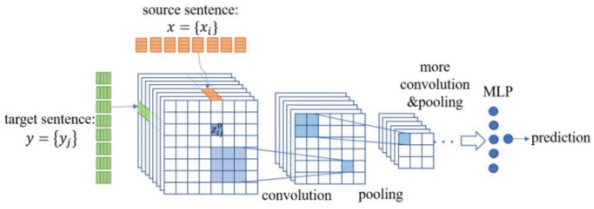


图 6.CNN 的拓扑结构图

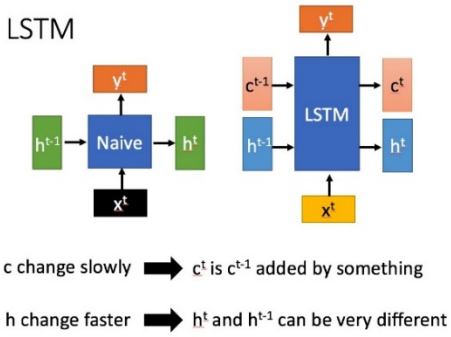


图 7.LSTM 的拓扑结构图

3.3.1 构建网络模型

我们使用 Pandas 库来读写数据；过滤停用词；分词；训练词向量，最后进行词向量转数字的流程来进行数据预处理。如图 8 所示。之后进行网络模型的构建。构建的多层卷积神经网络模型代码如表 2 所示。

图 8 数据预处理详细流程图

表 2. 构建的多层卷积神经网络模型代码

```
def num_cnn(self):#定义我们构建的多层卷积神经网络模型
input_layer = layers.Input(shape=self.X.shape[1])
        emb = layers.Embedding(self.max_nb_words, self.embedding_dim)(input_layer)
        spd = layers.SpatialDropout1D(0.2)(emb)
        #conv layers
        convs = []
        filter_sizes = [3,4,5]
        for fsz in filter_sizes:
            conv = layers.Conv1D(filters=250,
                                 kernel_size=fsz,
                                 padding='valid',
                                 activation='tanh',
                                 strides=1)(spd)
            pool = layers.MaxPooling1D(X.shape[1]-fsz+1)(conv)
            pool = layers.Flatten()(pool)
            convs.append(pool)
        merge = layers.concatenate(convs,axis=1)
        drop_1 = layers.Dropout(0.2)(merge)
        dense = layers.Dense(64,activation='relu')(drop_1)
        drop_2 = layers.Dropout(0.5)(dense)
        output_layer = layers.Dense(self.Y.shape[1], activation='softmax')(drop_2)
        model = Model(inputs = input_layer, outputs = output_layer)
        return model
```

我们采取构建多层卷积神经网络模型来进行求解问题一一级标签留言。

```
Model: "model"
_________________________________________________________________________________________________
Layer (type)                 Output Shape              Param #     Connected to
=================================================================================================
input_1 (InputLayer)         [(None, 40)]              0
_________________________________________________________________________________________________
embedding (Embedding)        (None, 40, 100)           5000000     input_1[0][0]
_________________________________________________________________________________________________
spatial_dropout1d (SpatialDropo (None, 40, 100)        0           embedding[0][0]
_________________________________________________________________________________________________
conv1d (Conv1D)              (None, 38, 250)           75250       spatial_dropout1d[0][0]
_________________________________________________________________________________________________
conv1d_1 (Conv1D)            (None, 37, 250)           100250      spatial_dropout1d[0][0]
_________________________________________________________________________________________________
conv1d_2 (Conv1D)            (None, 36, 250)           125250      spatial_dropout1d[0][0]
_________________________________________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 1, 250)            0           conv1d[0][0]
_________________________________________________________________________________________________
max_pooling1d_1 (MaxPooling1D) (None, 1, 250)          0           conv1d_1[0][0]
_________________________________________________________________________________________________
max_pooling1d_2 (MaxPooling1D) (None, 1, 250)          0           conv1d_2[0][0]
_________________________________________________________________________________________________
flatten (Flatten)            (None, 250)               0           max_pooling1d[0][0]
_________________________________________________________________________________________________
flatten_1 (Flatten)          (None, 250)               0           max_pooling1d_1[0][0]
_________________________________________________________________________________________________
flatten_2 (Flatten)          (None, 250)               0           max_pooling1d_2[0][0]
_________________________________________________________________________________________________
concatenate (Concatenate)    (None, 750)               0           flatten[0][0]
                                                                   flatten_1[0][0]
                                                                   flatten_2[0][0]
_________________________________________________________________________________________________
dropout (Dropout)            (None, 750)               0           concatenate[0][0]
_________________________________________________________________________________________________
dense (Dense)                (None, 64)                48064       dropout[0][0]
_________________________________________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0           dense[0][0]
_________________________________________________________________________________________________
dense_1 (Dense)              (None, 7)                 455         dropout_1[0][0]
=================================================================================================
Total params: 5,349,269
Trainable params: 5,349,269
Non-trainable params: 0
_________________________________________________________________________________________________
None
```

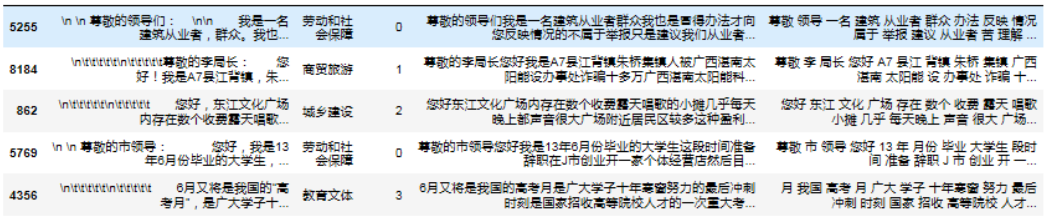
图 9 多层卷积的网络结构模型

3.4 算法分析及结论

表 3 停用词代码

Algorithm content

```
#删除字母，数字，汉子以外的所有符号
def remove_punctuation(line):
    line = str(line)
    if line.strip()==":
        return "
    rule = re.compile(u'[^a-zA-Z0-9\u4E00-\u9FA5]')
    line = rule.sub(",line)
    return line=
def stopwordslist(file):
    stopwords = [line.strip() for line in open(file, 'r', encoding='utf-8').readlines()]
    return stopwords
#加载停用词库:
stopwords = stopwordslist('ChineseStopWords.txt')
#分词，过滤停用词
df['cut_review'] = df['clean_review'].apply(lambda x: ' '.join([w for w in list(jb.lcut(x)) if w
not in stopwords]))
df.head()
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.921 seconds.
Prefix dict has been built successfully.
```



5255	\n \n 尊敬的领导们： \n\n 我是一名建筑从业者，群众。我也...	劳动和社会保障	0	尊敬的领导们我是一名建筑从业者群众我也是百得办法才向您反映情况的不属于举报只是建议我们从业者...	尊敬 领导 一名 建筑 从业者 群众 办法 反映 情况 属于 举报 建议 从业者 苦 理解 ...
8184	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的李局长： 您好！我是A7县江背镇，朱...	商贸旅游	1	尊敬的李局长您好我是A7县江背镇朱桥集镇人被广西湛南太阳能设办事处诈骗十多万广西湛南太阳能科...	尊敬 李 局长 您好 A7 县江 背镇 朱桥 集镇 广西 湛南 太阳能 设 办事处 诈骗 十...
862	\n\t\t\t\t\t\n\t\t\t\t\t 您好，东江文化广场内存在数个收费露天唱歌...	城乡建设	2	您好东江文化广场内存在数个收费露天唱歌的小摊几乎每天晚上都声音很大广场附近居民区较多这种盈利...	您好 东江 文化 广场 存在 数个 收费 露天 唱歌 小摊 几乎 每天晚上 声音 很大 广场...
5769	\n \n 尊敬的市领导： 您好，我是13年6月份毕业的大学生，...	劳动和社会保障	0	尊敬的市领导您好我是13年6月份毕业的大学生这段时间准备辞职在J市创业开一家个体经营店然后目...	尊敬 市 领导 您好 13 年 月份 毕业 大学生 段时间 准备 辞职 J 市 创业 开 一...
4356	\n\t\t\t\t\t\n\t\t\t\t\t 6月又将是我国的"高考月"，是广大学子十...	教育文体	3	6月又将是我国的高考月是广大学子十年寒窗努力的最后冲刺时刻是国家招收高等院校人才的一次重大考...	月 我国 高考 月 广大 学子 十年寒窗 努力 最后 冲刺 时刻 国家 招收 高等院校 人才...

图 10 过滤停用词

图 10 将原始数据中的换行（\n）、冒号（：）、逗号（，）、语气助词（的、都、又）等冗余数据进行了删除，原始的数据得到有效的提取与过滤，对后面分类模型的训练和预测起到了很大的作用，可以减少训练时间，提高模型训练效率，进而获得高精度的分类模型。

针对第一问建立关于留言内容的一级标签分类模型，我们建立了包括单向 LSTM、双向 LSTM+单层卷积、单层卷积+单向 LSTM、TM、单向 LSTM+单层卷积、单向 LSTM+多层卷积、双向 LSTM+多层卷积、多层卷积+单向 LSTM 在内的 13 类模型结构，并进行了仔细的测试和对比。结果如表 4 所示。

表 4 多种预测方法的分类精度

模型类型	词向量嵌入	精度（acc）
BP 神经网络	0	0.81
	1	0.64
单向 LSTM	0	0.69
	1	0.74
双向 LSTM	0	0.79
	1	0.73
单层卷积	0	0.84
	1	0.71
多层卷积	0	0.85
	1	0.71
单向 LSTM+单层卷积	0	0.78
	1	0.78
单向 LSTM+多层卷积	0	0.78
	1	0.77
双向 LSTM+单层卷积	0	0.75
	1	0.78
双向 LSTM+多层卷积	0	0.76
	1	0.75
单层卷积+单向 LSTM	0	0.74
	1	0.75
单层卷积+双向 LSTM	0	0.72
	1	0.75
多层卷积+单向 LSTM	0	0.72
	1	0.71
多层卷积+双向 LSTM	0	0.76
	1	0.73

对于是否使用词向量嵌入的多种分类模型的预测结果，预测精度得到提高的预测方法有4类：单向LSTM、双向LSTM+单层卷积、单层卷积+单向LSTM、单层卷积+双向LSTM，精度下降的5类：BP神经网络、双向LSTM、单层卷积、多层卷积、多层卷积+双向LSTM精度没有变化的有4类：单向LSTM+单层卷积、单向LSTM+多层卷积、双向LSTM+多层卷积、多层卷积+单向LSTM，从统计结果可以看出词向量的嵌入在对单项的LSTM预测精度中具有显著的提高作用，在 LSTM 混合卷积的神经网络中效果就不是很明显，在只有卷积神经网络中还降低了预测精度，词向量的嵌入是转意义，也就是相似含义的词在空间中的距离更近，所以能够提高 LSTM 的精度，但是受限与本文的数据导致卷积神经网络精度没有得到有效提高。

图11所示为在训练集和测试集上损失函数的变化。

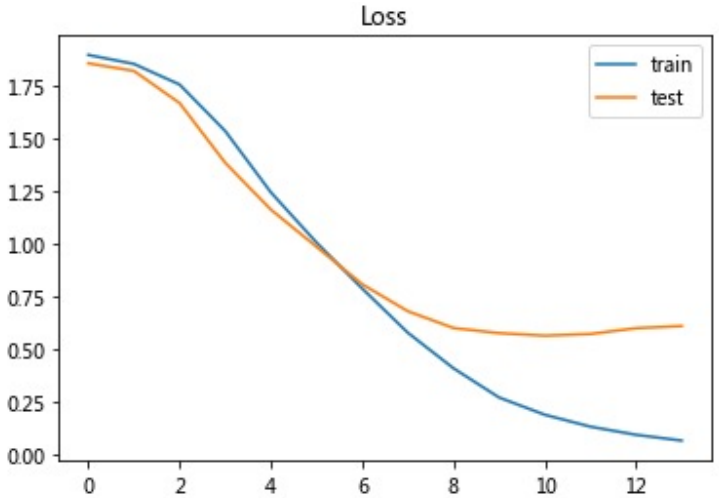
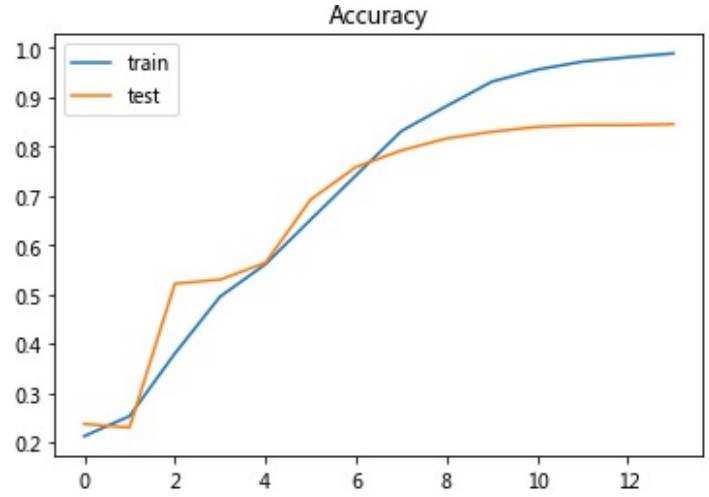


图11 训练和测试的损失函数

图12所示为在训练集和测试集上一级标签分类的准确率的变化。



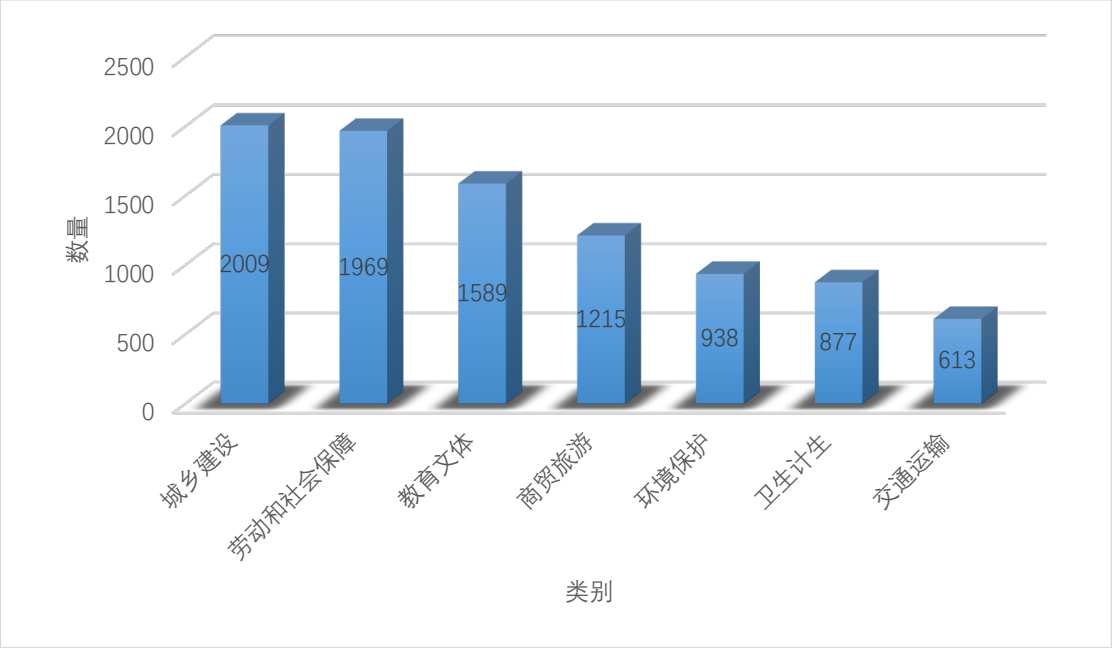


图 13 分类结果

本文选取了多层神经网络进行模型的搭建与预测。

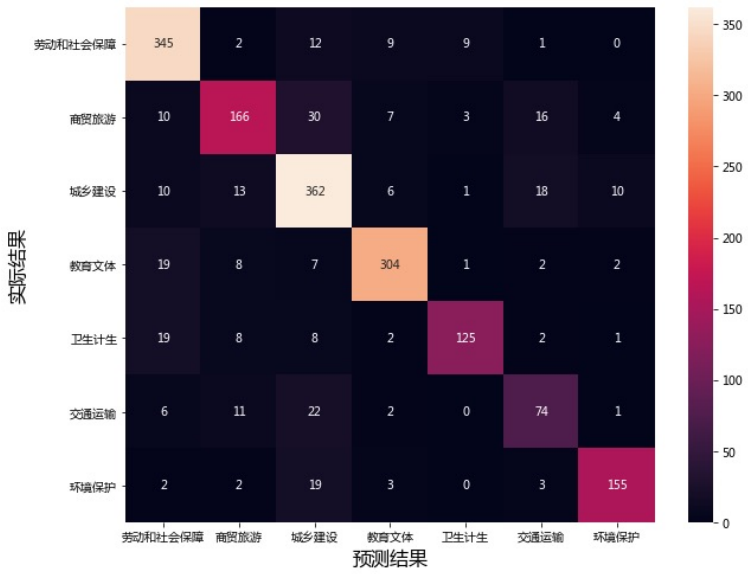


图 14 多层 CNN 预测的混淆矩阵结果

图 14 表示了为 7 类标签预测的混淆矩阵结果，从图中，实际结果为 378 个的劳动和社会保障一级标签结果，我们正确预测结果达到了 345 个。

4.问题二模型求解及算法设计

4.1 问题定义

针对热点问题挖掘问题，某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映 入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关 部门进行有针对性地处理，提升服务效率。请根据附件 3 将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 5 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 6 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”。

表 5.热点问题表

热度排名	问题 ID	指数热点	时间范围	地点/人群	问题描述
1	A	…	2019/08/18-2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	B	…	2017/06/08-2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

表 6.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	点赞数	反对数
1	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0

1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
1	360109	A0080252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这样搞……	0	0
…	…	…	…	…	…	…	…
2	360114	A0182491	A 市经济学院变相强制实习	2017/06/08 17:31:20	系里要求我们在实习前分别去指定的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口，不做满 6 个月不给实训分，不能毕业……	9	0

4.2 数学模型建立及求解

4.2.1 问题二数学模型建立

某一时段内反映特定地点或特定人群问题的留言进行归类，我们需要定义合理的热度评价指标，并给出评价结果。

首先我们需要制定是否能够作为热点问题的规则：

1、热门问题的热度数值是根据该留言问题的点赞数、反对数和问题留言时间等各项因素，算出热度基数，和热度权重相加，得出最终热度值。

3、所发留言问题反对数。

热度公式的计算思路，受银行本金、现值和终值启发，考虑到复利记息法，一笔资金的现值和终值的区别是在周期范围内利息的叠加，即利息算在下一个周期的本金中。类比银行利息算法，我们采取的热度计算与银行利息算法不同的是，我们需要考虑时间成本和衰减率因素。计算每一期的热度增长到期末的终值的叠加，使得热度数值终值的计算为每一期热度的现值到期末的终值的叠加，在此基础上构造了，我们的热度数值计算公式。

我们以绝对点赞数作为热度值计算公式中的变量数值，构造热度计算公式：

$$\sum_{i=1}^{n} x_0 \cdot (1-r)^{n-i} \tag{6}$$

式中，x_0表示每一期的平均绝对点赞数，绝对点赞数为留言数＋点赞数－反对数。i为热度周期，n为热度总期数。表示影响热度数值因素的权重值。

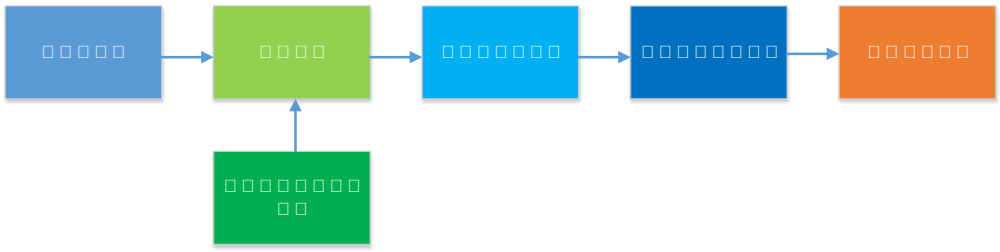


图 15 热点问题算法流程图

文本相似度的计算广泛的运用在信息检索，搜索引擎，文档复制等处：因此在各种不同的情况与任务中，有不同的文本相似度计算。

TF-IDF（项频率与逆文本频率）

TF 代表了词语在文档中出现的频率，一个词的权重由 TF * IDF 表示，其中 TF 表示词频，即一个词在这篇文本中出现的频率；IDF 表示逆文档频率，即一个词在所有文本中出现的频率倒数。当进行索引的时候，词语出现频率较高的文本，匹配度也会较高，但是某些停止词，例如 to 在文本中会出现相当多的次数，但这对匹配并没有起到很好的索引作用，因此需要引入另一个度量值 IDF（逆文本频率）。

$$IDF = \log \frac{N}{N(x)} \tag{7}$$

其中 N 为语料库中文本的总数，$N(x)$为文本中出现单词 x 的文本数量。可以度量该单词的重要程度。某些特殊情况下，x并未出现在语料库中（所有文本），则需要考虑将公式平滑为：

$$IDF = \log \frac{N+1}{N(x)+1} + 1 \tag{8}$$

最终的 TF-IDF 值为：

$$TF - IDF(x) = TF(x) * IDF(x) \tag{9}$$

受 TF-IDF（项频率与逆文本频率）计算文本相似度启发，我们提出基于 F1-score 的 F1-

相似度：对于给定文本 A 和 B，假设 A 中字符在 B 中的概率为 P，B 中字符在 A 中概率为 Q，则 F1-相似度=2PQ/(P+Q)。

4.3 算法设计

4.3.1 基于聚类的话题分类

计算机只能对结构化的数据进行处理，因此必须将非结构化的留言文本转化为结构化的数据。一般采用向量空间模型(VSM)技术，将每一个用户的内容文本映射为向量空间中的一个向量，向量由该文本的特征项组成。对群众留言中的每个话题，首先对其内容文本进行中文分词和过滤停用词，然后选择名词、简称、缩略语以及命名实体来建立相应的初始特征集合。其中由于留言主题一般直接代表了话题的主要内容，对在主题中出现过的词条(除停用词)，不论其词性征项。

用户的话题识别是为了政府人员进行热点话题检测与跟踪的基础，话题识别是指识别谈论同一话题的留言。已有的研究主要采用 K -Means 或 Single-Pass 聚类算法来进行话题识别，聚为一类的文本被认为有相同的话题，但是这些算法都属于硬聚类，因此本文提出基于 F1-近似度的聚类方法，得到的是一个给定对象属于一个类的程度，因此能够合理地改进硬聚类算。这里第一步进行粗分类，代码如下：

表 7 粗分类代码

Algorithm content

```
d = dict() #储存[留言数,点赞数,反对数]
for i in range(ndf.shape[0]):
    #跳过空评论
    if len(str(ndf['clean_review'][i])) == 0:
        continue
    else:
        if i == 0:
            d[ndf['clean_review'][i]] = np.array([1,ndf['点赞数'][i],ndf['反对数'][i]])
        else:
            if ndf['clean_review'][i] in d.keys():
                continue
            else:
                p = [] #储存当前留言跟字典中已有留言的相似度
                key = list(d.keys())
                #计算当前留言与字典中已有留言的相似度
                for w in key:
                    p.append(comp(ndf['clean_review'][i],w))
                if max(p) < 0.5:
                    d[ndf['clean_review'][i]] = np.array([1,ndf['点赞数'][i],ndf['反对数'][i]])
                else:
                    for pp in range(len(p)):
                        if p[pp] >= 0.5:
                            d[key[pp]] = d.get(key[pp]) + np.array([1,ndf['点赞数'][i],ndf['反对数'][i]])
                        else:
                            continue
```

4.3.2 热度的话题排序

将文本进行分类以后，再统计综合话题包含的各项信息（点赞数、反对数、留言时间），对相应的话题进行热度评分，最后输出得分最高的若干个话题，即为热点话题。考虑到热度的衰减和利息的计算方式类似，区别是，利息为叠加，热度为叠减。某一类的绝对点赞数为该类的。不妨令绝对点赞数为该类的初值，考虑到时间成本和热度衰

表 8 话题热度排序

Algorithm content

#计算热度
new_dict = hot_degree(new_array,t,0.5)
new_dict
new_list = sorted(new_dict.items(),key=lambda x:x[1][0],reverse=True)
new_list[:50]
new_list_2 = []
for i in range(len(new_list)):
  new_list_2.append([new_list[i][0],new_list[i][1][0],new_list[i][1][1]])
new_list_2 = np.array(new_list_2)
result_1 = pd.DataFrame(new_list_2,columns=['留言主题','留言热度','留言时间'])
result_1.to_excel('留言热度排序.xlsx')

4.3.2 话题热度数值计算

我们进行所有话题排序之后对话题进行热度数值的计算，以绝对点赞数作为热度值计算公式中的变量数值。公式如 10 所示，代码如下

表 9 热度数值排序

```
def hot_degree(A,t,r):
#计算热度数值排序，A 为所选类的样本，t 为指定终止时间
time_t = datetime.datetime.strptime(t,'%Y/%m/%d %H:%M:%S')
    d = dict()
    for i in range(len(A)):
        try:
            time_a = datetime.datetime.strptime(A['留言时间
'][i],'%Y/%m/%d %H:%M:%S')
        except:
            time_a = datetime.datetime.strptime(A['留言时间
'][i],'%Y-%m-%d %H:%M:%S')
        days = (time_t - time_a).days
        months = (days)//30 #计算月数
        if months < 0:
            continue
        elif months == 0:
            ht = 0
        else:
            ht = 0
            for j in range(1,months+1):
                ht += (eval(A['绝对点赞数'][i])/months) * ((1-r)**(months-j))
        d[A['留言主题'][i]] = [ht,A['留言时间'][i]]
    return d
```

$$\sum_{i=1}^{n} x_0 \cdot (1-r)^{n-i} \tag{10}$$

式中， x_0表示每一期的平均绝对点赞数，绝对点赞数为留言数＋点赞数－反对数。 i为热度周期， n为热度总期数。 r表示影响热度数值因素的权重值。

4.4 算法分析及结论

4.4.1 粗分类结果

在对文本进行分类的时候不是选用某一个模型就能得到很好的预期结果，于是本文采取先进行粗分类，在获得原始数据的基础上采用第一问所用到的 LSTM、多层卷积神经网络等学习深度模型进行分类，从得到的结果进行综合可以判断并筛选出 F1-相似度模型，更加适合本文所涉及的文本分类，由于文章的篇幅有限，本文列举了 6 类具有典型代表性的话题结果如表 10 所示。

表 10

留言主题	点赞数	反对数	留言时间
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2097	0	2019/8/19 11:34:04
A5 区五矿万境 K9 县的开发商与施工方建房存在质量问题	2	0	2019/5/5 13:52:50
A 市五矿万境 K9 县存在严重的消防安全隐患	1	0	2019/9/12 14:48:07
A 市五矿万境 K9 县交房后仍存在诸多问题	0	0	2019/9/11 15:16:02
A 市五矿万境 K9 县房屋出现质量问题	0	0	2019/9/19 17:14:49
A 市五矿万境 K9 县负一楼面积缩水	0	0	2019/9/10 9:10:22

反映 A 市金毛湾配套入学的问题	1762	5	2019/4/11 21:02:44
关于 A 市金晖优步花园相关问题的反映	9	0	2019/1/8 17:07:55
反映 A 市城市交通存在的诸多问题	8	0	2019/3/24 20:40:57
反映 A 市南雅中学初中秋游问题	2	0	2019/10/20 18:47:09
反映 A 市恒大江湾退房退款问题	1	0	2019/3/16 18:00:37
反映 A 市丁字湾街道规划问题	1	0	2019/9/2 22:20:10
反映 A6 区丁字湾的雅礼丁姜学校小学部建设问题	0	0	2019/9/19 18:23:48
反映 A 市枫树山大桥小学午餐的问题	0	0	2019/9/2 15:37:23
反映 A 市鑫华驾校的一些问题	0	0	2019/11/27 17:19:28
反映 A4 区金鹰小学生上学交通问题	0	0	2019/4/23 17:06:31
反映 A 市金座雅居的一些问题	0	0	2019/7/5 16:14:12
反映 A 市山水湾孩子上学问题	0	0	2019/12/31 17:06:53

反映 A 市人才补贴问题	0	0	2019/10/29 19:32:19
反映 A 市万科金域蓝湾物业收费等问题	0	0	2019/7/8 14:52:21

请书记关注 A 市 A4 区 58 车贷案	821	0	2019/2/21 18:45:14
承办 A 市 58 车贷案警官应跟进关注留言	733	0	2019/3/1 22:12:30
严惩 A 市 58 车贷特大集资诈骗案保护伞	790	0	2019/2/25 9:58:37

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	669	0	2019/9/5 13:06:55
A4 区绿地海外滩小区距渝长厦高铁太近了	1	0	2019/8/23 14:21:38

A 市富绿物业丽发新城强行断业主家水	242	0	2019/6/19 23:28:27
A 市新城国际都花物业公司限制业主买水	0	0	2019/8/28 16:52:30

关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	78	0	2019/3/26 14:33:47
关于 A6 区月亮岛路 110kv 高压线的建议	55	2	2019/4/9 17:10:01
关于 A6 区月亮岛路沿线架设 110KV 高压电线杆的投诉	10	0	2019/3/26 10:17:31
关于 A6 区月亮岛路沿线架设 110kv 高压电线杆的投诉	5	0	2019/4/3 17:36:58
A6 区月亮岛路 11 万伏高压线没用地埋方式铺设	2	0	2019/4/5 13:01:17

4.4.2 绝对点赞数

在通过了粗分类、细分类以后，就需要对点赞数和留言时间做一定的处理，一条留言被用户所提出来以后就会出现在网络上，如果有类似情况的用户会对其进行点赞评价，得到的点赞数越多能说明此观点是一个热点问题，但是水这时间的过去一个观点如果没有足够的用户去关注评价，其热度就会随着时间的延长而被忘记或者次问题已经得到了有效的解决了，这样的话就不应该得到和没有被解决问题同样的热度，不然会造成有效政府资源的浪费，因此，本文提出了一个绝对点赞数作为一个热度的计算基数，绝对点赞数是有点赞数和时间两个变量综合所得出的一个指标，可以更加准确高效的反映出某个话题的热度，对政府工作人员的工作具有一定的参考价值，绝对点赞数由下表所示。

表 11 绝对点赞数排序表

	留言主题	绝对点赞数	留言时间
0	A市A5区汇金路五矿万境K9县存在一系列问题	2106	2019/8/19 11:34:04
1	反映A市金毛湾配套入学的问题	1792	2019/4/11 21:02:44
2	请书记关注A市A4区58车贷案	2347	2019/2/21 18:45:14
3	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	672	2019/9/5 13:06:55
4	A市富绿物业丽发新城强行断业主家水	244	2019/6/19 23:28:27
5	关于A6区月亮岛路沿线架设110kv高压线杆的投诉	153	2019/3/26 14:33:47
6	请A市加快轨道交通建设力度	121	2019/3/1 15:19:28
7	建议西地省尽快外迁京港澳高速城区段至远郊	111	2019/1/10 15:01:26
8	反映A市地铁号线松雅湖站点附近地下通道问题	89	2019/5/22 23:37:38
9	问问A市经开区东六线以西泉塘昌和商业中心以南的有关规划	136	2019/1/11 15:46:04
10	关于加快修建A市南横线的建议	80	2019/5/10 18:01:52
11	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	79	2019/1/16 17:01:25
12	A市三一大道全线快速化改造何时启动	70	2019/9/15 15:31:19
13	A3区郝家坪小学什么时候能改扩建	67	2019/3/24 21:07:12
14	关于加快修建A市南横线的建议	106	2019/5/10 18:01:52
15	请A4区教育局尽快落实发放原A市七中01年后退休教师的文明单	55	2019/10/29 12:42:17
16	A市长房云时代多栋房子现裂缝质量堪忧	54	2019/2/25 15:17:38
17	居住在地铁3号线A7县松雅西地省站西北方向10万民众的心声	54	2019/4/17 11:13:12
18	反映A7县恒基凯旋门小区配套幼儿园公办或者普惠问题	52	2019/5/30 13:46:57
19	A市汽车南站何时能建好	51	2019/3/20 9:20:46
20	希望A市地铁四号线北延线同心路站设在雷峰大道上	48	2019/1/30 23:59:12
21	请问A7县东六路下穿长永高速在10月底能否如期通车呢	47	2019/10/9 10:14:08
22	请解决A7县松雅湖烂尾问题	46	2019/5/23 17:11:30
23	A2区丽发新城附近修建搅拌站污染环境影响生活	46	2020-01-02 00:00:00
24	建议A市经开区收回东六路恒天九五工厂地块打造商业综合体	45	2019/11/8 15:48:07
25	建议A市经开区泉星公园项目规划进一步优化	46	2019/8/12 13:15:05
26	A4区洪山公园的建设计划何时才能正式启动	43	2019/7/25 9:30:02
27	建议加大A7县东六线榔梨段拆迁力度	48	2019/1/17 19:25:45
28	关于A6区月亮岛路沿线架设110kv高压线杆的投诉	121	2019/3/26 14:33:47
29	请问A7县星沙圣力华苑出售地下车位是否合法	42	2019/3/11 14:52:12
30	建议A7县在漓楚路和东六路交汇处建地下通道或人行天桥	42	2019/5/26 18:23:42

经过高效的深度学习模型的粗分操作以后，再进行人为的细分，并通过公式&&的计算我们可以得到每一个话题所包含的绝对点赞数，由上表我们可以看到绝对点赞数超过 1000 以上的有 3 个话题分别为：A 市 A5 区汇金路五矿万境 K9 县存在一系列问题、A 市金毛湾配套入学的问题、请书记关注 A 市 A4 区 58 车贷案，留言的时间在 19 年离现在还是很近的时间点，但是绝对点赞超过 1000，间接说明这三个问题已经影响到老百姓的正常生活，而且还没有得到政府的关注，因此，应该得到工作人员的重视。超过 100 以上的绝对点赞数有 8 个话题：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、A 市富绿物业丽发新城强行断业主家水、关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉、请 A 市加快轨道交通建设力度、建议西地省尽快外迁京港澳高速城区段至远郊、问问 A 市经开区东六线以西泉塘昌和商业中心以南的有关规划、关于加快修建 A 市南横线的建议、关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉，以上 8 类话题在留言以后有一定的新增留言来保持

话题的热度但是相比于前三的热点问题，老百姓的关注还是较少。

本文在对没有进行分题随着人们不断关注的热度变化，工作人员应该要得到最有热度的话题，并用有限的政府资源解决，因此，绝对点赞数的提出对后面的话题热度评价提供了一个更准确的依据。

4.4.3 留言热度排序

对每一类的绝对点赞数进行求解以后，就需要对每一类的话题进行热度值的评价，评价结果如下

表 12 热度值的评价

	留言主题	留言热度	留言时间
1	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	990.46875	2019/8/19 11:34:04
2	反映 A 市金毛湾配套入学的问题	397.444444444	2019/4/11 21:02:44
3	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	315.0	2019/9/5 13:06:55
4	请书记关注 A 市 A4 区 58 车贷案	165.43828125	2019/2/21 18:45:14
5	A 市富绿物业丽发新城强行断业主家水	80.0625	2019/6/19 23:28:27
6	A 市三一大道全线快速化改造何时启动	40.833333333333	2019/9/15 15:31:19
7	请 A4 区教育局尽快落实发放原 A 市七中 01 年后退休教师的文明单位奖	39.75	2019/10/29 12:42:17
8	穿 A 市城而过的京港澳高速长楚高速什么时候可以外迁至远郊	36.0	2019/11/18 15:35:11
9	建议 A 市经开区收回东六路恒天九五工厂地块打造商业综合体	33.75	2019/11/8 15:48:07
10	请问 A7 县东六路下穿长永高速在 10 月底能否如期通车呢	27.416666666666664	2019/10/9 10:14:08

热度的评价值和分类以后的结果、绝对点赞数有关，依据公式 11 我们可以得到每一个话题的热度值，限于题目要求指出前五的热度话题，本文依据热度值将排名前 10 的话题用表 12 进行展示，

$$q=\sum_{i=0}^{n}(\frac{x}{n})(1-r)^{n-1} \tag{11}$$

由表格的结果可以得到排前五的话题为：A 市 A5 区汇金路五矿万境 K9 县存在一系列问题、反映 A 市金毛湾配套入学的问题、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、请书记关注 A 市 A4 区 58 车贷案、A 市富绿物业丽发新城强行断业主家水，其热度值分别为：990.46875、397.444444444、315.0、165.43828125、80.0625，从热度的结果可以看到 k9 县的问题是在所有的问题中居于榜首的问题，不仅在点赞数里面有 2097 个用户为其点赞，而且在表 5.4.1 里面可以发现在后面的分类中有 5 个类似的留言为此问题进行话题热度的刷新，而排在第四的车贷问题虽然在绝对点赞数的排名上处于很高的排名，但是在 5.4.1 的表格里可以看到，与他同一类的留言只有三条进行热度的延续，所以本文所提出的热度值评价体系是充分考虑了文本留言的相似基础上，得到绝对点赞数，最后观察其后续的变化情况进行综合的评价，因此本文提出的热度评价体系客观真实的反映出用户留言的

真实热度。

5.问题三模型求解及算法设计

5.1 问题分析

针对问题答复意见的评价，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，并尝试实现。

5.2 数学模型建立及求解

5.2.1 问题三数学模型建立

建立一个完整且合理的指标体系通常面临着以下三个问题：第一，指标体系具有很强的随意性和主观性，缺乏合理的论证过程；第二，指标体系量化中，有时存在数据搜集困难的问题，甚至有些末级指标数据收集根本不可行；第三，指标体系缺乏普适性。因此，本文构建的指标体系既包括了主观指标也包括客观性和时效性是由给出的表中数据计算得出，所以这三个指标为客观评价，可解释性需要人为主观的给出，为主观评价。

在四个指标已经确定，但每一个指标的权重是未知的，常见的方法为平均加权，这种方式不能体现某一个指标的相对重要性，因此，本文引用分析法进行具体的权重分析，它是指将一个复杂的多目标决策统，将目标分解为多个目标或准则，进而解为多指标（或准则、约束）的若干层次，通过定性指标模糊量化方法算出层次单排序（权数）和总排序，以作为目标（多指标）、多方案优化方案的顺序分解为不同的层次结构，然后用求解判断矩阵特征向量的办法，求得每一层次元素对上一层次某元素的优先权重，最加权和的方法递阶归并择方案对标的最终权重，此最终权重最大者即为案。因此考虑本文指标体系的特点，选择层次分析法进行的确定。

5.2.2 模型求解

模型的评价体系分为二级，包括了 4 个等级的评价结果分别为：好(>=4 分)，较好(>=3 分)，较差(>=2 分)，差(<2 分)。一级指标性评分，5 分制，由二级指标加权得来；每一个因素的具体权分析法算出，二级指标包关性、完整性、可解释性、时效性。

相关性：

相关性分析是指对或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。相关性素之间需要存在一定的联系或率才记为 5分。由于模型准确率为 0.85，则 0.15 是误差。对于两个句子 A B，A 和 B 进行对比，A

准确为 A1=0.85,不准确为 A2=0.15。B 准确为 B1=0.85,不准确为 B2=0.15。

表 13 两条语句的相关性

结果	概率
AB 都准确	A1B1（0.7225）
A 准 B 不准	A1B2（0.1275）
A 不准 B 准	A2B1（0.1275）
A 不准 B 不准	A2B2（0.0225）

现在将留言详情 A 和回复意见 B 两个句子进行比较分析，如果 A 和 B 匹配上了，则匹配结果里面实际包含 A和 A2B2 两种可能性，只统计 A1B1 的话，A1B1 的分数为5*0.7225/(0.7225+0.0225) = 4.84；A1B2 和 A2B1 的分数为 4.84*(0.1275*2)/0.7225 = 1.71 因此，在使用第一问所建立的卷络模型进行计算，当输出结果一致时，就将此条留言对应的相关系数值赋为 分，不一样，则为 1.71 分，

完整性：

回复的完整性是指在用户留言详情里面所涉及到的问题，留言回复具有针对性的回答越多，所体现的就越全面，越完整，因此本文利用查全率和查准率的计算方法：A 中字符在 B 中的概率为 P，B 中字符在 A 中的概率为 Q，则相似度计算公式为：2PQ/(P+Q)。以 0.5 的相似度视作两句子相似，完整。通过观察第二所提出的分类模人工干后的结果，发现该算法存在 20%的误差，因此，相似度>=0赋值为 4（=5*0.8）分，相似度>=0.4赋值为 3.2（=4*0.8）分，相似度>=0.3 赋值 2.4，相似度>=0.2 赋值 度<0.2 赋值 1 分。

可解释性：

广义上的可解释性指在我们需要了解或解决一件事情的时候，我们可以获得我们所需要的足够的可以理解的信息。比如我们在调试 bug 的时候，需要通过变量审查和日志信息定位到问题出在哪里。比如在科中面临一个新问题的研究时，我们需要查阅一些资料来了解这个新问题的基本概念和研究现状，以获得对研究方向的正确认识。反过来理解，如果在一些情境中我们无法得到的足够的信么这些事情对我们来说都是不可解释的此，可解释性指标为主观评价指标，本文将该指标由专家集进行主观评价后值。评语集为可解释性好（5 分），较好（4 分），较差（3 分），差（1 分）。

时效性：

信息的时效性是指从信息源发送信息后经过接收、加工、传递、利用的时间间隔及其效率。时间间隔越短，使用信息时，使用程度越高，时效性越强，本文所定义的时效性是指在某一个用户留完言以后，政府工作人员给与回复的时间差值，这个参数可以通过观察具体的留言情况，将这一段时间进行划分：一周以内：快（5 分），一周到一个月：较快（4 分），1 个月-2 个月慢（3 分），2-3 个月为：很慢（1 分）四个等级。

层次分析法：

1、 构造层次结构模型

将决策的目标、考虑的因素（决策准则）和决策对象按它们之间的相互关为最高层、中间层和最低层，绘出层次。最高层是指决策的目的、要解决的问题。最低层是指决策时的备选方案。中间指考虑的因决策则。对于相邻的两层为目标层，低层为因素层，本文建立次结构模型如下图 16 所示：

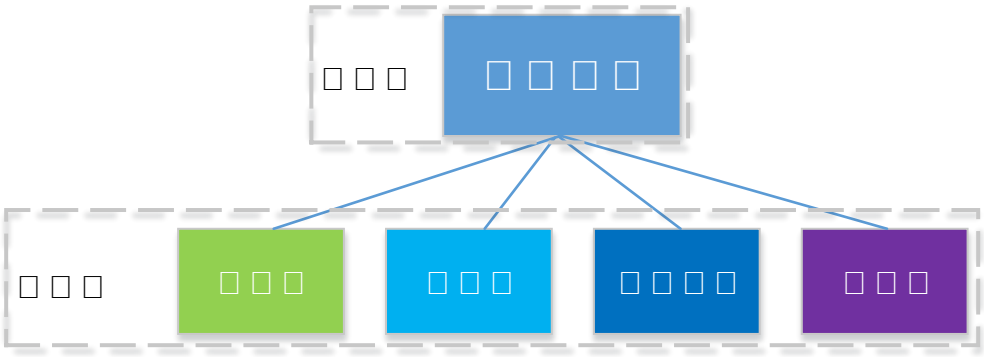


图 16 层次结构模型

2、 构造判断（成对比较）矩阵

确定每一层的各个因素的权重时，如果采取定性的结果，则常常不容易被别人接受，因而 Saaty 等人提出一致矩阵法所有因素放在一起比较，而是两两相互比较，对此时采用相对尺度，以尽可能减少性质不同的诸因素相互比较，以提高。如对某一准则，对其下的各方案进行两两对比，并按其重要性程度评定等级。a_{ij}为要素 i 与要素 j 重要性比较结果，表 6-2-2 列出的 9 个重要性等级及其。按两两比较结果构成的矩阵称作判断矩阵。判断矩阵具性质：

表 14 判断矩阵性质

因素 i 对因素 j	比较结果
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断中间值	2、4、6、8

在本文所讨论的因素中只有 4 个指标，因此，选取 1、3、5 三个等级，构造的结构矩阵如下表 15 所示：

表 15 构造的结构矩阵

	相关性	完整性	可解释性	时效性
相关性	1	3	3	5
完整性	1/3	1	1/2	2
可解释性	1/3	2	1	1
时效性	1/5	1/2	1	1

3、 层次单排序及其一致性检验

对应于判断矩阵最大特征根λ_{max}的特征向量，经归一化(使向量中各元素之和等于 1)后记为 W。W 的元一层次因素对于上一层次因素某因素相对重要性的排序权值，这一过程称为层次单排序。能否确认层次单排序，则需要进行一致性检验，所谓一致性检验是指对 A 确定不一致范围。其中，n 阶一致阵的唯一非零特征根为 n；n 阶正互反阵 A 的最大特征根 ， 当且仅当 时，A 为一致矩阵。

由于λ连续的依赖于a_{ij}，则λ比 n 大的越多，A 的不一致性越严重，一致性指标用 CI 计算，CI 越小，说明一致性越大。用最大特征值对应的特征向量作为被比较因素对上层某因素影响程度的权向量，其不一致程度越大，引起的判断大。因而可以用的大小 λ-n 数值来衡量 A 的不一致。定义一致性指标为：

$$CI = \frac{\lambda - N}{N - 1} \tag{12}$$

CI=0，有完全的一致性；CI 接近于 0，有满意的一致性；CI 越大，不一致越严重。

为衡量 CI 的大小，引入随机一致性指标 RI

$$RI = \frac{CI_1 + CI_2 \cdots CI_N}{N} \tag{13}$$

其中，随机一致性指标 RI 和判断矩阵的阶数有关，一般情况下，矩阵阶数越大，则出现一致性随机偏离的可能性也越大。

考虑到一致性的偏离可能是由于成的，因此在检验判断矩阵是否具有满意的一致性时，还需将 CI 和随机一致性指标 RI 进行比较，得出检验系数 CR，公式如下：

$$CR = \frac{CI}{RI} \tag{14}$$

一般，如果 CR<0.1 ，则认为该判断矩阵通过一致性检验，否则就不具有满意一致性

4、 层次总排及一次性检验

计算某一层次所有因素对于最)相对重要性的权值，称为层次总排序。这一过程是从最高层次到最低层次依次进行的

5.3 算法设计

根据 6.2 中的模型建立，首先使用分析法，确定各指标所占权重如下表 16：

表 16 指标所占权重

二级指标	相关性	完整性	可解释性	时效性
权重	0.5247	0.1612	0.1944	0.1197

接下来对每个二级指标进行计算，加权平均后即可得到评论的评价得分，以及对应的评语。

5.4 数据分析及结论

表 17 每一个用户的留言详情表

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	的物业公司却以交20万保证金，不能	车管理费，在业主大会结束后	2019/5/10 14:56:53
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好？	2019/4/24 16:03:40	和店面的生意带来很大影响，里面体	换墙，且换墙后还有三趟雨水	2019/5/9 9:49:10
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	园的同时更是加大了教师的工作压力	园聘任教职工要依法签订劳动	2019/5/9 9:49:14
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗？	2019/4/24 15:07:30	政落户A市，想买套公寓，请问购买	岁以下（含），首次购房后	2019/5/9 9:49:42
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	为“马坡岭小学”，原“马坡岭小学	坡岭”的问题。公交站点的设	2019/5/9 9:51:30
2759	A00077538	A3区含浦镇马路卫生很差	2019/4/8 8:37	后再把泥巴冲到右边，越是上下班上	中没有说明卫生较差的具体路	2019/5/9 10:02:08
2849	A000100804	A3区教师村小区盼望早日安装电梯	2019/3/29 11:53:23	出台为老社区惠民装电梯的规范性文	人民政府办公室下发了《关于	2019/5/9 10:18:58
3681	UU00812	反映A5区东澜湾社区居民的集体民生诉求	2018/12/31 22:21:59	要跑好远，天寒地冻的跑好远，眼看	准备及设施设备采购等工作。	2019/1/29 10:53:00
3683	UU008792	映A市美麓阳光住宅楼无故停工以及质量问	2018/12/31 9:55:00	。也没得到相关准确开工信息。同时	实分户检查后，西地省楚江新	2019/1/16 15:29:43
3684	UU008687	洋湖新城和顺路洋湖壹号小区路段公共绿化	2018/12/31 9:45:59	在立交桥等地方做立体绿化，取得	按规划要求完成了建设，其中	2019/1/16 15:31:05
3685	UU0082204	反映A2区大托街道大托新村违建问题	2018/12/30 22:30:30	城乡规划局审批通过《温室养殖大棚	一笔耕地征收补偿款给原大托	2019/3/11 16:06:33
3692	UU008829	A5区鄱阳村D区安置房人防工程的咨询	2018/12/29 23:27:51	村D区安置房地下室近两万平方米，	人防发[2014]7号文件要求，	2019/1/29 10:52:01
3700	UU00877	国城小区段请求修建一座人行天桥或者地	2018/12/29 11:55:34	高峰，大量从小区开车出去的业主经	合进行具体选址，招标（邀标	2019/1/14 14:34:58
3704	UU0081480	举报A市芒果金融平台涉嫌诈骗	2018/12/28 17:18:45	在贵省相关政府部门的大力支持下	关警情，已由银盆岭派出所立	2019/1/3 14:03:07
3713	UU0081227	建议增开A市261路公交车	2018/12/28 7:53:25	半小时以上！天寒地冻，其他公交车	由于驾驶员工作时间长，劳动	2019/1/14 14:33:17
3720	UU008444	新开铺路与披塘路交叉路口通行安全问题	2018/12/27 15:18:07	址地址：https://baidu.com/。但	披塘路路口西端各拆除20米中	2019/3/6 10:26:14
3727	UU0081194	投诉A3区桐梓坡路益丰大药房以次充好	2018/12/27 1:55:21	他们便以各种理由拒绝退货，并将可	提供的信息进行投诉信息的登	2019/1/3 14:02:47
3733	UU008706	建议在A市梅溪湖开办一个图书馆	2018/12/26 16:51:40	不相称。建议在艺术中心先期借一个	梅溪湖二期金菊路与雷松路对	2019/1/14 14:32:40
3747	UU008201	部门治理A3区中海国际社区一期旁边工地的	2018/12/25 19:35:12	早上很早就施工，严重影响居民的	工单位由于需要夜间连续作业	2019/1/8 16:19:16
3755	UU0081681	望A市社保卡、医保卡、居民健康卡尽快合	2018/12/25 16:23:27	更，希望可以尽快合一。让社保卡尽	同机构，需三万或三万以上不	2019/1/4 15:48:23
3756	UU0081681	望A市潇楚一卡通尽快支持手机nfc虚拟公交	2018/12/25 16:19:49	卡、华为、苹果等手机都无法开通n	线时间请关注潇洒支付公司官	2019/1/4 15:49:46
3760	UU0081500	市北盛镇对泉水村塘下组土地征收存在的违	2018/12/25 14:40:13	了向有权国家行政机关进行了申请用	签订了土地补偿协议，并按协	2019/1/8 16:18:00
3762	UU0081057	于A5区交警大队纠正电子交通警察的错误执	2018/12/25 13:56:31	妨碍直行车辆和行人通行，此路口也	例》第三十八条第一款第三项	2019/1/16 15:22:16

表 18.模拟评价

留言时间	留言详情	答复意见	答复时间
2019/4/25 9:32:09	业公司却以交20万保证金，不能	意收取停车管理费，在业主大会结束后业委会也	2019/5/10 14:56:53
2019/4/24 16:03:40	面的生意带来很大影响，里面的	是，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	时更是加大了教师的工作压力，	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	户A市，想买套公寓，请问购买	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“马坡岭小学	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 8:37	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，也没	2019/5/9 10:02:08

表 19 模拟评价的平均评价得分

留言编号	相关性	完整性	可解释性	时效性	评价得分	评语
2549	4.84	4	5	4	4.6351	好
2554	4.84	3.2	4	4	4.3601	好
2557	4.84	4	5	4	4.6835	好
2574	4.84	4	3	4	4.2947	好
2759	1.71	1	4	3	2.2122	较差

6.模型算法评价

1：本文在正确、清楚地分析了题意的基础上，建立了真实客观的深度学习模型，能够实现文本的精确分类和相同话题的聚类，得到每一个话题的具体热度排名。
2：本文针对第一二三问的算法，特别是第一问算法，我们对多种模型的组合进行实验验证，文章中提出的模型都是所有测试效果中最佳的测试结果构建的。
3：建立的分类与聚类模型能与实际紧密联系，结合实际情况对问题进行求解，使得模型具有很好的通用性和推广性；
4：模型的计算采用专业的数学软件，可信度较高
5：对模型中涉及到的众多影响因素进行了量化分析，使得论文有说服力

7.谢辞

春风化雨，润物无声。受到疫情影响，我和队友们没法在一起工作，但这一切的外界因素并没有影响到我们彼此之间的信任和配合，能够圆满完成比赛项目。

首先，我们要感谢的是出题人，得益于这次项目，我们自身的学术能力，应用水平得以提高。并且从题目中可以看到，出题人以及我们的政府为了更加高效的为人民服务，想民之所想，急民之所急，很让我们感动，也坚定了我们做好这道题的决心。

其次，我们要感谢我们的指导老师，在无数个日夜关注我们的进程，指导我们的项目，仔细阅读我们的每一句话，每一个词，然后提出最好的建议。老师辛苦了！

最后，我们要感谢我们自己，经过精诚合作，夜以继日的努力，我们构建了庞大的模型，用来支撑我们的项目，各方面能力都有所提高。

8.参考文献

[1]周伟枭,蓝雯飞.融合文本分类的多任务学习摘要模型[J/OL].计算机工程:1-10[2020-05-08].https://doi.org/10.19678/j.issn.1000-3428.0057448.

[2]彭俊利,谷雨,张震,耿小航.融合改进型 TC 与 word2vec 的文档表示方法[J/OL].计算机工程:1-7[2020-05-08].https://doi.org/10.19678/j.issn.1000-3428.0056370.
[3]张衡,马明栋,王得玉.基于聚类网络的文本-视频特征学习[J/OL].计算机科学:1-10[2020-05-08].http://kns.cnki.net/kcms/detail/50.1075.TP.20200330.1326.017.html.
[4]杨锐,陈伟,何涛,张敏,李蕊伶,岳芳.融合主题信息的卷积神经网络文本分类方法研究[J].现代情报,2020,40(04):42-49.
[5]杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术与信息化,2020(03):146-148.
[6]赵容梅,熊熙,琚生根,李中志,谢川.基于混合神经网络的中文隐式情感分析[J].四川大学学报(自然科学版),2020,57(02):264-270.
[7]姚佳奇,徐正国,燕继坤,熊钢,李智翔.基于标签语义相似的动态多标签文本分类算法[J/OL].计算机工程与应用.
[8]艾楚涵，姜迪，吴建德。关键词：基于主题模型和文本相似度计算的专利推荐研究 信息技术，2020,44（04）：65-70。
[9]王新新. 面向模式的文本数据描述模型[J]. 科技创新与应用,2020(10):28-30.
[10]殷硕，王卫亚，柳有权。关键词：基于语义特征抽取的文本聚类研究 计算机技术与发展，2020,30（03）：46-50.
[11]Yongchang Wang,Ligu Zhu. Research on improved text classification method based on combined weighted model[J]. Concurrency and Computation: Practice and Experience,2020,32(6).
[12]Park,Hong,Kim. A Methodology Combining Cosine Similarity with Classifier for Text Classification[J]. Applied Artificial Intelligence,2020,34(5).
[13]Mehdi Emadi,Maseud Rahgozar. Twitter sentiment analysis using fuzzy integral classifier fusion[J]. Journal of Information Science,2020,46(2).