第八届“泰迪杯”数据挖掘挑战赛 C 题论文

题目 ：“智慧政务”中的文本挖掘应用

关键词：文本分类 循环卷积神经网络 核心词簇 AHP—熵权综合评价模型 灰色模糊评价

摘要

随着我国互联网事业的不断发展，中国的网民数量已跃居世界第一位，为了更好的适应时代的发展，我国政府等机构也推出了微博、微信、市长信箱等网络问政平台来了解民意，汇聚民声，带来便利的同时也出现了许多挑战。文本数据的数量不断扩大，不同类别、不同热点的分类显得尤为重要。本文针对不同类型的留言文本,聚焦留言中的各种热点问题,对不同文本内容进行文本挖掘，提取语义关键特征，从而实现对文本的精准分类，针对文本在文本分类时提取语义关键特征难度大,分类效果差等问题,建立基于循环神经网络变体和卷积神经网络（BGRU-CNN)的混合模型,实现留言文本的准确分类。首先,通过循环层神经网络捕捉文本信息,并将其作为神经网络的输入；然后,建立 BGRU-CNN 模型,经双向门控循环单元(B-GRU)实现文本的序列信息表示,利用卷积神经网络(CNN)提取文本的关键特征,通过 Softmax 分类器实现文本的准确分类。本文针对三个问题分别建立 3 个模型，并结合附件数据给出相应的分类。

针对问题 1：建立分类模型，并使用 F-score 对分类模型进行评价。

针对问题 2：对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。

针对问题 3：建立 AHP—熵权综合评价模型和构造灰色模糊评价。

关键词:文本分类 循环卷积神经网络 核心词簇 AHP—熵权综合评价模型 灰色模糊评价

Abstract.

With the continuous development of China's Internet industry, the number of Internet users in China has become the largest in the world. In order to better adapt to the development of The Times, China's government and other institutions have launched micro-blog, WeChat, mayor's mailbox and other network political platform to understand public opinion, gather people's voices, bring convenience but also many challenges. The amount of text data is constantly expanding, and the classification of different categories and hot spots is particularly important. In view of the different types of message text, this paper focus on the message in various hot issues, for different text mining text content, extract the semantic key characteristics, so as to realize the accurate classification of the text, for the text in a text classification key features is difficult to extract semantic, poor classification results, neural network based on cycle variation and convolutional neural network (BGRU - CNN) hybrid model, realize accurate classification of text messages. Firstly, text information is captured through a loop layer neural network and used as an input of the neural network. Then, the BGRU-CNN model was established, and the sequence information representation of text was realized by the bidirectional gated loop unit (b-gru), and the key features of text were extracted by the convolutional neural network (CNN), and the text was accurately classified by the Softmax classifier. In this paper, three models are established for the three problems, and the corresponding classification is given based on the attached data.
For question 1: establish a classification model and use f-score to evaluate the classification model.
For question 2: categorize the comments reflecting the problems of specific places or specific people in a certain period of time, define a reasonable heat evaluation index, and give the evaluation results.
Aiming at problem 3: establishing AHP - entropy weight comprehensive evaluation model and constructing grey fuzzy evaluation.

Keywords:Text Classification circular convolution Neural Network Core Word Cluster AHP---Entropy Weight Comprehensive Evaluation Model Grey Fuzzy Evaluation

一、问题重述

1.1 问题背景

新时代的中国互联网事业迅猛发展，网络已成为百姓日常生活中必不可少的一部份，习近平总书记也强调“通过网络走群众路线”，让互联网成为同群众沟通交流的新平台，成为了解群众，贴近群众，为群众排忧解难的新途径，成为发扬人民民主，接受人民监督的新渠道。微博、微信、政府平台留言等网络问政平台便成为人民群众抒心声，表民意的重要途径，如何处理庞大的留言数据也成为了有关部门的一大挑战。

1.2 目标任务

1.2.1 利用附件 2 所给数据，提取并分析留言内容中与分类标签相关的特征语义，并建立一级标签分类模型，使用 F-score 对分类方法进行评价。

1.2.2 利用附件 3 所给数据，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.

1.2.3 综合考虑相关部门的答复意见，并从相关性，可解释性等角度，为相关部门给出留言答复的综合评价指标体系与综合评价模型。

二、问题分析

题目要求利用所给留言内容对文本进行分类挖掘，对于问题一而言可根据文本的特征性，语言特征含义，通过卷积循环对留言进行一级标签分类，并以此分析出问题二中的热点问题，建立核心词簇模型，给出热度评价结果。在留言答复的过程中不确定性因素对答复意见有一定影响，在问题三中，人的主观情感意识，事件热度等都是影响因素之一，对于此问而言我们建立出 AHP—熵权综合评价模型，并通过灰色模糊综合评价方法得出综合评价模型，达到求解问题的结果。

三、问题假设

1.假设题中所给附件中的数据真实、可靠、具有普遍性。
2.假设除题中所给干扰因素外其它因素不对评价结果造成影响。

四、符号说明

x_i	文本特征向量
P_i	评估指标查准率
R_i	查全率
$\vec{d}_{title}$	报道的标题向量

$\vec{d}_{text}$	报道的正文向量
tf_i	特征t_i在当前话题T_k下所有报道中出现的频率
N	当前话题中所有的报道数
n_i	当前话题中含有特征t_i的报道数
θ_1	一次聚类阈值
θ_2	二次聚类阈值
β	偏好系数

五、模型建立与求解

5.1 问题一 的分析与模型建立：

5.1.1 对问题一的分析：

针对附件 3 的相关留言内容，可根据留言主题短文本确定词向量，通过双向循环结构输出的文本特征词向量进入卷积层从而确定句向量从而确定留言一级分类标签，对模型的评价则使用 F-score 评价方法对文本一级分类的查全率和查准率进行评价。

5.1.2 对问题一的模型建立及求解：

5.1.2.1 建立 BGRU-CNN 模型

BGRU-CNN 混合模型的整体结构如图 1 所示。BGRU-CNN 混合模型的输入为文本 T,经过循环层捕捉句子上下文信息,卷积层提取文本的关键特征,输出为文本 T 属于类别 h 的概率。

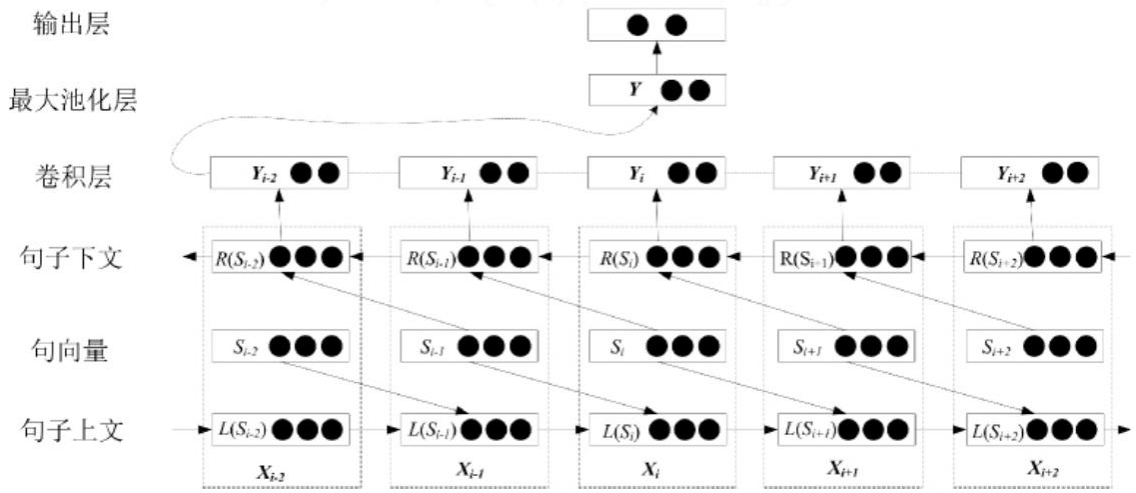


图 1　BGRU-CNN 混合模型

（1）循环层

采用双向循环结构捕获第i个句子S_i;上文$L(s_i)$和句子下文$R(s_i)$的信息,计算方式为

$$L(s_i)=f(W_1,L(s_{i-1})+W_{s1},e(s_{i-1})) \tag{1}$$

$$R(s_i)=f(W_r),R(s_{i+1})+W_{sr},e(s_{i+1})) \tag{2}$$

式(4)中,$e(s_{i-1})$ 表示句子S_{i-1}的句向量$L(S_{i-1})$表示句子S_{i-1}的上文信息,W_{s1}表示S_i句子和S_{i-1}句子语义信息组合的权阵,W_1为隐藏层的转换权阵,f为激活函数。然后,通过式(6)将上文信息$L(S_i)$、句向量$e(S_i)$和下文信息$R(S_i)$拼接,构成文本特征向量x_i作为循环层的输出。

$$x_i=\left[L(S_i),e(S_i),R(S_i)\right] \tag{3}$$

双向循环结构输出的文本特征向量x_i;学习了文本上下文的语序信息,作为卷积层的输人,进一步提取文本语义特征。

（2）卷积层

在循环层获取文本特征向量x_i后,使用CNN网络进行特征y_i的提取,计算方式为

$$y_i=f\left(w\bullet x_{i:h}+b\right) \tag{4}$$

其中,卷积核用$w\in R^{hk}$表示,h和k分别表示卷积核的窗口高度和宽度,用来对循环层的输出x_i进行卷积.$x_{i:h}$表示输入特征向量第i行到第h行的特征值。b为偏置,f为激活函数。获取所有y_i后,通过式(5),构建关键特征图Y。

$$Y=\left[y_1,y_2,y_3,\cdots,y_n\right] \tag{5}$$

然后使用最大池化层来确定文本的最佳特征,计算方式如式(6)所示,

$$y=\max\left(y_i\right),i=1,2,\cdots,n \tag{6}$$

得到最佳特征y,将这些特征输入分类层分类。分类层采用dropout方式将最佳特征y连接到Softmax分类器中。dropout算法随机将最佳特征y按一定比例置0,其他没有置0的元素参与运算。由于每一次输入特征向量后,置0的方式都是随机的,因此网络权重参数每一次都得到了更新。直到所有样本都被训练完成。因为每次网络权重参数都不相同,dropout算法将神经网络变成了多种模型组合,有效地防止了过拟合,提升了模型预测的精度。W_c和b_c分别表示Softmax

分类器的权重参数和偏置项,通过 dropout 产生的向量记为c_d,则其输出O的计算方式为

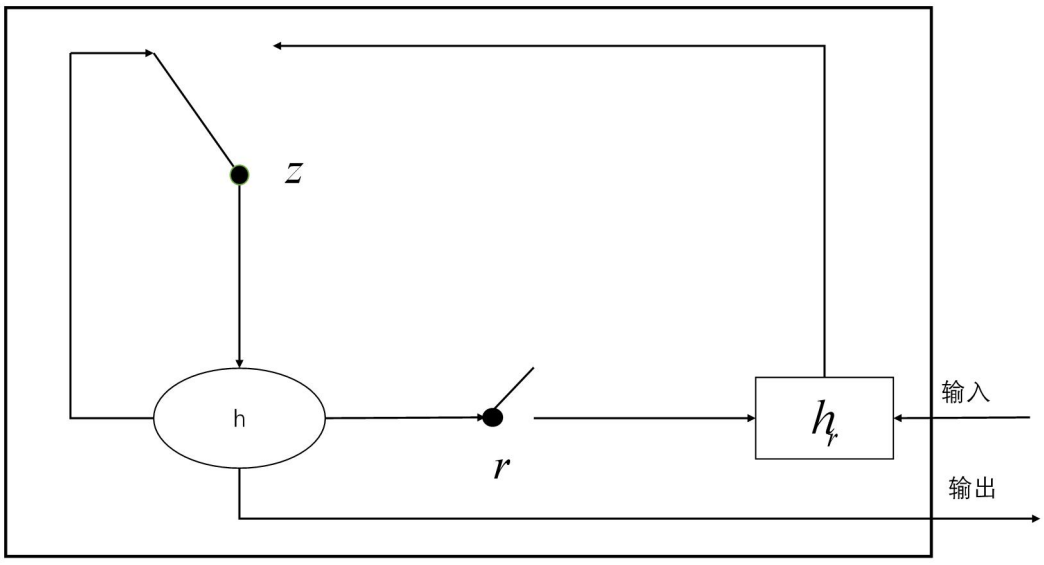
$$O = W_c c_d + b_c \tag{7}$$

最后,预测文本属于第k类的概率的计算方式如式(8)所示。其中,O_k表示输出向量O中的第k个元素,N表示类别数。

$$p(k|T) = \frac{e^{O_k}}{\sum_{j=1}^{N} e^{O_j}} \tag{8}$$

5.1.2.2 BGRU-CNN 模型循环层计算节点

本文对长文本进行分类，在一篇长文本中，往往包含十几甚至几十个句子，以当前句子为中心，中距离或者长距离句子的语序与当前句子所要表达的语意相关。而 RNN 网络采用的计算节点无法很好地处理这种依赖，未能完全识别长文本子语序特征，因此，将 RNN 网络的计算节点替换为 GRU 计算节点，使模型学习到更完整的长文本句子语序特征。图 5 为所采用的 GRU 结构，GRU 网络节点通过重置门r和更新门z对输入信息进行处理。t时刻的激活状态h_t,计算方式如式(9)



图二 GRU 计算节点

$$h_t = (1 - z_t) \otimes h_c + z_i \otimes h_{t-1} \tag{9}$$

h_{t-1}是$t-1$时刻的激活状态,与h_t呈线性关系，⊗表示向量矩阵对应元素相乘,z_i表示t时刻更新门的状态,计算方式为式(10)。t时刻激活状态h_c的计算方式如式(11)，重置门r_t的计算方式如式(12)。

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right), \tag{10}$$

$$h_c = \tanh\left(W_x x_t + U\left(r_t \otimes h_{t-1}\right)\right), \tag{11}$$

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right)。 \tag{12}$$

σ为sigmoid函数,x_i是t时刻该节点输人的句向量,W_z, W_x, W_r和U_z, U, U_r是更新门z、当前候选的激活状态h_c和重置门r要训练的权重参数。更新门z和重置门r都使用了σ函数,将两个门的输出都限定在[0,1]之间。h_c使用r来控制包含过去时刻信息的上一个激活状态的流入，如果r近似于0,那么上一个激活状态将被丢弃，因此,r决定了过去有多少信息被遗忘。h_t使用z来对上一个激活状态和候选状态进行更新，如果z近似于0,那么h_t不会被更新,z可以控制过去激活状态在当前t时刻的重要性。通过GRU计算节点更新门和重置门的设置,改善RNN在识别长文本语序时,无法很好地处理中远距离句子依赖的问题,识别到更加完整的上下文语序特征。

5.1.2.3 测验BGRU-CNN模型

将所有BGRU-CNN模型的参数定义为集合θ。循环层所有参数定义为集合W，包括初始，上下文信息$L(S_1)$和$R(S_n)$、权重参数W_{S1}, W_1, W_{sr}, W_r。卷积层所有参数定义为集合U，包括$W_z, W_x, W_r, U_z, U, U_r$。则$\theta$包括句向量$S, W, U$,偏置项$b_c, W_c, w$,如式(13)所示。

$$\theta = \{S, W, U, b_c, W_c, w\}, \tag{13}$$

经过神经网络训练，找到最小代价的θ,如式(14) 所示，

$$\theta \leftarrow \sum_{T \in D} \log p\left(class_T | T, \theta\right) \tag{14}$$

其中,D为训练的文档集,$p\left(class_T | T, \theta\right)$ 表示文档T在参数θ下属于目标类别$class_T$的概率。在实验中采用随机梯度下降方法训练BGRU-CNN模型,则θ的更新

式为

$$\theta \leftarrow \theta + \alpha \frac{\partial \log p\left(class_D \middle| D, \theta\right)}{\partial \theta} \tag{15}$$

其中,α 为学习率。如果α 太小，导致网络收敛速度太慢,太大会导致无法收敛。经过多次实验验证,α 设置为 0.01 时效果最好。

本测验采用 MATLAB 对 BGRU-CNN 模型进行检验，BGRU 循环网络实验参数设置的隐层节点数设置为 600,学习率为 0.01,BGRU 的隐层状态作为 CNN 卷积层的输入;CNN 卷积网络模块的卷积核窗口大小设置为 3x150,4x150,5x150;选择 relu 作为激活函数。池化层通过采样,选择最佳特征;Softmax 层采用 L2 正则项防止过拟合，参数为 0.2;dropout 参数设置为 0.5。具体实验参数设置如表 2 所示。

采用 F-score 对文本分类进行评价，评估指查准率(P_i)、查全率(R_i)和 F1 值对 BGRU-CNN 模型的有效性进行验证。令 TP,FN, FP 和 TN 分别代表正阳性、假阴性、假阳性和正阴性的分类数量,则评估指标的表示如式(16)所示，

$$P = \frac{TP}{TP + FP},$$
$$R = \frac{TP}{TP + FN},$$
$$F = \frac{2PR}{P + R}. \tag{16}$$

5.2 问题二的模型建立与求解：

5.2.1 对问题二的分析：

在不同类别的网络留言中,对于热点留言的特征是多方面的,从时间上来看,在一定时间范围内反映相同的事件；从信息源来看不同网民所反映的相同事件;从网民角度来分析不同网民对事件的关注度点赞数等都是热点留言的相关特征,本文主要以附件中的相关留言材料为研究对象，综合考虑热点留言中的四个特征，通过对留言进行去重、排序、合并以及热度计算等调整，实现对留言热度的有效检测。

5.2.2 问题二模型的建立及求解：

5.2.2.1 热点留言发现模型

首先采用切词系统 ICTCLAS 对采集的留言语句进行切词,并采用一个人工整理的停用词表对切词结果进行过滤，去掉助词、连词、介词等虚词，以及词语长度为 1 的无实际含义的字和无意义的字符串。对于每一篇报道d，本文将其标题和正文分别进行文本表示，假设$\vec{d}_{title}$代表报道的标题向量，$\vec{d}_{text}$代表报道的正文向量，则报道d经过文本预处理后可以表示为:$\vec{d} = \left\{\vec{d}_{title}, \vec{d}_{text}\right\}$。对于每一个话题，本文采用中心向量来构建话题模型，选取附件的第一条留言作为初始话题中心，后续每加入一篇留言，话题模型动态更新，话题T的中心向量可表示

为: $T=(t_1,w_1,t_2,w_2,\cdots,t_i,w_i,\cdots t_m,w_m)$，$w_m$ 为话题 T 中特征项 t_m 的权重，计算公式如下:

$$w_i=\frac{tf_i\times\log\left(\frac{N}{n_i}+0.01\right)}{\sqrt{\sum_{t_m\in T_k}tf_m^2\times\log^2\left(\frac{N}{n_m}+0.01\right)}} \quad (17)$$

其中，tf_i 为特征 t_i 在当前话题 T_k 下所有报道中出现的频率，N 为当前话题中所有的报道数，n_i 为当前话题中含有特征 t_i 的报道数。当话题内有新的报道加入时，则重新计算话题模型内每个特征的权重，从而动态的调整话题模型。此外，为了降低空间维数,提高算法的运算速率,对于话题模型中权重小于 0.001 的特征项,考虑到其对话题的贡献度很小，本文均将其去掉。

5.2.2.2 建立核心词簇

核心词簇，是指一组具有特定文本表示能力的词组集合，由话题的若干个热点主题词组成。经过问题一的标签分类分析得出，在分类标签下一般 2 个热点主题词就可以描述一个话题，如：“占道施工”，“安全隐患”等，当热点主题词个数为 3-8 个时，最能描述话题的特性。因此，对于每一个留言内容，本文提取 5 个热点主题词作为该话题的核心词簇，其中留言主题中 3 个，留言详情中 2 个，共包括 2 个命名实体，1 个名词和 2 个动词。

如何准确的提取出话题的热点主题词是热点话题识别的关键,本文采用基于统计信息的主题词提取方法，并考虑到词性和词的位置因素:在词性上，命名实体的重要程度要高于动词和名词;在词的位置上，位置越靠前的词重要程度越高，而且在标题中出现的词语的重要程度要比在正文中的高的多。综合上述考虑，本文给出了热点主题词的提取方法，具体公式如下:

$$Imp(t_i)=\alpha\times w(t_i)\times pos(t_i)+\beta\times loc(t_i) \quad (18)$$

其中，$Imp(t_i)$ 为特征项 t_i 的重要性分值 α、β 为调整因子，令 α=0.8，β=0.2

$w(t_i)$ 为特征项的权重，由公式(18)可得 $pos(t_i)$ 和 $loc(t_i)$ 分别表示特征项的词性和位置，具体定义如下:

$$pos(t_i)=\begin{cases}2 & \text{命名实体}\\ 1.5 & \text{动词}\\ 1 & \text{其它}\end{cases}$$

$$loc(t_i)=\begin{cases}4 & \text{留言主题中}\\ 1 & \text{留言详情中}\end{cases}$$

通过公式(18)计算可得到特征项的重要性分值，按照重要程度从高到低进行排序，提取出符合条件的热点主题词，得到话题的核心词簇。

5.2.2 相似度计算

5.2.2.1 词簇之间相似度

由于核心词簇本身的文本维数很低，采用夹角余弦公式计算词簇之间的相似度时，会导致词簇之间不同特征词的权重值很大而相同特征词的权重值很小，影响聚类的效果。因此，对于低维的词组集合，本文采用统计的方法来计算其相似度，具体公式如下:

$$Sim\left(C_i,C_j\right)=\frac{C_i\left(t\right)\cap C_j\left(t\right)}{C_i\left(t\right)\cup C_j\left(t\right)} \tag{19}$$

其中C_i,C_j表示两个不同的词簇，它们之间的相似度值为两者共同含有的特征词语个数与两者特征词语个数总和的比值。

5.2.2.2 留言与话题之间相似度

报道和话题之间的相似度计算方法采用经典的余弦夹角公式，具体计算公式如下:

$$Sim\left(d,T\right)=\frac{\sum_{i\in N}d_iT_i}{\sqrt{\left(\sum_{i\in N}d_i^2\right)\left(\sum_{i\in N}T_i^2\right)}} \tag{20}$$

其中，公式（20）中 的相似度值为介于 0 至 1 之间的小数，值越接近 1，则报道d与话题T的相似度越高，说明报道d属于话题T的可能性就越大，反之，值越接近 0,报道d属于话题T的可能性就越小。

5.2.2.3 留言热度评估

话题热度是衡量网络热点话题的主要指标，对于快速了解网络风向标，政府部门及时了解民声，具有十分重要的意义。影响话题热度的因子主要有:相同话题的留言数量、相同话题留言间隔时间、留言点赞量、用户评论数等。对于一个话题来说，如果该话题在某一段时间内被重复的频率比较高，则说明网民对于该话题的关注度比较高;如果该话题在某--段时间内参与讨论的人数比较多，则说明网民对该话题的关注度比较高。一个热点事件，广大网民就是政府部门的眼睛，对于基层建设而言，政府关注百姓心中的相关热点问题具有十分重要的意义。因此，本文根据热点话题的特点，综合考虑政府和网民这两个方面，以天为时间单位，给出了话题热度的计算方法，具体公式如下:

$$H_T\left(t\right)=\sum_i\left(\alpha\frac{rn}{rN}+\beta\log\ pn\right) \tag{21}$$

其中，rN为单位时间内所有的相同话题留言数，rn为只与话题T相关的留言数，pn为话题T的参与人数，α,β为本文设置的调整因子$\alpha=0.8,\beta=0.1$。

现已有的话题热度评估方法仅仅考虑了过去某一段时间内话题的关注度，忽略了对话题未来发展趋势的预测，不利于网络舆情的监测及有效控制。为此，本文引入了“话题热度指数”，具体定义如下:

$$E_x = \frac{H_T(t_x)}{H_T(t_1)} \times 100\% \qquad (22)$$

其中$H_T(t_x)$为话题T在第x天的热度，$H_T(t_1)$为话题T在第 1 天的热度。通过话题热度指数，可以对话题的产生、发展、高潮、衰落等各个阶段进行实时的跟踪和预测，并可以得到一条话题的发展变化曲线，对于分析和预测网络热点话题具有十分重要的意义。

5.2.2.4 核心词簇的热点发现算法描述

该算法按报道的输入先后顺序依次处理数据流中的报道，直到所有的报道处理完毕，具体过程如下：

输入：新闻报道、微聚类阈值θ_1和二次聚类阈值θ_2。

输出：热点话题簇

算法过程：

（1） 将第一篇报道作为初始话题中心，建立话题模型；

（2） 抽取话题模型的热点主题词，构建话题的核心词簇；

（3） 对每个后续到来的报道，利用公式(19)计算报道的标题向量与已有话题的核心词簇的相似度；

i.若相似度值超过预设的聚类阈值 ，则将该报道加入到最相似话题的候选集中，并更新话题的核心词簇，转到 （4）；

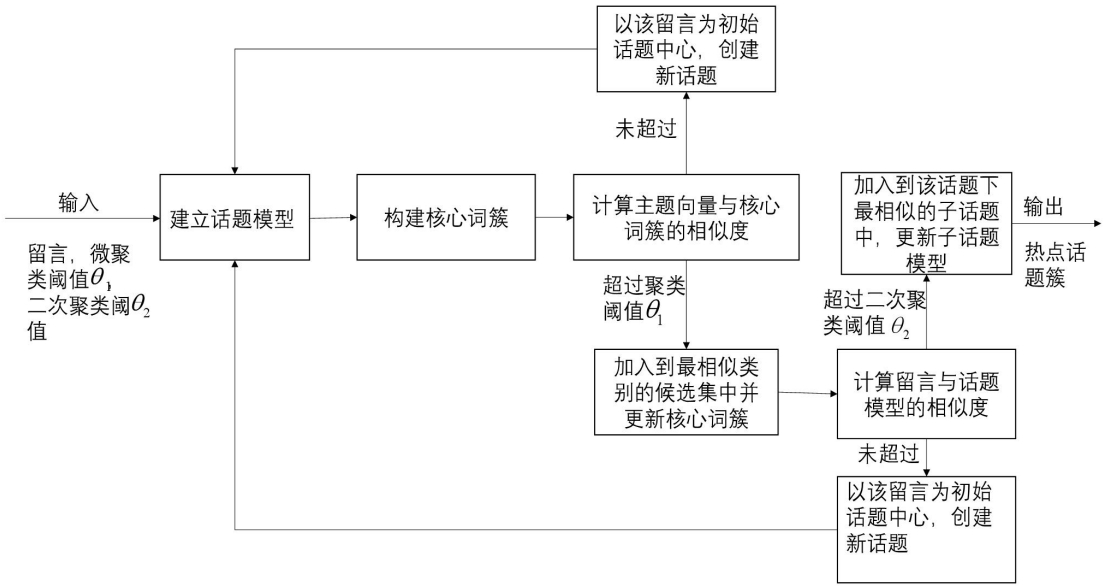
ii.否则，以该报道为初始话题中心，创建新话题；

（4）利用公式(20)计算报道与话题模型的相似程度；

i. 若超过二次聚类阈值，则将其加入到该话题下最相似的子话题中，并更新该子话题模型；

ii.否则，以该报道为初始子话题中心，创建新的子话题；

（5） 重复上述过程，直到所有的文本处理完毕。



图三 核心词簇的热度发现步骤

本实验采用 TDT 评测标准中的漏检率、误检率以及检测代价来评价话题检测性能。假设对于话题i，系统已检测到与话题i相关的报道数为a，未检测到的与话题i相关的报道数为c，已检测到与话题i不相关的报道数为b，未检测到与话题i不相关的报道数为d，则话题i的漏检率P_{miss}、误检率P_{fa}以及系统检测代价$(C_{Det})_{Norm}$定义如下:

漏检率: $P_{miss}=\frac{c}{a+c}$，

误检率：$P_{fa}=\frac{b}{b+d}$

归一化检测开销：$(C_{Det})_{Norm}=\frac{C_{miss}\times P_{miss}\times P_{target}+C_{Fa}\times P_{Fa}\times P_{non-target}}{\min\left(C_{Miss}\times P_{target},C_{Fa}\times P_{non-target}\right)}$，

其中，C_{miss}为漏报一个话题的代价，C_{Fa}为误报一次的代价，P_{target}和$P_{non-target}$为目标话题的先验概率$P_{non-target}=1-P_{target}$，它们的值需要预先设定，本文认为漏检的代价要比误报的高的多，此处取$C_{Miss}=1.0,C_{Fa}=0.1$，P_{target}采用 TDT2004 中设定的值 0.02,则$P_{non-target}$ =0.98,$(C_{Det})_{Norm}$的值越小表示系统的识别性能越好。

5.2.2.5 实验结果与分析

5.2.2.5.1 确定最佳聚类阈值

本组实验的目的是验证系统的最佳识别性能,通过多次实验寻找系统的最小错误识别代价$(C_{Det})_{Norm}$，从而确定话题的最佳聚类阈值θ_1和θ_2。其中，阈值 0 的范围由核心词簇和报道标题向量含有相同特征词的个数界定,阈值θ_2的范围--般为[0.4,0.6]。实验分为两部分，每部分采用不同的数据集，实验结果如表所示。

表一 聚类阈值比较表

聚类阈值θ_1	P_{miss}	P_{fa}	$(C_{Det})_{Norm}$	聚类阈值θ_2	P_{miss}	P_{fa}	$(C_{Det})_{Norm}$
0	0.0000	1.000	1.0000	0.40	0.0820	0.1933	0.2842
1/9	0.0971	0.2012	0.2786	0.42	0.1392	0.1728	0.2767
2/8	0.1182	0.1627	0.2017	0.44	0.1577	0.1401	0.2245
3/7	0.1401	0.1105	0.1705	0.46	0.1600	0.0994	0.2023
4/6	0.3127	0.0098	0.4415	0.48	0.1897	0.0821	0.2149
1	1.0000	0.0000	1.0000	0.50	0.2133	0.0533	0.2751

通过实验结果可以看出，随着聚类阈值θ_1和θ_2的不同，系统的检测代价上下波动。当聚类阈值θ_1从 0 上升到 3/7 时，漏检率略微升高，误检率明显下降，系统检测代价呈下降趋势;而当聚类阈值θ_1从 3/7 上升到 1 时，漏检率上升幅度增大，误检率略微下降，系统检测代价呈.上升趋势，由此可以看出，当聚类阈值θ_1=3/7 时，即核心词簇与报道标题向量具有 3 个相同的主题词时，系统具有最小的系统检测代价，因此令θ_1=3/7。同理，可以确定最佳聚类阈值θ_2为 0.46。

5.3 问题三的模型建立与求解

5.3.1 对问题三的分析与建立模型：

5.3.1.1 基于 AHP—熵权留言回复综合评价模型

网络留言回复往往与多种因素有关，多种因素间相互关联，相互制约，故在评价留言答复意见质量的主要从答复是否与留言相匹配、留言与答复间的时间间隔、公众留言所反映问题是否得到解决、答复内容中有关单位是否进行相关处理四个方面入手。

题目是要综合考虑留言答复质量，并结合附件内的部分留言答复，建立综合评价体系。本文将答复相关性、完整性、可解释性作为行车安全综合评价体系的三个一级指标，将政府处理能力、领导重视程度、地方政府层级、网络问政制度化水平、公众对事件的情感倾向、网民诉求量、互联网普及率、相关地区的经济发展水平等 8 个因素作为综合评价的二级指标。因为互联网的开放性会影响一级指标“状态”，本文将互联网中的一些不确定性因素设置为扰动变量。

5.3.1.2 建立评价指标体系

如果留言答复受到L个因素影响，这L个因素又各自受到若干个二级影响因

素的影响，则综合评价体系的指标设为$(U_{i1},U_{i2},...,U_{il})$，其中，$i=1,2,...,L$。$l$为相应的一级指标下二级指标的数量。

5.3.1.3 计算评价体系各指标权重

将灰色理论与模糊评价结合起来，建立一种基于模糊灰色综合评价方法，对综合评价答复质量进行更加准确、有效的评价。其中，确定各个评价指标的权重是进行灰色模糊综合评价的重要步骤，获取尽可能准确的权重对于最终结果的准确性有很大影响。层析分析法是一种常用的指标权重确定方法，是指将与决策总是有关的元素分解成目标、准则、方案等层次，例如：本文采用的三层分级结构，一层为综合质量评价，二级为影响综合评价的 3 个主要因素，三级为影响二级因素的 8 个指标。在此基础之上进行定性和定量分析的决策方法，但该方法存在主观性较强的缺点。为此，可引入熵权法对其三级指标权重进行修正。

5.3.1.4 层次分析法确定权重

层次分析法是由美国运筹学家 T. L. Satty 提出，是将决策问题按总目标、各层子目标、评价准则直至具体的备投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，求得每一层次的各元素对上一层次某元素的优先权重，最后再加权和的方法递阶归并各备择方案对总目标的最终权重，此最终权重最大者即为最优方案。其主要步骤包含建立层次结构、构造判断矩阵、计算各因素指标权重和矩阵一致性检验。假设由层次分析法确定的第i个指标的权重为w^i。

5.3.1.5 熵权法确定权重

熵本来是热力学中的概念，它由 C. E. Shannon 引入到信息论中，用来度量信息量。不确定性与信息量成反相关，因而可以用熵值来度量不确定性情况。指标对结果的影响程度随着其集中程度增高而降低，即一个系统越有序，信息熵就越低；系统越无序，信息熵越高熵权法是一种客观的确定权重的方法，使用时根据指标的变异程度来确定指标权重，能够尽可能消除人的主观干扰。熵权法的主要步骤如下。

指标数据标准化。假设由m个评价指标、n个评价对象构建的原始矩阵为：

$$X=(x_{ij})_{m\times n}=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \tag{23}$$

将该原始数据矩阵进行标准化处理得到$P=(p_{ij})_{m\times n}$，其中p_{ij}为第j个评价对象在第i个指标上的标准值。对于越大与优的指标而言

$$p_{ij}=\frac{x_{ij}-\min\limits_{j}\{x_{ij}\}}{\max\limits_{j}\{x_{ij}\}-\min\limits_{j}\{x_{ij}\}} \tag{24}$$

对于越小越优的指标而言

$$p_{ij}=\frac{\max\limits_{j}\{x_{ij}\}-x_{ij}}{\max\limits_{j}\{x_{ij}\}-\min\limits_{j}\{x_{ij}\}} \tag{25}$$

（2）定义熵。定义第i个指标的熵为

$$e_i=-k\sum_{j=1}^{n}z_{ij}\ln z_{ij}\qquad 1\le i\le m \tag{26}$$

式中：$z_{ij}=\dfrac{p_{ij}}{\sum\limits_{j=1}^{n}p_{ij}}$，$k=\dfrac{1}{\ln n}$，且当$z_{ij}=0$时，规定$z_{ij}\ln z_{ij}=0$

（3）定义熵权。将第i个指标的熵权定义为

$$w_i^{''}=\frac{1-e_i}{m-\sum\limits_{i=1}^{m}e_i} \tag{27}$$

其中：$0\le w_i^{''}\le 1$，$\sum\limits_{i=1}^{m}w_i^{''}=1$

5.3.1.6 确定指标综合权重

层次分析法确定的权重包含专家的主观判断，而熵权法侧重原始数据本身的客观信息。因而结合层次分析法和熵权法来确定权重是一种更加科学合理的方法。AHP-熵权法确定的综合权重为

$$w_i=\beta w_i^{'}+(1-\beta)w_i^{''} \tag{28}$$

式中，β为偏好系数，根据主客观相结合，客观权重优先的原则，将其取为

0.4，设最终确定一级指标权重为$(w_1,w_2,...,w_L)$，二级指标权重为$(w_{i1},w_{i2},...,w_{il})$，其中，$i=1,2,...,L$.

5.3.1.7 构造灰色模糊评价

灰色模糊综合评价模型能够充分考虑专家评判过程中评判信息的灰性特点。通过灰色相关理论计算得到灰色评价权值，利用灰色评价权值来进行构造模糊评价矩阵，最终使用模糊算法进行模糊综合评价。

5.3.1.8 确定评价权矩阵

（1）首先确定评价等级，本文将对行车安全的综合评价等级分为：火爆、较高、一般、正常、较低 5 个等级，各个等级得分用向量$u=(9,7,5,3,1,)$表示。通过查阅文献，得到满分为 10 分制的二级指标分数，假设第 j 个专家对第i个二级指标的评价为d_{ij}。

（2）确定评价灰类。由于各专家存在认知差异及经验限制，因而只能给出一个灰数的白化值。为了准确判断评价对象所属类别的程度，还需要确定评价灰类等级、灰数和白化权函数等。将评价等级分为 5 级，由灰数确定白化权函数个数。因此，设白化权函数为$f_k(d_{ij})$其中$k=1,2,3,4,5$

$k=1$时的百化权函数为

$$f_1(d_{ij})=\begin{cases}0 & d_{ij}\in(-\infty,0)\\ d_{ij}/9 & d_{ij}\in[0,9)\\ 1 & d_{ij}\in[9,+\infty)\end{cases} \qquad (29)$$

$k=2$时的百化权函数为

$$f_2(d_{ij})=\begin{cases}0 & d_{ij}\notin(0,14)\\ d_{ij}/7 & d_{ij}\in[0,7)\\ 2\text{-}d_{ij}/7 & \in[7,14]\end{cases} \qquad (30)$$

$k=3$时的百化权函数为

$$d_{ij})=\begin{cases}0 & d_{ij}\notin(0,10]\\ d_{ij}/5 & d_{ij}\in[0,5)\\ 2\text{-}d_{ij}/5 & \in[5,10]\end{cases} \qquad (31)$$

$k=4$时的百化权函数为

$$f_4(d_{ij})=\begin{cases}0 & d_{ij}\notin(0,6]\\ d_{ij}/3 & d_{ij}\in[0,3)\\ 2\text{-}d_{ij}/3 & \in[3,6]\end{cases} \tag{32}$$

$k=5$时的百化权函数为

$$f_5(d_{ij})=\begin{cases}0 & d_{ij}\notin(0,2]\\ d_{ij} & d_{ij}\in[0,1)\\ 2\text{-}d_{ij} & \in[1,2]\end{cases} \tag{33}$$

（3）计算灰色模糊评价权矩阵。设C_{ik}是答复相关性、完整性、可解释性三个一级指标下的第i个二级指标属于第$k(k=1,2,...,p)$个灰类的灰色评价系数，总的灰色评价系数为C_i。则根据和白化权函数可以得到：

$$C_{ik}=\sum_{j=1}^{n}f_k(d_{ij}) \tag{34}$$

$$C_i=\sum_{k=1}^{p}C_{ik} \tag{35}$$

由此得到评价权$r_{ik}=C_{ik}/C_i$，则该指标的灰色模糊评价向量$r_i=(r_{i1},r_{i2},...,r_{ip})$。计算得到该一级指标下面的所有二级指标灰色模糊评价向量，组成的灰色模糊评价权矩阵为：

$$R=\begin{bmatrix}r_{11} & \cdots & r_{1p}\\ \cdot & & \cdot\\ \cdot & & \cdot\\ r_{n1} & \cdots & r_{np}\end{bmatrix} \tag{36}$$

5.3.1.9 进行模糊综合评判

模糊综合评判需要综合分析各种影响因素，对评判目标进行客观评估，基本思想是根据最大隶属度原则，对各个影响因素进行全面评判。获得某个一级指标的灰色模糊评价权矩阵后，要对答复质量进行灰色模糊综合评价。建立评价指标体系阶段，假设留言答复的相关质量受L个一级指标影响，则一级指标集合为$U=\{U_1,U_2,...,U_L\}$；在确定指标权重阶段，一级指标综合权重为$w_i=(w_{i1},w_{i2},...,w_{il})$。假设某个一级指标包含 L 个二级指标，则该二级指标集合

表示为$U_i=\{U_{i1},U_{i2},...,U_{il}\}$，相应的综合权重为$w_i=(w_{i1},w_{i2},...,w_{il})$。

对该一级指标的灰色模糊评价权矩阵 R 和二级评价指标综合权重w_i进行模糊运算，得到该一级指标的模糊评价向量$B_I=w_i\bullet R$，其中“$\bullet$”是模糊运算的符号。

综合一级指标的模糊评价向量，可以得到答复质量指标体系的模糊评价矩阵R'。模糊评价矩阵R'结合一级指标的综合权重$w=(w_1,w_2,...,w_L)$，进行模糊运算得到答复质量评价总目标的结果为$B=w\bullet R'$，最终答复质量评价得分可以表示为：

$$Z=\alpha B\bullet v^T \tag{37}$$

其中，α是考虑了互联网不确定性的扰动变量，考虑互联网中某些因素的影响。

5.3.2 模型求解

5.3.2.1 建立留言答复综合评价指标体系

由前述：将答复相关性、完整性、可解释性作为行车安全综合评价体系的三个一级指标，将政府处理能力、领导重视程度、地方政府层级、网络问政制度化水平、公众对事件的情感倾向、网民诉求量、互联网普及率、相关地区的经济发展水平等 8 个因素作为综合评价的二级指标。因为互联网的不确定性因素会影响一级指标“状态”，本文将互联网中的某些因素为扰动变量。

表二 行车安全综合评价指标体系

系统层	要素层	指标层
留言答复质量综合评价指标体系U	相关性U_1	地区经济状况U_{11}
		互联网普及率U_{12}
	完整性U_2	网络问政的制度化水平U_{21}
		领导重视程度U_{22}
		地方政府层级U_{23}

		政府处理能力U_{24}
	可解释性U_3	网民诉求量U_{31}
		公众对事件的情感倾向U_{32}

4.3.2.2 计算指标权重

通过查阅文献结合专家打分，得到评价体系中的一级评价指标和二级评价指标的得分，打分结果用于层次分析法中构造判断矩阵，进而确定主观权重。

表三二级指标评价结果

二级评价指标	专家 1	专家 2	专家 3	专家 4	专家 5
U_{11}	6	6	7	6	5
U_{12}	6	7	6	5	6
U_{21}	4	6	5	7	5
U_{22}	5	5	5	5	4
U_{23}	7	6	5	6	6
U_{31}	7	6	7	6	6
U_{32}	5	4	5	4	6
U_{33}	7	6	5	6	7

根据表二中数据使用熵权法计算各个指标的信息熵，计算得到各个指标的客观权重。最后根据主观权重和客观权重，获得各个指标最终的综合权重。最终层次分析法获得的权重，熵权法获得的权重及综合权重，评价体系指标权重如表三所示。

5.3.2.3 灰色模糊评价模型并求解

以一级指标因素下的所有二级指标为例，构建灰色灰色模糊评价模型求解。根据二级指标分数，计算文中因素下的灰色评价系数，当 k=1 时，

C_{11}=$f_1(6)+f_1(6)+f_1(7)+f_1(6)+f_1(5)=5$，同理可得：$C_{12}$=7.012。总的灰色评价系数为$C_1=C_{11}+C_{12}=12.012$。然后计算得到第一个指标的评价权向量为$r_1=(0.464\ 0.536)$。同理可得其余的二级指标的评价向量为$r_2=(0.346\ 0.324\ 0.33)$，$r_3=(0.396\ 0.201\ 0.403)$。由此得到评价体系指标权重表如下表所示：

表四 评价体系指标权重

评价目标	一级评价指标	一级指标权重	二级评价指标	二级指标权重	信息熵	熵权	综合权重
U	U_1	0.431	U_{11}	0.463	0.705	0.604	0.566
			U_{12}	0.537	0.698	0.396	0.434
	U_2	0.294	U_{21}	0.336	0.768	0.431	0.349
			U_{22}	0.324	0.698	0.269	0.365
			U_{23}	0.34	0.674	0.3	0.286
	U_3	0.275	U_{31}	0.342	0.678	0.324	0.432
			U_{32}	0.213	0.657	0.264	0.151
			U_{33}	0.445	0.862	0.412	0.417

接下来根据第三章各公式计算各指标的模糊评价向量，综合所有，模糊评价向量可以得到质量评价体系的综合评价矩阵，再计算该指标体系综合模糊评价向量，得到答复质量综合得分，根据综合得分来判断综合答复水平。

六、模型评价及优缺点

（1）本文在正确、清楚地分析题意地基础上，建立了对应问题的各个模型。这些模型都有效的考虑了全局优化的问题，从而同时得到多个未知参数的最优组合解并且通过软件实现的参数解精度很高，因此得到的结果可信度较高。不仅如此，本文建立的模型能结合实际情况对问题进行求解，实用性强，具有很好的推广性。
（2）由于我们的能力有限，对建立 BGRU-CNN 混合模型等一系列模型的研究还不够深入、不够系统、不够完善，对于模型还需要进一步查资料验证。

（5）本文存在的不足是模型较依赖 MATLAB 的实现，这对模型求解造成了一定的误差。
（6）本文虽然将留言的热度给出了评价，但由于现实生活下的一些不确定因素仍会影响最终结果。

七、参考文献

[1]孙飞显. 政府负面网络舆情热度定量评价方法* ———以新浪微博为例[J]. 情报杂志, 2015(08):137-141.
[2]刘西林, 陈红捷. 电子政务留言反馈系统中的信息管理研究*[J]. 数据分析与知识发现, 2007, 2(2).
[3]张一文, 齐佳音, 方滨兴,等. Research on the Index System of Public Opinion on Internet for Abnormal Emergency%非常规突发事件网络舆情热度评价指标体系构建[J]. 情报杂志, 2010, 029(011):71-75,117.
[4]唐贤伦, 林文星, 杜一铭,等. 基于串并行卷积门阀循环神经网络的短文本特征提取与分类[J]. 工程科学与技术, 2019, 051(004):P.125-132.
[5]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(12):1278-1283.
[6]刘腾飞, 于双元, 张洪涛,等. 基于循环和卷积神经网络的文本分类研究[J]. 软件, 2018.
[7]李云红, 梁思程, 任劼,等. 基于循环神经网络变体和卷积神经网络的文本分类方法[J]. 西北大学学报:自然科学版, 2019(4):573-579.
[8]何跃, 蔡博驰. 基于因子分析法的微博热度评价模型 [J]. 统计与决策, 2016(18):52-54,共 3 页.
[9]赵爱华. 面向网络新闻的话题检测技术研究[D]. 山东师范大学.
[10]黄怡璇, 谢健民, 秦琴, et al. 影响网络舆情热度评价的主要因素识别研究[J]. 情报科学, 2017(10):51-56+64.
[11]殷风景. 面向网络舆情监控的热点话题发现技术研究[D]. 国防科学技术大学.

八、附录