摘要

近年来，随着互联网的广泛应用以及网络问政平台的出现，为政府了解民意提供了方便快捷的渠道，群众留言的分类和热点问题挖掘的重要性日益凸显，因此大量的文本分类和热点问题挖掘技术应运而生，通过对群众留言的分类，将群众留言分派至相应的职能部门处理，为人工处理节省时间提高效率；通过对热点问题的挖掘，找到相应的热点问题，以便相关部门及时解决，提升服务效率；针对留言问题的答复意见的评价，可以提高相关部门的服务，更好更快的解决问题。本文针对网络问政平台提供的群众留言数据，运用相应的文本挖掘技术进行了分析，首先我们对相应的文本进行相应的预处理，预处理的主要内容有处理缺失值、去掉重复的数据，然后利用 jieba 分词包对文本内容进行分词，分词完成后利用相应的停用词表去除多余的不必要的停用词，然后进行下一步的文本挖掘工作：

针对问题一，我们构建了基于 LSMT 的多分类模型，它是一种时间循环神经网络，利用附件二分好类的一级标签的数据，对该模型进行训练，随着训练周期的增加,模型在训练集中准确率越来越高，而在训练集中的损失越来越小。经过不断的优化，最终得到该模型的 F1 值为 0.85。

针对问题二，我们建立了基于 word2vec 的文本相似度模型，为了保证模型的完整性，我们将切词未去停用词的附件三的留言主题作为语料放入模型进行训练，得到 word2vec 模型。通过对测试集的测试，我们将文本的相似度定义为 0.75，对输出大于 0.75 相似文本的数量进行排序，得到相似话题数量最多的文本为 47 条。然后通过相应话题总的点赞数和反对数的差除以它们的乘积相当于权重一个权重指标，由于点赞数和反对数有可能为 0 ，所以我们通过话题数量乘以权重指标再加上话题数量，然后得到热度评价指标，最终得到热度排名前五的热点问题。

针对问题三，在问题二建立的 word2vec 模型的基础上，我们将切词后的附件四中的留言详情和答复意见放入模型中进行训练，计算出关于留言详情与对应答复意见的相似度，再对留言时间和答复时间的时间间隔取权重，以天为单位，赋值范围为 1-2 之间，权重越大表示时间间隔越短，再计算答复意见的字长，并且给字长的长度也进行赋值，为了保证和时间的一致性，我们给长度的赋值范围也是 1-2，权重越大则表示长度越长，最后通过相似度、时效性、完整性这三个指标的乘积得到一个综合评价指标，利用综合评价指标来对留言回复进行打分。

关键词：文本挖掘；多分类模型；文本相似度；热度评价指标；综合评价指标

目录

1 挖掘目标

文本挖掘是抽取有效、新颖、有用、可理解的、散布在文本文件中的与价值的知识，并且利用这些知识更好地组织信息的过程。也是信息挖掘的一个研究分支，用于基于本文信息的知识发现。文本挖掘利用智能算法，如神经网络、基于案例的推理、可能性推理等，并结合文字处理技术，分析大量的非结构化文本源（如文档、电子表格、客户电子邮件、问题查询、网页等，抽取或标记关键词概念、文字间的关系，并按照内容对文档进行分类，获取有用的知识和信息。文本挖掘的基本技术有五大类，包括文本信息抽取、文本分类、文本聚类 、文本数据压缩、文本数据处理。

随着网络问政平台的逐步发展，已经成为政府了解民意、汇聚民智、凝聚民气、解决民难的重要渠道，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整理的相关部门带来了极大的挑战。同时，随着大数据、人工智能等技术的发展，建立基于文本挖掘和自然语言处理技术的智慧政务系统已经是社会治理的创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。本文主要解决以下几个问题：

一是留言分类问题，目前大部分电子政务系统还是依靠传统的人工分类，存在工作量打、效率低，且差错率高的问题，建立关于留言内容的分类模型为人工处理节省时间提高效率。二是热点问题挖掘问题，建立聚类模型，将热点问题聚类，及时发现热点问题，有助于相关部门进行有针对性的处理，提高服务效率。三是对答复意见的评价，答复时间是否及时、答复的内容是否相关和完整，构建对评价质量打分的评价方案，从而提高相关部门的服务体系。

2 总体流程

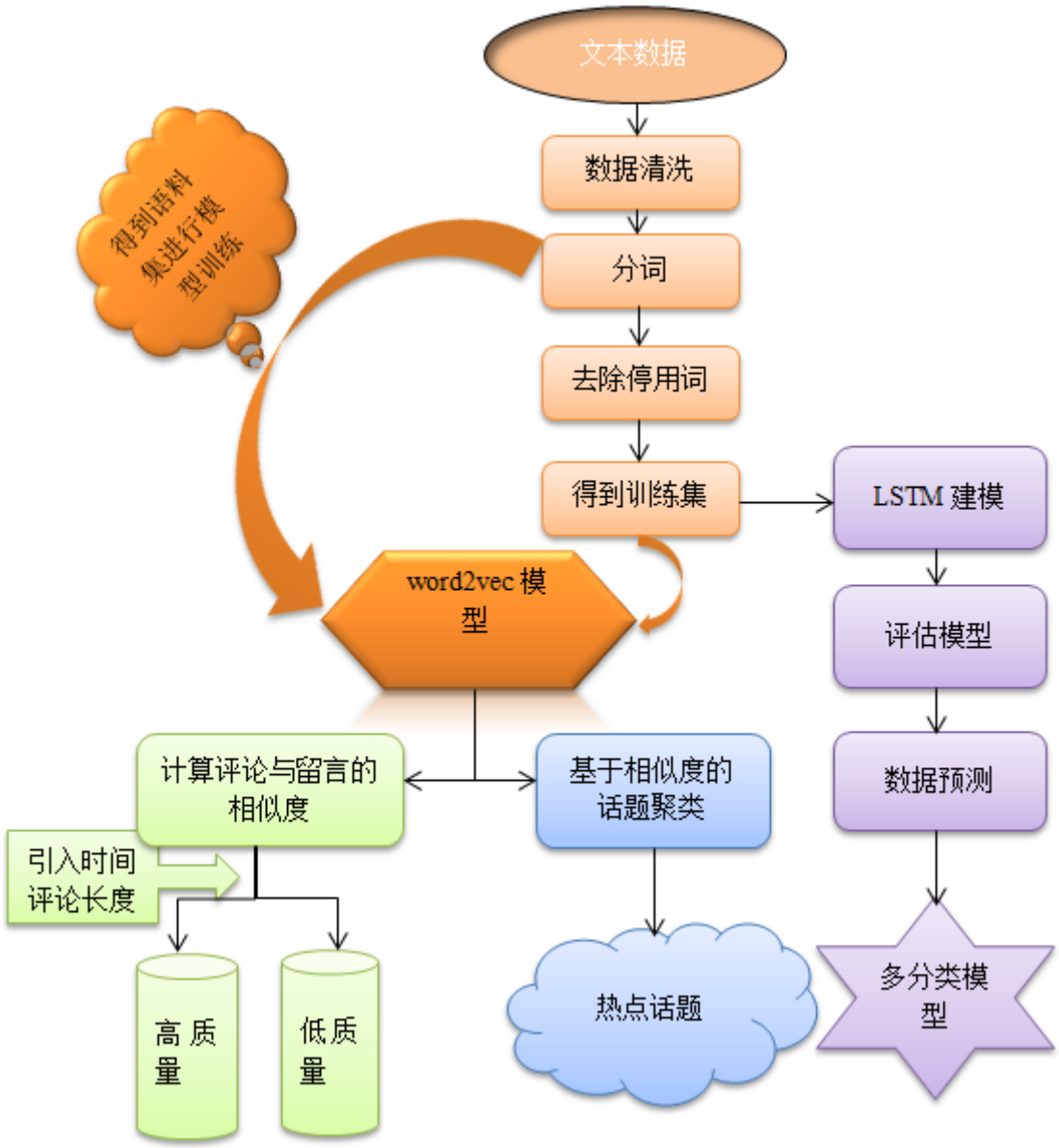


图 1 总体流程

3 数据预处理

3.1 数据清洗

取到文本后，我们首先对要进行的文本数据进行预处理，预处理的第一步为数据清洗，通过数据清洗对数据进行重新审查和校验的，查看是否存在缺失数据，

删除重复信息、纠正存在的错误，并提供数据一致性，以便数据的后续使用。

3.2 分词

中文分词：指以词作为基本单元，使用计算机自动对中文文本进行词语的切分，即使词之间有空格，这样方便计算机识别出各句的重点内容。

在预处理后，在这里我们用到 jieba 中文分词包对评论进行分词处理，将评论分成由空格隔开的一个个单独的单词。

A市西湖建筑集团占道施工有安全隐患	A 市 西湖 建筑 集团 占 道 施工 安全隐患
A市在水一方大厦人为烂尾多年安全隐患严重	A 市 在水一方 大厦 人为 烂尾 多年 安全隐患 严重
投诉A市A1区苑物业违规收停车费	投诉 A 市 A1 区苑 物业 违规 收 停车费
A1区蔡锷南路A2区华庭楼顶水箱长年不洗	A1 区 蔡锷 南路 A2 区华庭 楼顶 水箱 长年 不洗
A1区A2区华庭自来水好大一股霉味	A1 区 A2 区华庭 自来水 好大 一股 霉味
投诉A市盛世耀凯小区物业无故停水	投诉 A 市 盛世 耀凯 小区 物业 无故 停水
咨询A市楼盘集中供暖一事	咨询 A 市 楼盘 集中 供暖 一事
A3区桐梓坡西路可可小城长期停水得不到解决	A3 区 桐梓 坡 西路 可可 小城 长期 停水 得不到 解决
反映C4市收取城市垃圾处理费不平等的问题	反映 C4 市 收取 城市 垃圾处理 费 平等 问题
A3区魏家坡小区脏乱差	A3 区 魏家坡 小区 脏乱差
A市魏家坡小区脏乱差	A 市 魏家坡 小区 脏乱差

图 2 分词处理

而有些特定词组不需要分开，故将我们需要用到的特定词组写入到字典里，例如某市，某区。

```
0
2549                                [A2区, 景蓉华苑, 物业管理, 有, 问题]
2554                [A3区, 潇楚, 南路, 洋湖, 段, 怎么, 还, 没, 修好, ？]
2555                  [请, 加快, 提高, A市, 民营, 幼儿园, 老师, 的, 待遇]
2557          [在, A市, 买, 公寓, 能, 享受, 人才, 新政, 购房, 补贴, 吗, ？]
2574                     [关于, A市, 公交站点, 名称, 变更, 的, 建议]
2759                        [A3区, 含浦镇, 马路, 卫生, 很差]
2849                [A3区, 教师, 村, 小区, 盼望, 早日, 安装, 电梯]
3681             [反映, A5区, 东澜湾, 社区, 居民, 的, 集体, 民生, 诉求]
3683       [反映, A市, 美麓, 阳光, 住宅楼, 无故, 停工, 以及, 质量, 问题]
```

图 3 分词处理

3.3 去停用词

因为需要处理的评论内容为中文，所以要针对中文文本进行一系列处理，中文表达中最常用的功能性词语是限定词，如“的”、“一个”、“这”、“那”等。这些词语的使用较大的作用仅仅是协助一些文本的名词描述和概念表达，并没有太多的实际含义。而大多数时候停用词都是非自动生产、人工筛选录入的，因为需要根据不同的研究主题人为地判断和选择合适的停用词语。

4 构建算法

4.1 基于 LSTM 的文本多分类

4.1.1 LSTM 网络

LSTM 网络为长短期记忆网络的简称，LSTM 它是一种时间循环神经网络，LSTM 结构图如下所示

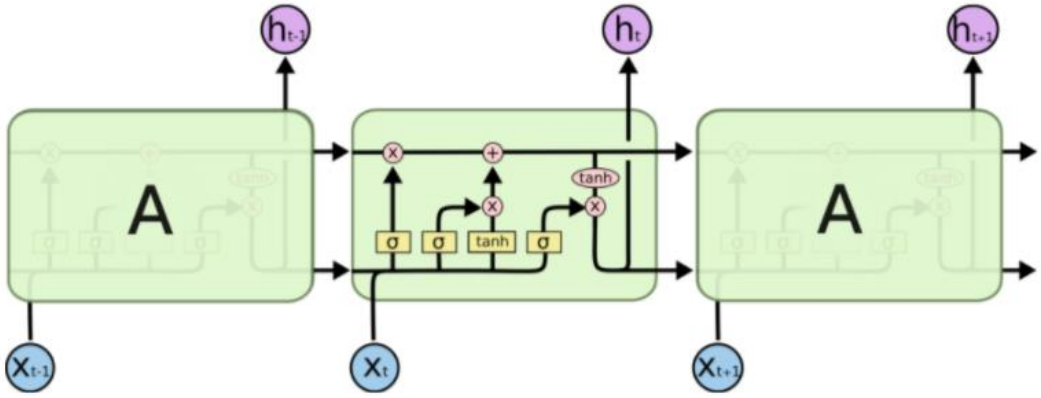
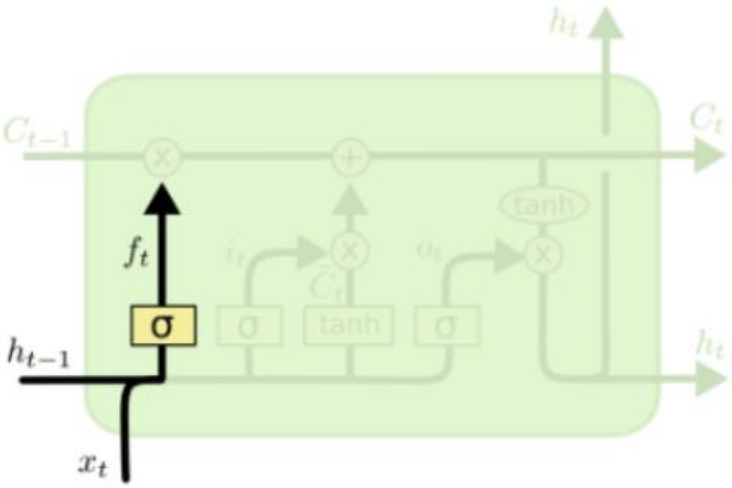


图 4 LSTM 网络

在上图中，每箭头都携带一个向量，从上一个节点的输出到其他节点的输入。粉色圆圈表示逐点运算，如矢量加法，而黄色框表示神经网络层。 箭头合并表示连接，而箭头分叉表示其内容被复制，副本将转移到不同的位置。

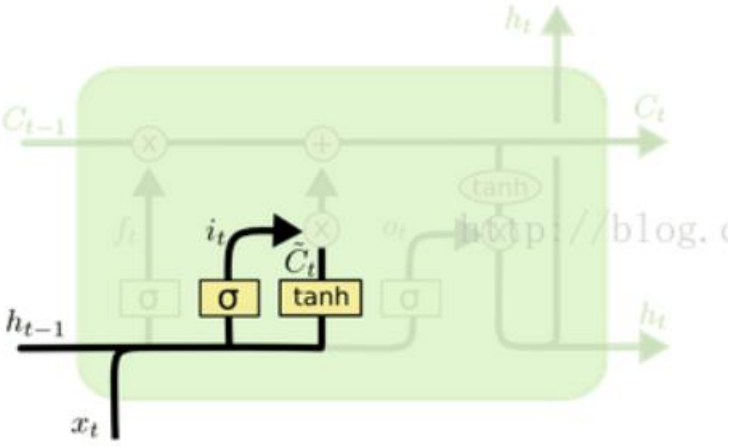
LSTM 工作原理：

（1） forget gate：选择忘记过去某些信息：



$$f_t = \sigma\left(W_f \cdot \left[h_{t-1},\ x_t\right] + b_f\right) \tag{1}$$

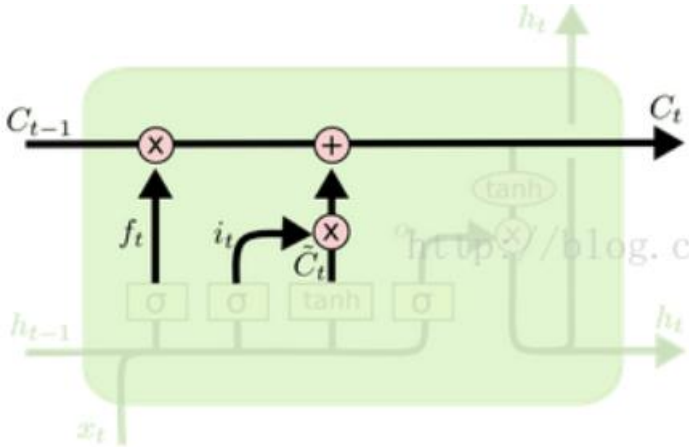
（2） input gate：记忆现在的某些信息：



$$i_t = \sigma\left(W_i \cdot \left[h_{t-1},\ x_t\right] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1},\ x_t\right] + b_C\right) \tag{3}$$

（3） 将过去与现在的记忆进行合并：



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

（4） output gate：输出

$$O_t = \sigma\left(W_o \cdot \left[h_{t-1},\ x_t\right] + b_o\right) \tag{5}$$

$$h_t = O_t * \tanh\left(C_t\right) \tag{6}$$

4.1.2 模型建立

我们首先将分词处理后的评论向量化处理,设置最频繁使用的 8000 个词，设置每条评论的最大词语数为 100，超过部分会被截去，不足部分会补 0。

取 10%的数据作为训练集，其余 90%为测试集

定义一个 LSTM 的序列模型:

- 模型的第一次是嵌入层(Embedding)，它使用长度为 100 的向量来表示每一个词语
- SpatialDropout1D 层在训练中每次更新时， 将输入单元的按比率随机设置为 0， 这有助于防止过拟合
- LSTM 层包含 100 个记忆单元
- 输出层为包含 10 个分类的全连接层
- 由于是多分类，所以激活函数设置为'softmax'
- 由于是多分类，所以损失函数为分类交叉 categorical_crossentropy

- 定义好 LSTM 模型后，令训练周期为 5，batch_size 为 64，训练结果如下图所示。

```
Epoch 1/5
7460/7460 [==============================] - 41s 5ms/step - loss: 1.6823 - accuracy:
0.3780 - val_loss: 1.1215 - val_accuracy: 0.6128
Epoch 2/5
7460/7460 [==============================] - 41s 6ms/step - loss: 0.6953 - accuracy:
0.7796 - val_loss: 0.6225 - val_accuracy: 0.8191
Epoch 3/5
7460/7460 [==============================] - 42s 6ms/step - loss: 0.3487 - accuracy:
0.9107 - val_loss: 0.5556 - val_accuracy: 0.8420
Epoch 4/5
7460/7460 [==============================] - 42s 6ms/step - loss: 0.1665 - accuracy:
0.9570 - val_loss: 0.5688 - val_accuracy: 0.8432
Epoch 5/5
7460/7460 [==============================] - 42s 6ms/step - loss: 0.0937 - accuracy:
0.9788 - val_loss: 0.6239 - val_accuracy: 0.8372
921/921 [==============================] - 1s 2ms/step
```

图 5 训练结果

通过上面训练数据损失函数趋势图和准确率趋势图

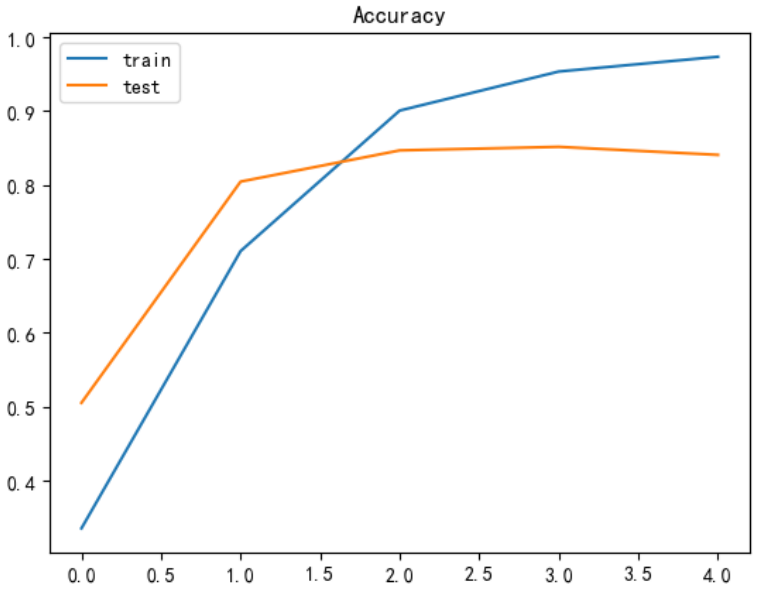


图 6 准确率趋势图

随着训练周期的增加,模型在训练集中准确率越来越高。

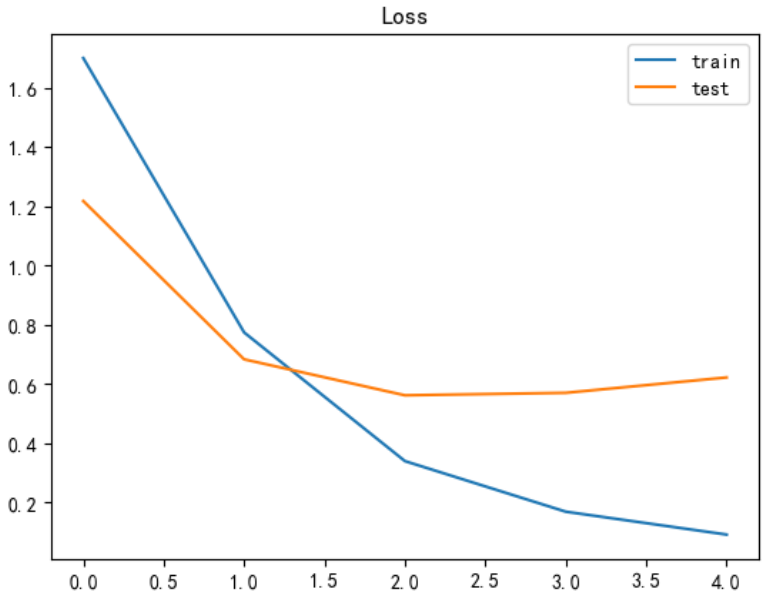


图 7 损失趋势图

随着训练周期的增加,模型在训练集中损失越来越小。

4.1.3 LSTM 模型的评估

通过画混淆矩阵和求 F1 分数来评估模型

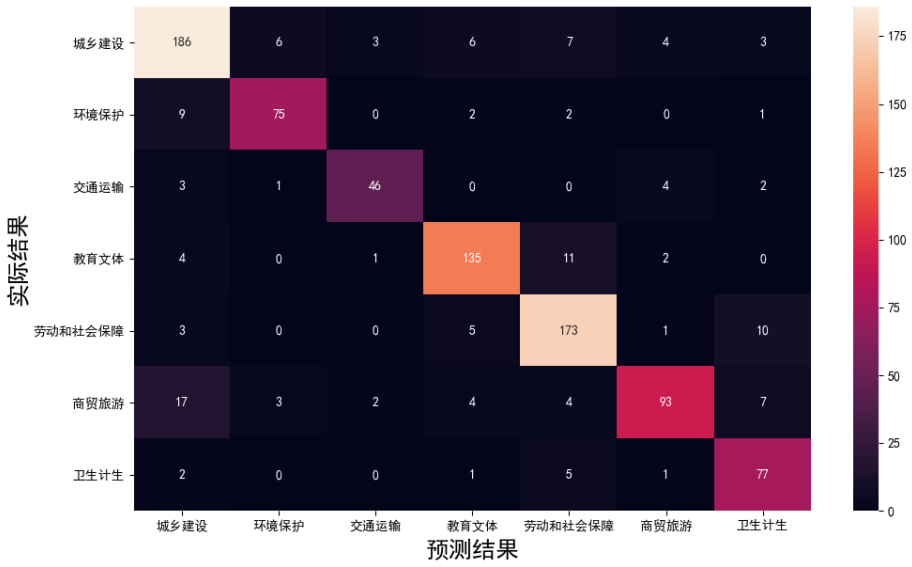
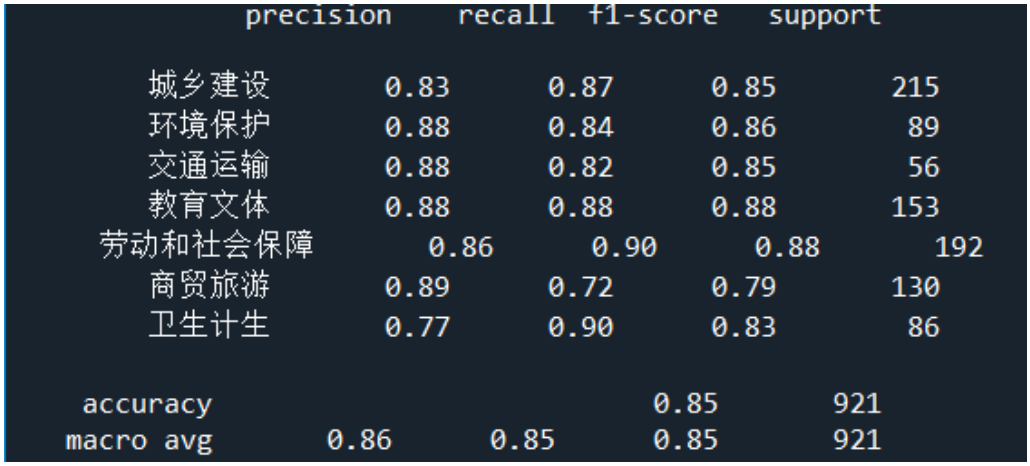


图 8 混淆矩阵

混淆矩阵的主对角线表示预测正确的数量,除主对角线外其余都是预测错误的数量。

多分类模型一般不使用准确率(accuracy)来评估模型的质量,因为accuracy 不能反应出每一个分类的准确性,因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，accuracy 不能反映出模型的实际预测精度,这时候我们就需要借助于 F1 分数。



	precision	recall	f1-score	support
城乡建设	0.83	0.87	0.85	215
环境保护	0.88	0.84	0.86	89
交通运输	0.88	0.82	0.85	56
教育文体	0.88	0.88	0.88	153
劳动和社会保障	0.86	0.90	0.88	192
商贸旅游	0.89	0.72	0.79	130
卫生计生	0.77	0.90	0.83	86
accuracy			0.85	921
macro avg	0.86	0.85	0.85	921

图 9 F1 分数

从上图可以看到，模型的 F1 分数为 0.85 。对于一般得分类模型，评价指标达到 0.8，我们就认为模型可靠，所以我们认为该分类模型可靠。

4.2 基于 word2vec 的文本相似度计算

4.2.1 CBOW 神经网络模型

CBOW 是一种根据上下文的词语预测当前词语出现概率的模型,即预测 $P(w_t \mid w_{t-k}, w_{(t-k-1)}, \ldots w_{t-1}, w_{t+1}, w_{t+2} \ldots, w_{t+k})$

$$L = \sum_{w \in C} \log w(Context\ x\ (\tag{7}$$

如果出现上下文，词 w 我们希望它出现的概率应该是越大越好的。如下图所示为 CBOW 模型：

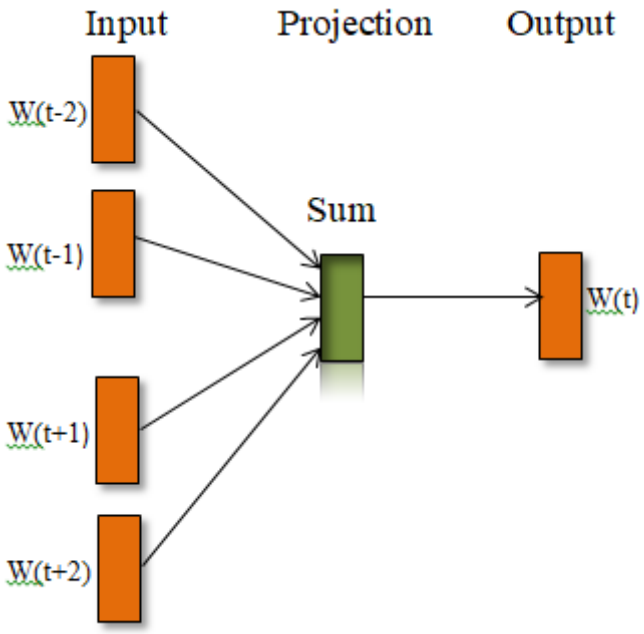


图 10 CBOW 模型

输入层是上下文的词语的词向量，在训练 CBOW 模型，词向量是一个副产品，确切来说，是 CBOW 模型的一个参数。训练来时的时候，词向量是个随机值，随着训练的进行不断被更新。隐藏层是对其求和，所谓求和就是简单的加法。输出层输出最可能的 W 。输入向量通过一个 $V*N$ 维的权重矩阵 W 连接到投影层；隐藏层通过一个 $N*V$ 的权重矩阵 $W^{'}$ 连接到输出层。假设知道输入和输出权重矩阵的大小。

第一步：计算隐藏层 h 的输出

$$h=\frac{1}{C}W\cdot(\sum_{i=1}^{C}x_i) \tag{8}$$

该输出就是输入向量的加权平均。

第二步：计算输出层每个结点的输入

$$u_j=v_{\omega_j}^{'T}\cdot h \tag{9}$$

其中 $v_{\omega_j}^{'T}$ 是输出矩阵 $W^{'}$ 的第 j 列。

最后我们计算输出层的输出，输出 y_j 如下：

$$y_{c,j}=p\left(\omega_{y,j}|\omega_1,\ldots,\omega_c\right)=\frac{\exp\left(u_j\right)}{\sum_{j=1}^{V}\exp\left(u'j\right)} \tag{10}$$

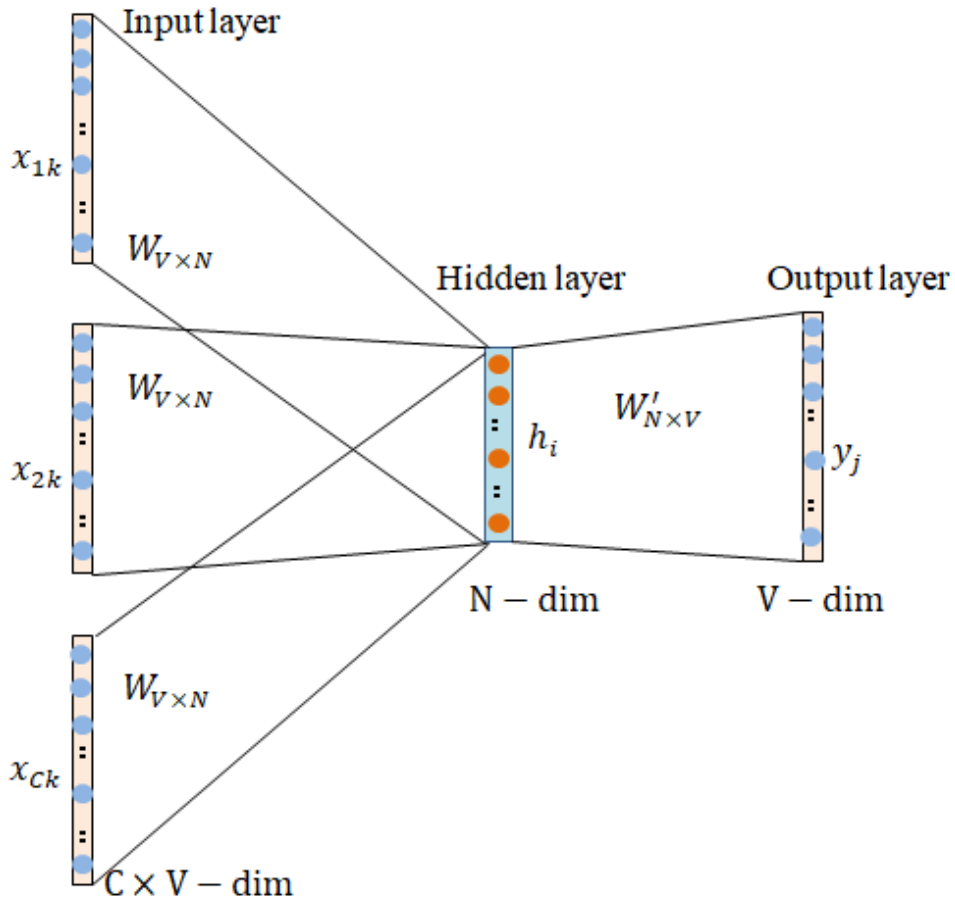


图 11 CBOW 神经网络模型

由于语料中的词汇量是固定的|C|个，所以上述过程其实可以看做一个多分类问题，给定一个特征，从|C|个分类中挑一个。

4.2.2 Skip-Gram 模型

Skip-Gram 模型正好与 CBOW 相反，从图 10 可以看出 Skip-Gram 应该预测概率 $P(w_i \mid w_t)$，其中 $t-c \le i \le t+c$且 $i \ne t$，c 是决定上下文窗口大小的常数，c 越大则需要考虑的 pair 就越多一般能够带来更精确的效果，但是训练的时间也会增加。假设存在一个 $w_1, w_2, \ldots, w_t$ 的词组序列，Skip-Gram 的目标是最大化：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c,\, j\ne 0}\log p(w_{t+j} \mid w_t) \tag{11}$$

基本的 Skip-Gram 模型定义 $P(w_0 \mid w_I)$

$$P(w_0|w_t) = \frac{e^{v'^{T}_{w_0} v_{w_I}}}{\sum_{w=1}^{w} e^{v'^{T}_{w_0} v_{w_I}}} \tag{12}$$

Skip-Gram 中的每个词向量表征了上下文的分布。Skip-Gram 中的 skip 是指在一定窗口内的词两两都会计算概率，就算他们之间隔着一些词，这样的好处时“白色汽车”和“白色的汽车”很容易被识别为相同的短语。

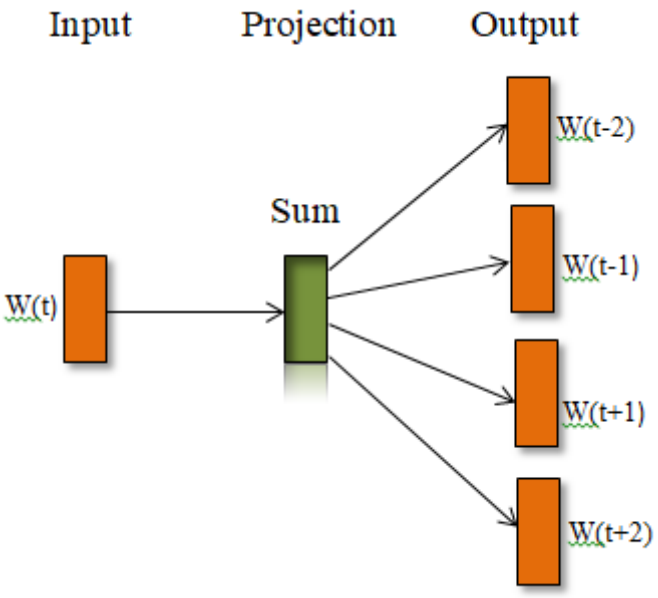


图 12 Skip-Gram 模型

4.2.3 构建模型

基于 CBOW 和 Skip-Gram 的思想，word2vec 实际上分为了两部分，第一步是建立模型，第二部分是通过模型获取嵌入词向量。Word2vec 先基于训练数据构建一个神经网络，当这个模型训练好以后，得到这个模型通过训练数据所得的参数，例如隐层的权重矩阵。Word2vec 通过学习文本来用词向量的方式表征词的语义信息，即通过一个嵌入空间使得语义上相似的单词在该空间内距离很近，从而得到我们想要的结果。

我们直接使用 python 中的 gensim 库来构建模型，为了保证模型的完整性，我们将分好词没有去停用词的留言主题作为语料放入 word2vec 中进行模型的训练，得到训练好的模型；再将经过分词停用词处理后的文本输入，输出文本与语料相似的文本。为了减小误差，在这里我们将相似度的阈值设定为 0.75，为保证模型的准确性，我们先输入一条文本进行测试，如图所示为输出结果：

```
[{3415: 'A5 区 劳动 东路 魅力 之 城 小区 底层 餐馆 油烟 扰民', 'score': 1.0},
 {4313: 'A5 区 劳动 东路 魅力 之 城 小区 底层 餐馆 油烟 扰民', 'score': 1.0},
 {2041: 'A5 区 劳动 东路 魅力 之 城 小区 油烟 扰民', 'score': 0.9814384},
 {4312: 'A5 区 劳动 东路 魅力 之 城 小区 油烟 扰民', 'score': 0.9814384},
 {2449: 'A5 区 劳动 东路 魅力 之 城 小区 临街 门面 烧烤 夜宵 摊', 'score': 0.9156687},
 {4314: 'A5 区 劳动 东路 魅力 之 城 小区 临街 门面 烧烤 夜宵 摊', 'score': 0.9156687},
 {272: '魅力 之 城 小区 临街 门面 油烟 直排 扰民', 'score': 0.8577281},
 {4311: '魅力 之 城 小区 临街 门面 油烟 直排 扰民', 'score': 0.8577281},
 {4057: 'A5 区 劳动 东路 魅力 之 城 小区 一楼 的 夜宵 摊 严重 污染 附近 的 空气', 'scor
e': 0.8562383},
 {4318: 'A5 区 劳动 东路 魅力 之 城 小区 一楼 的 夜宵 摊 严重 污染 附近 的 空气', 'scor
e': 0.8562383},
 {3549: 'A5 区 劳动 东路 魅力 之 城 小区 一楼 的 夜宵 摊 严重 污染 附近 的 空气 ， 急需
处理 ！',
  'score': 0.8161653},
 {4319: 'A5 区 劳动 东路 魅力 之 城 小区 一楼 的 夜宵 摊 严重 污染 附近 的 空气 ， 急需
处理 ！',
  'score': 0.8161653},
 {2381: '万科 魅力 之 城 小区 底层 门店 深夜 经营 ， 各种 噪音 扰民', 'score': 0.789952
16},
 {4320: '万科 魅力 之 城 小区 底层 门店 深夜 经营 ， 各种 噪音 扰民', 'score': 0.789952
16},
 {2288: 'A5 区 魅力 之 城 小区 一楼 被 搞 成 商业 门面 ， 噪音 扰民 严重', 'score': 0.7
5876933},
 {4316: 'A5 区 魅力 之 城 小区 一楼 被 搞 成 商业 门面 ， 噪音 扰民 严重', 'score': 0.7
5876933}]
```

图 13 文本相似度

从以上结果可以看出，word2vec 模型的预期结果不错，相似度在 0.75 以上的文本基本上与输入文本相似，可归为同一类问题，说明该模型可靠。

5 结果分析

5.1 自定义预测

由上面部分，我们建立的 LSTM 模型的 F1 分数为 0.85，

	precision	recall	f1-score	support
城乡建设	0.83	0.87	0.85	215
环境保护	0.88	0.84	0.86	89
交通运输	0.88	0.82	0.85	56
教育文体	0.88	0.88	0.88	153
劳动和社会保障	0.86	0.90	0.88	192
商贸旅游	0.89	0.72	0.79	130
卫生计生	0.77	0.90	0.83	86
accuracy			0.85	921
macro avg	0.86	0.85	0.85	921

图 14 F1 分数

下面我们通过定义一个预测函数：通过输入一段话，来判断它属于哪一个一级标签。

```
In [11]: predict('A市西湖建筑集团占道施工有安全隐患')
Out[11]: '城乡建设'

In [12]: predict('举报M3县卢家村公路硬化工程和砂石路面建设工程的招投标有违规行为')
Out[12]: '城乡建设'

In [13]: predict('请求解决原G市公路局合同工社保、医保、工伤保险的报告')
Out[13]: '劳动和社会保障'

In [14]: predict('D2区雷公塘皓源实验中学寒假违规收费补课')
Out[14]: '教育文体'

In [15]: predict('西地省人事厅废止的法规文件还执行吗？')
Out[15]: '劳动和社会保障'

In [16]: predict('举报补课行为')
Out[16]: '教育文体'
```

图 15 预测函数测试

5.2 热度评价指标

人们在评价一个话题的热度时，比如说微博话题的热度，评价的指标主要有发布次数、转发量、点赞量等，所以在这里我们将热点话题的指标分为三个——相似话题出现的总的次数、该话题总的点赞数、该话题总的反对数。

定义相似话题出现的总次数设为 N，相似话题总的点赞数设为 p，相似话题总的反对数设为 q，由于社区热点话题不像微博热点话题随时随地都有点击量，所以社区话题的点赞量和反对量跟话题出现次数显然相比起来就没有那重要，所以我们将点赞数减去反对数再除以两者的乘积作为一个权重指标，故我们将热度评价指标定义为：

$$y = N \times (1 + \frac{p - q}{p * q}) \tag{13}$$

y 的值越高，则说明该话题的热度越高。通过 word2vec 模型我们得到相似度数量最多的前十个话题为：

表 1 前十热点话题

话题	出现次数	地点/人群	问题描述
1	47	A 市伊景园滨河苑	A 市伊景园滨河苑捆绑销售车位
2	37	A 市 A2 区丽发新城	A2 区丽发新城小区遭搅拌站严重污染
3	30	A 市多个小区	A 市多小区工地长期施工扰民
4	17	西地省	西地省多个集团涉嫌诈骗
5	12	A5 区魅力之城小区	A5 区魅力之城小区底层餐馆油烟扰民
6	9	A 市	A 市人才购房补贴政策
7	9	A 市	增加 A 市公交车
8	9	A 市经济学院	A 市经济学院强制学生实习
9	9	A3 区西湖街道茶场村	A3 区西湖街道茶场区五组拆迁
10	9	A7 县星沙四区	A7 县星沙四区凉塘路旧城改造

通过对这些话题的热度指标进行计算排序得到前五的热度话题为：

表 2 前五的热度话题

热度排名	热度指数	时间范围	地点/人群	问题描述
1	91.96	2019/7/28 至 2019/9/1	A 市伊景园滨河苑	A 市伊景园滨河苑捆绑销售车位
2	54.71	2019/11/2 至 2020/1/25	A 市 A2 区丽发新城	A 市 A2 区丽发新城小区遭搅拌站严重污染
3	30.00	2019/1/16 至 2019/12/6	A 市多个小区	A 市多地小区工地长期施工扰民
4	24.73	2019/1/8 至 2019/10/27	西地省	西地省多个集团涉嫌诈骗
5	12.00	2019/7/21 至 2019/09/25	A5 区劳动东路魅力之城小区	A5 区劳动东路魅力之城小区底层餐馆油烟扰民

5.3 答复意见质量评价

题目是根据答复的相关性、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。我们从评价的时效性、完整性、相似度来对评价的质量进行打分。

1) 时效性

假设 t 为时间间隔，单位是天，我们给时间间隔权重的值为 1-2:

表 3 按时间间隔赋值

答复时间间隔	权重
$t \geq 50$	1
$20 \leq t < 50$	1.2
$10 \leq t < 20$	1.4
$5 \leq t < 10$	1.8
$t < 5$	2

表 2 时间间隔分布情况

权重值	数量	百分比
1	219	7.78%
1.2	612	21.73%
1.4	727	25.82%
1.8	594	21.09%
2	664	23.58%
总值	2816	100%

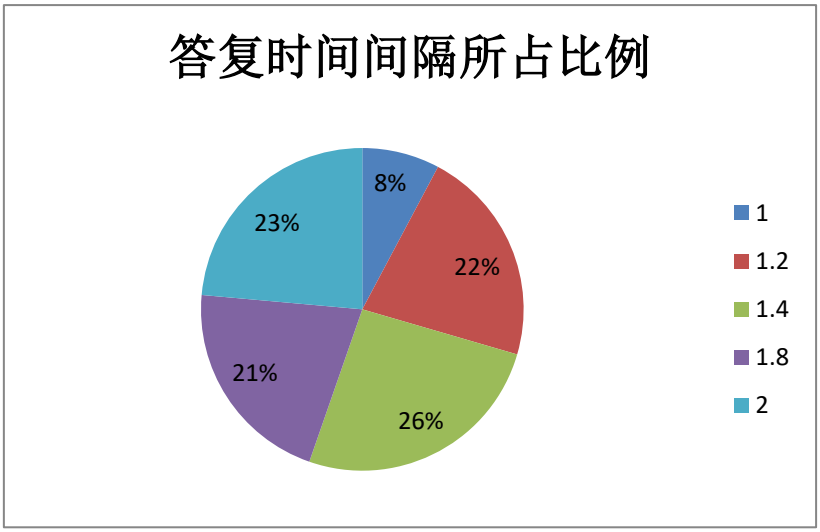


图 16 答复时间间隔所占比例

从图 13 可以看出，答复时间超过 50 天的所占比例比较少，不超过 8%，在 5 天之内、10 天之内、20 天之内就会答复网友问题的占比较高的比例，有的甚至高达 25.82%。这就充分体现了及时性，也可以看出相关部门对网友的问题比较重视。

2) 完整性

大致浏览附件 4 的答复意见可以发现提出问题所给的答复意见有很多内容，但是有些答复意见上面写着“网友：您好！留言已收悉”、“网友：您好！2019 年 3 月 5 日”、“网友：您好！您反映的问题，我们已转交相关单位。2019 年 8 月 12 日”等信息，相关部门根本没有回复网友所提出的问题。根据这些可以发现没有具体回复问题的答复意见的字符长度比较短，那就可以以一个长度为临界值，小于这个临界值（设为 100）的就是在答复意见上面写着“网友：您好！留言已收悉”等之类的信息，而大于临界值的都是有具体的回复。因此根据答复意见的不同长度设不同的权重值。如表 5、表 6 所示。

设 x 为答复意见的字符长度

表 5 答复意见字符长度赋值

答复意见长度	权重
$x<100$	1
$100\le x<200$	1.2
$200\le x<300$	1.4
$300\le x<400$	1.5
$400\le x<500$	1.6
$600\le x<800$	1.7
$800\le x<1000$	1.8
$1000\le x<2000$	1.9
$x>2000$	2

表 6 长度分布情况

权重值	数量	百分比
1	477	16.94%
1.2	514	18.25%
1.4	549	19.50%
1.5	432	15.34%
1.6	440	15.63%
1.7	214	7.60%
1.8	79	2.81%
1.9	90	3.20%
2	21	0.75%
总值	2816	100%

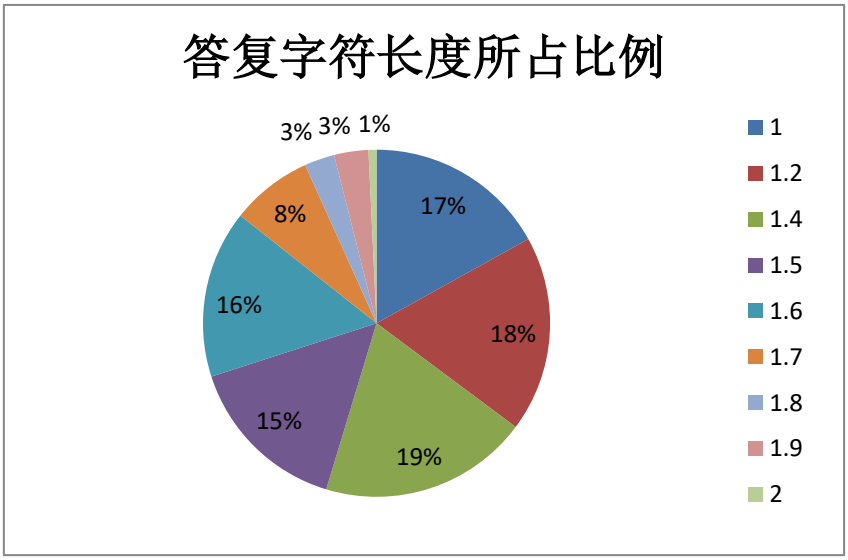


图 17 答复字符长度所占比例

根据上图的图 14 所示，可以发现相关部门具体答复了网友问题的占 83%左右，这是一个相当高的比例，也为验证答复意见的完整度作为一个依据，从而验证相关性。在剩下的 17%左右的比例中，可以直观地看出留言主题、留言详情和答复意见之间的完整度比较低，因为没有回答网友的问题。

3) 相似度

利用 wode2vec 计算数据中评论留言和评论回复的相似度，在得到的相似度数据中我们发现有相似度很高以及相似度为负数的情况。如表 7 所示为前 20 条答复的相似度：

表 7 前 20 条答复的相似度

编号	相似度	编号	相似度
1	0.52	11	0.59
2	0.26	12	0.67
3	0.68	13	0.13
4	0.71	14	-0.07
5	0.66	15	0.38
6	0.20	16	0.60
7	0.05	17	-0.01
8	0.08	18	0.58
9	0.49	19	0.46
10	0.66	20	0.16

以下列出相似度为正数、负数的情况：

◆ 相似度为负数的情况：

- 答非所问

留言主题	投诉 A 市外国语学校国庆补课
留言详情	A 市外国语学校国庆补课，只放四天，剩下三天全部考试，下发的国庆安全通知上面明明确确写清了放假从 10.1-10.7，但只放四天，学生好不容易盼来的国庆七天假，现在不仅不让我们放松，放松时间段外，还在一收假就考月考，这种事为什 么教育局不管啊，不是犯法吗补课?难道学生就没有自己的权利吗，中秋节也是只放了一天半，难道中国高考政策不进行改革，教育局就永远对这种非法补课不至于理睬?睁一只眼闭一只眼?我希望我们的高中生活是非富多彩的而不是阴暗，无趣，筋疲力竭的，但我请你们让学生好好放一次假吧，上个星期连上六天，上上个星期连上七天。教育局你们该管管了。
答复意见	网友“UU008310”您好！您的留言已收悉。现将有关情况回复如下：经查，该工地是龙湖春江郦城项目工地，目前已进入渣土回填、基底施工阶段。该工地办理了由 A2 区渣土管理办公室审批的《A2 区工程项目回填审结许可》（有效期：2018.09.28-10.28）办理了渣土回填证，在此期间内夜间施工，属于合法施工。该工地办理了 3 次夜间混凝土浇筑施工许可证（有效期 2018.08.23-08.25、09.06-09.08、09.17-09.19，施工时间：22:00-06：00），近期因天气原因暂未施工。2018 年 10 月 17 日 23 时 20 分，A2 区城管执法大队渣土中队执法队员赶到该工地现场核实情况，当天夜间没有施工。下一步，A2 区城管执法大队渣土中队将加大对该工地的巡查力度，确保该工地有证施工并控制施工噪音在规定范围内，如该工地出现无证施工情况，将依法处理。感谢您对我们工作的支持、理解与监督！2018 年 10 月 22 日。

- 没有回答

留言主题	关于尽快开通 12 路公交线路的情况反映
留言详情	尊敬的易书记：您好！我是高新先导区的一名普通企业员工，现有如下问题向您反映，我们公司位于金洲大道与雷高路相交处，处于先导区发展前沿，近两年来发展迅猛，有大批的企业相继建成、投产，在这里上班的企业员工也越来越多，但公交线路却只有 916 路一条且是绕行雷锋镇至三汊矶大桥方向，如要沿金洲大道去保利麓谷林语小区、中联重科、高新区管委会等方向则没有公交车，如乘坐沿金洲大道开行的 A8 县至西站方向大巴车白天 5 元/人次、夜晚 10 元/人次则超出了我们普通员工的经济承受能力，今年 5 月份的时候在潇楚晨报上看到一则消息：“A 市在 6 月底前会新开通多条公交线路”，其中包括新开通一条“12”路线路，由雷高路首末站经金洲大道往保利麓谷林语方向，这令我们欢欣鼓舞，但现在时间已经到 9 月下旬了，盼望中的 12 路公交车却迟迟没有到来，不知将会拖到什么时候，现通过问政窗口向您反映，希望能得到您的关注，在百忙之中要求有关部门能提高办事效率，尽快开通 12 路公交车，在此万分感谢！
答复意见	网友：您好！留言已收悉

- 转交给其他部门

留言主题	咨询 M 市电信携号转网的问题
留言详情	本人想携号转网，由电信转出，但与电信签有靓号协议，2037 年才能到期，如要携号转网，先要解除靓号协议，电信公司要求支付 30%违约金**********.3 近 6000 的违约金。要他们拿依据出来都没有！当时协议上也没有写明违约金的问题！这是明显坑老百姓！
答复意见	网友：您好，您的留言已转入 M 市人民政府门户网站书记市长信箱办理，请耐心等待回复。您也可以在 M 市人民政府门户网站书记市长信箱写信反映您的诉求或查询本条留言的办理进度。

◆ 相似度大于 0.8 的情况

留言主题	严重质疑 M 市经开区违规采石污染环境
留言详情	严重质疑M市经开区违规采石破坏环境我们是M市红星美凯龙建材家具城商户，自进驻红星美凯龙经商一年多来，饱受商城隔壁采石场工地噪音、灰尘的困扰，我们多次向 M 市环保局等相关部门反映，但上级部门互相推诿，事情一直没得到重视和解决。具体表现为：1、工地直接现场将石头破碎，造成灰尘弥漫，噪音环绕，严重污染空气。2、就该问题已经多次向环保部门、市政府反应，但没有具体人员下来核实调查，实有糊弄老百姓嫌疑。政府态度模棱两可，助长了工地项目负责人的歪风邪气，他们变本加厉，加紧开采，给老百姓生活带来严重影响。我们盼星星、盼月亮，本以为这次政府可以帮我们老百姓办大实事，结果盼来的却是这样的答复，这令我们业主极度不满意。中央指出，金山银山不如绿水青山，党中央将环境治理作为头等大事在抓，容不得半点马虎。试想，红星美凯龙商城也算是位于城区中心

	地段，原本就应该严格治理环境，如果因周边环境恶劣，灰尘噪音超标，给商户们带来的隐形损失谁来买单？并且我们严重质疑，市政府、经开区及环保部门的不作为，今天，我们再次通过平台表达我们的诉求：1、 整个工地山头开挖裸露土地，必须用绿网覆盖；2、 工地内两处采石场必须关闭，采石场没有合法手续；3、 请省、市级相关部门重视此事，派人到现场查看并答复为感。如果此事再不得到彻底解决，我们全体商户将继续向省级相关部门反映！
答复意见	网友：您好，留言收悉，现回复如下：1、红星美凯龙建材城旁的施工工地属楚中物流园 e 地块平基土石方工程，位于秋蒲东街。该工程于 2018 年 7 月开工，平基期间通过间接洒水降尘等措施有效地减少了路面扬尘污染现象；2、信访件中所指的采石场是楚中物流园 e 地块平基土石方工程，不是专门的采石场，不属于独立主体，因此未单独办理相关审批手续；3、4 月 17 日上午，我局组织产业环保局、投资集团工程部及 e 地块施工方召开了扬尘、噪声污染整改会，提出如下整改意见：一是施工方在一周内将原渣土运输车辆抛洒在主次街道的渣土及碎石全部清洗到位；二是对裸露地进行覆盖并对长期不使用的裸露地进行绿化；三是安排专门洒水车辆及人员对其区域进行洒水降尘并购买了两台雾炮机对工地内扬尘采取湿润作业方式控制扬尘；四是对碎石机进行助力缓冲来减少噪声的产生，同时合理安排作业时间，晚间禁止高噪声作业，不对周边居民造成影响。4、我局会联同产业环保局等相关部门进一步督促施工方文明施工，确保周边商户、居民生活环境良好。网友：您好，您的留言收悉，已于 2019 年 4 月 12 日转交 M 市经开区办理。谢谢您对我们工作的关心、理解和支持！您还可以拨打 M 市经开区办公室电话:0738-8652618，咨询了解有关具体情况。

4) 综合评价指标

我们定义：

综合评价指标=间隔时间权重*答复意见长度权重*相似度

表 8 综合评价指标分布情况

指标区间	数量
[-1,0)	237
[0,1)	1131
[1,2)	1227
[2,3)	217
[3,4]	4
总数	2816

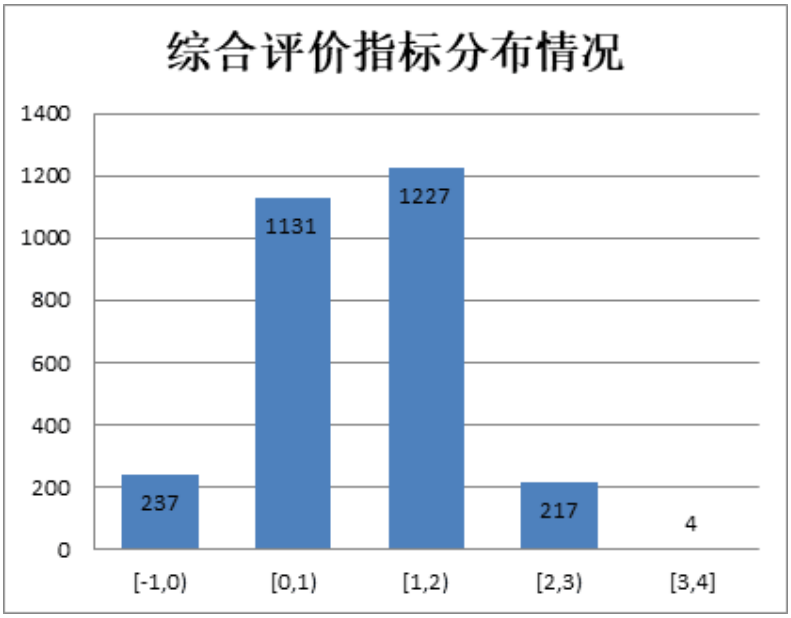


图 15 综合评价指标分布情况

根据图 15 可以很清晰的看到综合评价指标的分布情况。我们可以发现在[-1,0)这个区间的数量是 237，这就说明有 237 个相似度是为负的，可能是因为答复意见答非所问，可能是没有回答，还有可能是转交给相关部门；综合评价指标大于 0 的我们都可以认为是回复质量较好的。

6 结论

本文通过对网络问政平台提供的用户留言数据分别利用 LSTM、CBOW 和 Skip-Gram 等思想分别构造了文本多分类模型和文本相似度计算模型，得到了具有一定价值的结果，实现了对文本留言数据的多分类和对热点问题的提取，并结合平台上相似留言的点赞数和反对数给出热度评价指标，以及结合时间间隔、回复长度和文本相似度来对回复评价的质量打分。这些结果对于政府和相关部门处理用户留言数据具有一定的使用价值，相关部门可以通过多分类模型来实现对数据的快速分类，以及通过文本相似度计算模型快速提取热点话题，从而及时解决群众反映的问题，可以通过评论的综合评价指标来判断问题是否被解决，提高相关部门的执行能力。

但是从我们的分析结果可以看出总体来说效果并不是特别显著，分类模型的 F1 值为 0.85，可能在实际运用的过程中会存在一些误差；在定义热度评价指标和综合评价指标时，由于是人为的定义，可能会与最后的得出结果不相符；在提取热点话题时我们只计算出来了热点话题大于 0.75 的文本，可能在小于 0.75 的范围也存在一部分相似的的话题，并且在提取热点话题里还加入了一些人工的因素，少了一点智能呢个，这也是我们在后期进一步的对文本数据的研究过程中可以继续深入探讨的地方。

7 参考文献

[1] 马存.基于 Word2Vec 的中文短文本聚类算法研究与应用[D].中国科学院大学(中国科学院沈阳计算技术研究所),2018.

[2] 章成志.文本聚类结果描述研究综述[J].现代图书情报技术,2009 年 02 期.

[3] 周练.Word2Vec 的工作原理及应用探究[J].科技情报开发与经济,2015,25(2):145-148.

[4] 姜霖,王东波.采用连续词袋模型(CBOW)的领域术语自动抽取研究[J],2016 年 02 期.

[5] 派神.基于 LSTM 的中文文本多分类实战[J],2019.

[6] 邓澍军,陆光明,夏龙.Deep Learning 实战之 word2vec[J],网易有道,2014

[7] Mikolov T,Chen K,Corrado G,et al.Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[8] Mikolov T,Sutskever I,Chen K,et al.Distributed Representations of Words and Phrases and their Compositionality[J]. 2013, 26:3111-3119.

[9] https://blog.csdn.net/weixin_39672386/article/details/82189923