“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，各类相关文本数据量不断攀升。传统的依靠人工进行留言划分、热点整理、政府反馈的方式面临极大挑战。同时，随着大数据、自然语言处理、分类算法等技术的不断发展，建立基于自然语言处理技术的智慧政务系统已成为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进作用。本文将基于自然语言处理、分类算法技术对收集自互联网公开来源的群众问政留言记录，以及相关部门对部分群众留言的答复意见进行内在的信息挖掘，提取其文本信息以进行深度挖掘和分析。

针对问题 1：本文首先对附件 2 中的非结构化数据进行去重去空，标点删除。然后用 jieba 分词库对数据进行分词，并进行停用词过滤等数据预处理操作。随后，基于 One-hot 算法和 TF-IDF 权重法分别提取留言主题和留言详情的候选特征词。接着，我们将留言主题和详情的候选特征词进行合并，形成词汇-文本矩阵，再通过深度学习算法对文本进行训练和分类，最后我们在多层感知机（MLP）+TF-IDF 分类器上取得了最高的 F1 值（0.91）。

针对问题 2：本文对附件 3 中的留言主题和留言详情进行与问题 1 相同的数据预处理和文本向量化操作。然后我们运用问题 1 中设计的分类算法（MLP + TF-IDF）对附件 3 中的文本进行一级标签的标注。然后，我们以每条留言的点赞数和反对数为指标计算每条数据的热度，并过滤掉热度较低的数据。之后我们用 K-means 聚类算法对每一个一级指标下的数据进行聚类，提取热点事件。最后，我们采用用户参与度（用户的评论数量）、时间跨度（事件的持续时间）和用户关注度（点赞数和反对数之和）三个指标，运用熵值法确定各指标的权重，构建热度评价指标体系，对各热点事件进行指标体系的评估。

针对问题 3：本文首先进行和问题 1 相同的数据预处理和文本向量化操作。然后运用 LDA 计算回复意见的完整性、bm25 算法计算回复与用户留言的相关性，同时通过回复和留言之间的时间间隔，计算回复的及时性。随后运用熵值法确定三个指标的权重，构建回复意见指标体系，并计算每条回复意见的质量。

关键词：TF-IDF；多层感知机；主成分分析；K-means 聚类；熵值法；bm25；LDA

The application of text mining technique in “Intelligent government affairs”

Abstract: Recent years, with the development of We-chat, microblog, and the mayor's mailbox, e-government has become an essential way for the government to understand the will, gathering the intelligence of residents, and giving feedback. Collecting and classifying the category of the message from residents merely by human are becoming more and more difficult as the soaring of the number of related data. Meanwhile, with the development of big data, natural language processing, classifying techniques, building intelligent e-government platforms based on natural language processing techniques is becoming a new trend for improving the efficiency and effectiveness of management and administration for the government. The current research uses natural language processing and classification algorithms to dig out and analyze useful information from comments data, which is written by residents and is available to the public.

Aiming at question 1: firstly, we proposed a data-preprocessing step to remove blank, repetition, punctuation, and stopwords for the data in file 2. Next, we applied the One-hot algorithm and the TF-IDF algorithm to extract keywords from textual contents, and we combined the keywords of topics and details to construct the word matrix. Finally, we applied several deep learning algorithms for the classification task. We reached the highest F1-score of 0.91 using MLP + TF-IDF classifier.

Aiming at question 2: We firstly applied the same data pre-processing and word vectorization steps for the data in file 3. Then we proposed the classifier (MLP + TF-IDF) to classify the data in file 3 according to the first level labels. Then, we calculate the heat of each review using its agree and disagree numbers, and remove these reviews with low heat. Next, we proposed the K-means algorithm to classify the review and extract the events or issues people concern about. Finally, we selected three indicators: degree of participation (The number of reviews), Time span (The duration of an issue), and degree of attention (The number of agrees and disagrees). And we applied the entropy method to calculate the weight of each indicator. After that, we were able to count the score of each data.

Aiming at question 3: firstly, we also applied the data pre-processing and word vectorization steps for the data in file 4. Then we applied the LDA and bm25 algorithm respectively to calculate the degree of integrality and correlation for each feedback. We also calculate the timeliness by calculating the time difference between the messages and their feedbacks. After that, we used the entropy method to calculate the weight of each indicator and figure out the quality of each feedback.

At last, we will give our conclusion based on the results and the fact.

Keywords: TF-IDF; MLP; PCA; K-means; Entropy method; bm25; LDA

目录

1. 挖掘目标

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。因此，利用自然语言处理和文本挖掘的方法对群众留言进行划分和热点整理研究具有重大的意义。

本次研究的挖掘目标是利用互联网公开来源的群众问政留言记录，和相关部门对群众留言的答复意见等非结构化文本数据，对其进行深入挖掘和分析，以达到以下三个目标。

第一，对用户留言数据进行一级标签分类：在对数据进行基本的去除停用词、标点删除等预处理，以及 TF-IDF 权重法赋权后，根据留言主题和内容，运用多层感知机、RNN、CNN 等深度学习分类算法对留言按一级标签进行分类。

第二，运用 k-means 聚类算法对群众留言中的热点问题进行归类：在对用户留言主题和留言详情数据进行处理以及分词操作后，用训练好的分类器对数据进行一级标签分类，同时过滤掉关注度较低的留言数据（点赞数+反对数）。之后对于每个类别分别运用 K-means 聚类算法提取热点问题，最后采用熵值法确定指标权重，构建热度评价指标体系，对热点问题的热度进行排序。

第三，设计指标体系，以判断回复内容的质量：首先，对用户留言主题和留言详情以及有关部门的回复数据进行处理。然后，我们分别运用 lda 算法、bm25 算法计算回复意见的完整性和相关性，并通过回复时间与留言时间的差值，计算回复的及时性。最后运用熵值法对各指标进行赋权，对回复数据进行打分。

2. 分析方法与过程

2.1 问题 1 分析方法与过程

问题 1 的处理过程包含以下四部分：数据预处理、文本向量化、文本分类以及输出结果，其流程图如图 1 所示：

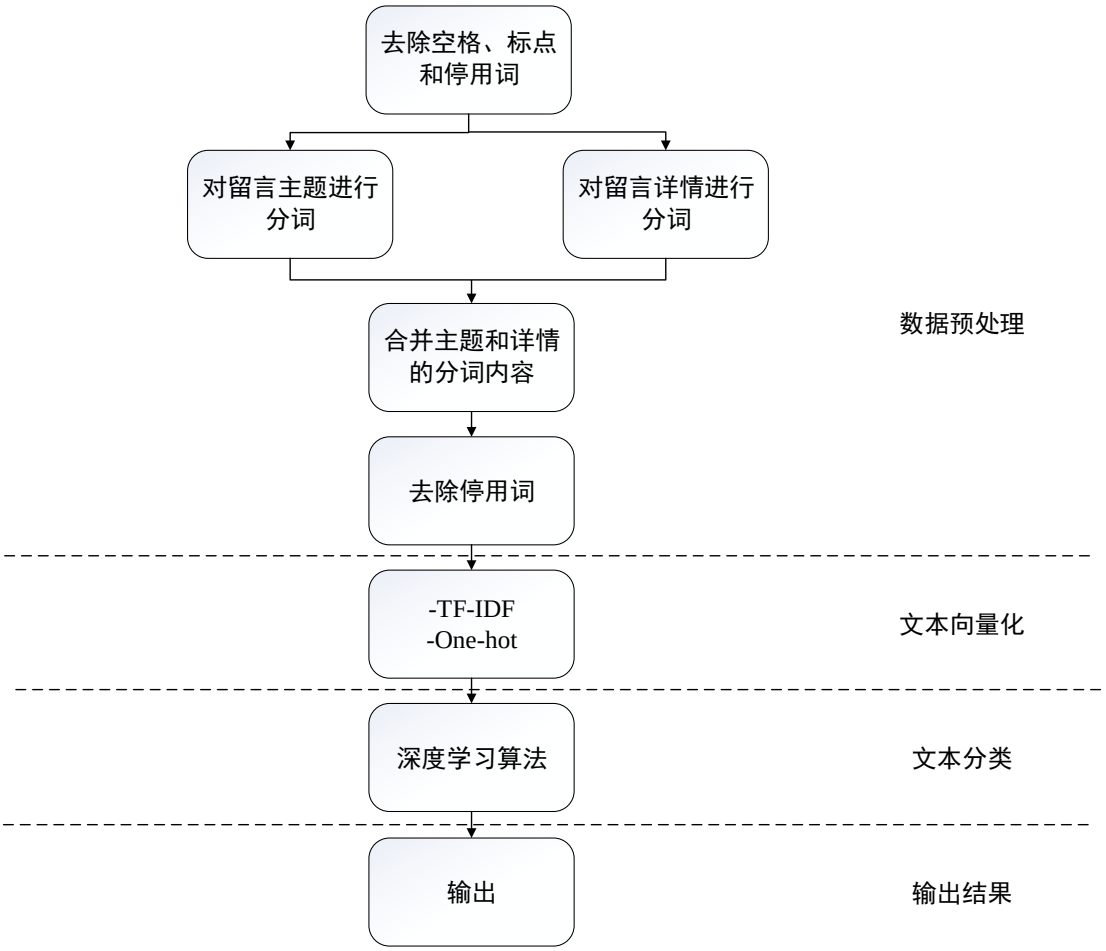


图 1：问题 1 分析流程图

2.1.1 数据预处理

问题 1 的数据预处理包含三个过程：（1）去除标点、空格；（2）对文本数据进行中文分词；（3）停用词删除。

2.1.2 去除标点、空格

在题目给出的数据中，出现了大量空格、标点。这些内容会对后续的候选词提取等步骤造成干扰。因此我们首先对附件 2 中所有的留言主题和留言详情文本数据进行空格、标点和停用词的删除。

2.1.3 对文本数据进行中文分词

在去除文本数据的标点和空格后，我们对留言主题和详情进行中文分词。分词采用 python 中的中文分词包 jieba。jieba 采用了基于前缀词典实现的词图扫描，实现句子中汉字可能成词的所有情况的有向无环图，并运用动态规划法查找最大概率路径，从而保证了其中文分词的效果。

2.1.4 停用词删除

停用词（停用词指在信息检索中，为节省存储空间或提高搜索效率，在处理自然语言数据时自动过滤掉的字或词）等。停用词在几乎所有的文本内容中都普遍存在，因此它们对于文本的分类并无作用，甚至可能产生噪音，影响文本分类效果。所以我们对分词后的文本数据进行停用词的删除，停用词的词库来自互联网公开数据。

2.1.5 文本向量化

所谓文本向量化，就是将文本表示成一系列能够表达文本语义的向量，在本次实验中，我们选用 One-hot 编码和 TF-IDF 两种算法分别对留言主题和留言详情信息进行文本向量化操作。

2.1.6 One-hot 编码

所谓 one-hot 词向量，就是用一个很长的向量来表示一个词。一般 One-hot 编码会根据自己的训练文档来构建一个词汇表，再对单词进行 one-hot 编码。比如说在我们的所有的训练集中存在 10000 个互不相同的单词，那么就可以利用这 10000 个单词构成词汇表。我们对这 10000 个单词进行 one-hot 编码时，每个单词的向量表示都是 10000 维，且其中只有 1 和 0 的表示方式。假设一个词“我”出现在第 5 个位置，那么“我”这个单词的 10000 维的向量表示中，只有第 5 个位置为 1，其余的都为 0，即[0,0,0,0,1,0,0，.....]的形式。

2.1.7 2.1.2.2 TF-IDF 算法

TF-IDF 算法是一种从文本数据中提取特征的文本挖掘算法。特征词的 TF-IDF 值计算步骤如下：

第一步：计算词频，即 TF 权重（Term Frequency），他描述某个词在一个文本中出现的频率。其计算公式为：

$$\text{词频(TF)} = \frac{\text{某词在某文本中出现的次数}}{\text{该文本的总词数}} \qquad (1)$$

第二步：计算逆文档频率，即 IDF（Inverse Document Frequency），他描述一个词在整个语料库中出现的频繁程度，IDF 值越大，则该特征词在文本中的分布越集中，从而其区分文本内容的能力越强。IDF 的计算公式为：

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库中的文本总数}}{\text{包含该词的文本数} + 1}\right) \qquad (2)$$

第三步：计算 TF-IDF 值（Term Frequency Inverse Document Frequency），计算公式为：

$$\text{TF} - \text{IDF} = \text{词频（TF）} \times \text{逆文档频率（IDF）}$$

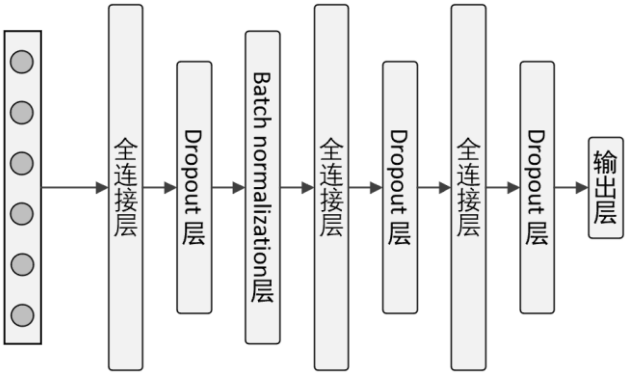
对于每个特征词来说，TF-IDF 值越大，该词的重要程度越高。在本次实验中，我们采用 TF-IDF 算法分别对留言主题和留言详情数据的特征词进行提取，并将其结合构造词汇-文本矩阵。

2.1.8 深度学习文本分类

留言分类是电子政务中的一个重要的问题，但是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。近几年， 深度学习在自然语言处理领域取得了巨大的成功，因此本文利用多种深度学习算法对留言进行分类。

(一) 多层感知机分类

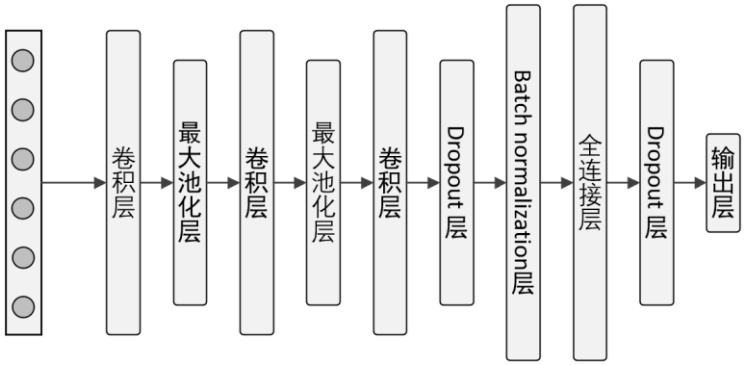
多层感知机（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network），多层感知机包括输入层，输出层，以及隐藏层，其中每一层都可以包括多个神经元。本文将向量化的文本作为输入，设置了 3 个全连接层，三个 dropout 层以及一个 batch _normalization 层并将输出层的神经元个数设为 7。具体的模型如下：



我们将三层全连接层神经元个数分别设置为 512，256，128 个，并统一使用$relu$作为激活函数。在 dropout 层，我们将 drop 的比例都设置为 0.5。在输出层，我们使用$softmax$作为输出函数。

(二) 卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）是一类包含卷积计算且具有深度结构的前馈神经网络。CNN 最大的特点是局部感知和权值共享，这样的特点使其在文本挖掘任务中取得了良好的表现。CNN 通常会包含 4 个部分：卷积层，池化层和全连接层。具体的模型如下：

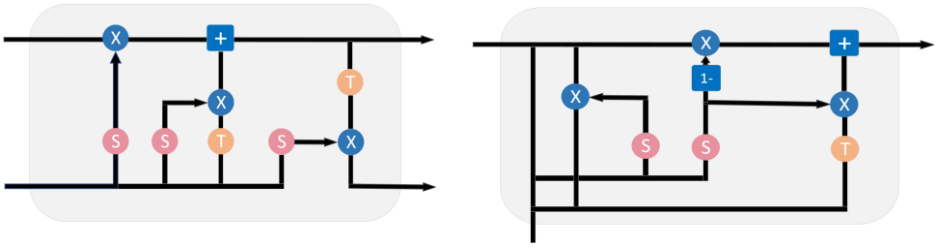


在三个卷积层中，我们都使用$relu$作为激活函数并以“same”的方式进行填充，神经元的个数分别为 256，128，64 个且卷积的步长为 3。在池化层，我们都选择最大池化并以“same”的方式进行填充。在 dropout 层，我们将 drop 的比例都设置为 0.1。在最后的输出层，设置 7 个神经元且以$softmax$作为输出函数。

(三) 循环神经网络

循环神经网络（Recurrent Neural Network, RNN）是一类以序列数据为输入，在序列的演进方向进行递归且所有节点按链式连接的递归神经网络。RNN 中的长短记忆机制（Long

Short-Term Memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）都是非常重要的算法，这两个模型的结构图如下：



(a) LSTM　　(b)GRU

2.2　问题 2 分析方法与过程

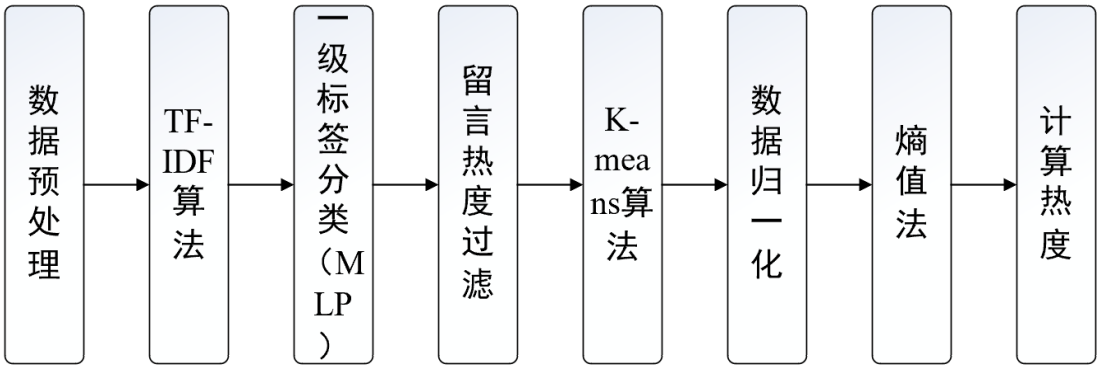


图 2：问题 2 分析流程图

2.2.1 数据预处理

在解决问题 2 时，我们对附件 3 中的数据进行与问题 1 相同的数据预处理操作，并运用 jieba 分词库对留言的主题和文本进行分词。随后我们用 TF-IDF 算法提取特征词，形成词袋，并分别构造留言主题和留言详情的词汇-文本矩阵。然后我们用主成分分析法（PCA）将训练集和测试集的文本维度一致。

2.2.2 一级标签标注

由于在问题一中，我们设计的一级标签分类模型具有良好的分类效果，因此我们用附件 2 中的数据作为训练集对分类模型进行训练。然后我们用分类模型对附件 3 中的数据进行一级标签的标注（附件 2 中总共有 7 类一级标签）。在标注完毕后，我们随机选取 100 条留言数据进行人工检验，以判断分类的准确性。

2.2.3 留言热度过滤

由于大部分的留言热度较低，会对数据产生大量的噪声，所以在每一类的数据中设置一个阈值，如果热度低于这个阈值就会被去除。

2.2.4 K-means 聚类算法

对于问题 2，我们采用 K-means 聚类算法对数据进行分类。K-means 聚类算法是一种迭代求解的无监督学习算法，它能在无标签的情况下对数据进行分类。K-means 聚类算法的步骤如下：

1. 从 X 个元素中随机选取 K 个元素（K ≤ X），作为 K 个簇的中心。
2. 依次计算其他元素到 K 个簇中心的相异度（用欧式距离作为衡量指标），并将这些元素归为与其相异度最低的簇。
3. 根据步骤 2 的聚类结果，对每个簇中所有元素的各自维度的算术平均数进行计算，作为每个簇的新的中心点。
4. 重复步骤 2，对所有元素按照新的簇中心重新聚类。
5. 重复步骤 4，直到聚类结果不再改变。
6. 输出聚类结果。

2.2.5 热度评价

2.2.5.1 热度评价指标

对于第二次聚类结果中的每个热点事件，我们选取了三个热度评价指标，分别为用户参与度、时间集中度和用户关注度如表 1 所示：

表 1：热点事件热度评价指标

评价指标热度	计算公式	指标含义
用户参与度	热点事件中用户的留言数量	反应热点事件所涉及的用户规模
时间跨度	Max(时间) – Min(时间)	反应热点事件的时间跨度
用户关注度	点赞数+反对数	反应用户对热点事件的关注程度

2.2.5.2 指标赋权

在进行指标赋权之前，我们需要对各指标的得分进行归一化，归一化公式为：

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

完成归一化操作后，我们采用熵值法作为热度指标的赋权方法。熵值法是一种客观赋权方法，它通过判断指标的离散程度，确定指标对综合评价的影响程度。指标的离散程度越大，则其对综合评价的影响就越大。其具体步骤为：

1.选取 n 个指标

2.对 n 个指标进行标准化处理

3.计算第 i 项指标的熵值（离散程度）

4.通过各指标的熵值比例确定各指标的权值
5.计算综合得分，得分越高，热点事件的热度越高

2.3 问题 3 分析方法与过程

问题 3 的处理过程包括五个部分：数据预处理、文本向量化、指标得分计算、指标赋权以及输出回复得分，其分析流程图如图 3 所示：

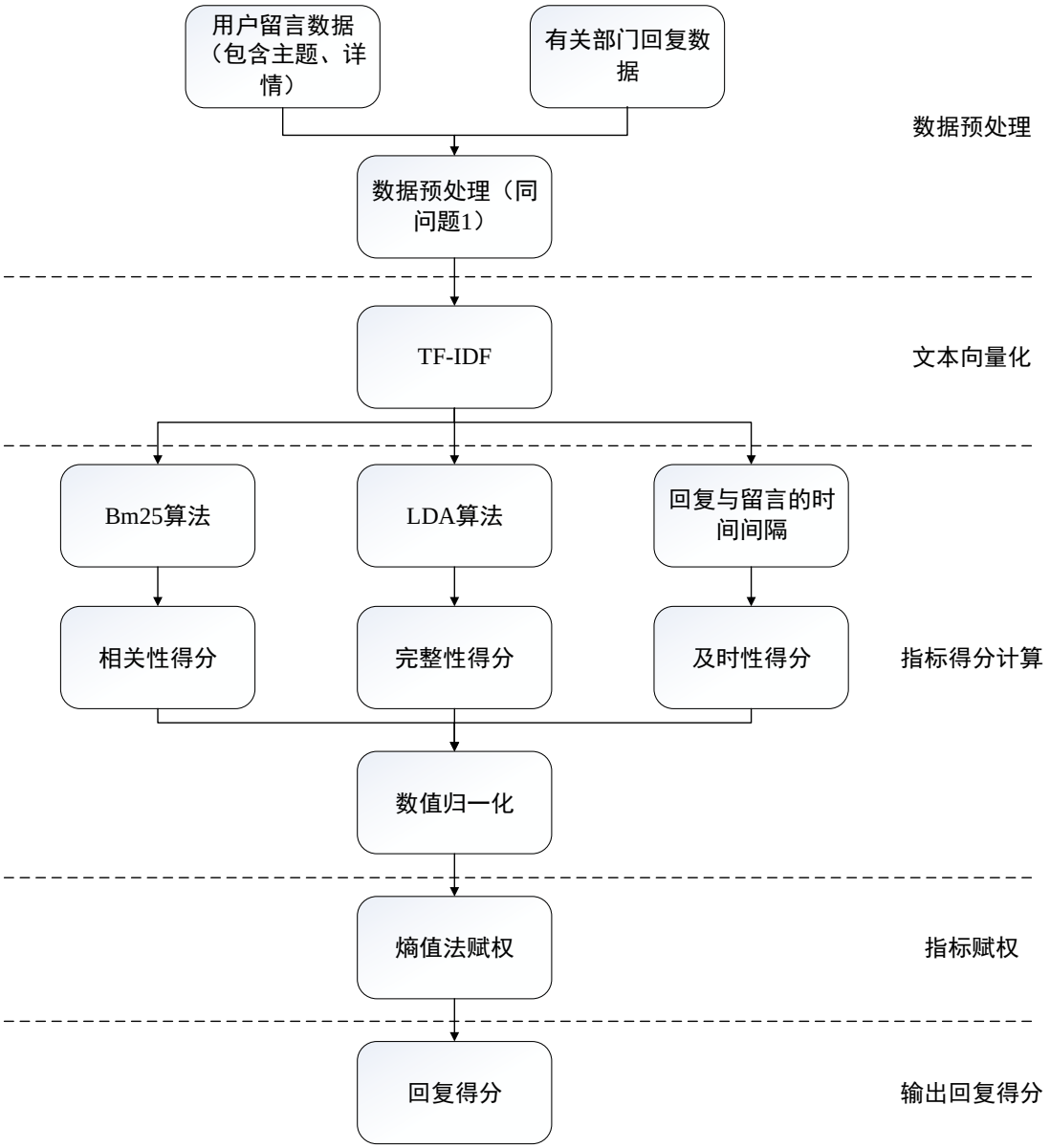


图 3：问题 3 分析流程图

2.3.1 数据预处理

对于用户留言的主题、留言详情以及有关部门的回复文本数据，我们进行与问题一相同的去空、去除标点和停用词以及分词的数据预处理操作。随后我们分别计算回复意见的及时性、相关性。

2.3.2 及时性

对于用户的留言问题，相关部门的留言应具有时效性，如果拖延太久，那么回复给用户带来的帮助很可能就不会太大。因此，为了保证回复的有用性，相关部门的回复时间间隔不宜太长。通过附件 4 中的数据可知，用户的留言时间和相关回复时间都可获得，因此我们采用两者的时间间隔反应回复意见的及时性，其计算公式为：

$$\begin{cases} 1-\Delta t/90, & \Delta t \le 90 \\ 0, & \Delta t \ge 90 \end{cases} \quad (3)$$

根据《信访条例》，信访事项的办理时限为：

（一）对群众信访事项，承办单位在受理之日起 60 日内办结；

（二）情况复杂的，经本级行政机关负责人批准，承办单位可以适当延长办理期限，但延长不得超过 30 日，并须告知信访知人延期理由

因此，我们将最长回复间隔设为 90 天（申请延期 30 天，总共 90 天），则及时性随回复时间的增加而减少；如果回复时间超过 90 天，则回复视为不及时，得分为 0。

2.3.3 相关性

回复的相关性反应回复内容与用户留言问题的相关程度，一般情况下，留言的相关性越高，留言的质量就会越高。在本次实验中，我们采用 bm25 算法来计算回复意见与用户留言的相关性。Bm25 算法是一种用来评价文本数据之间相关性的算法，它是一种基于概率检索模型提出的算法，其具体步骤如下：

1. 对文本进行分词，得到单词，单词的得分由以下三部分构成：

（1）计算单词与另一文本的相关性。

（2）计算单词与该文本的相关性。

（3）计算每个单词的权重。

2. 计算该文本中的每个单词的得分之和，即为回复的 bm25 值，当 bm25 的值大于 0 时，则说明回复与留言具有相关性，且 bm25 值越高，两个文本的相关性就越高，说明回复针对性高，能很好地针对用户留言给出解释。而当 bm25 小于 0 时，则说明回复答非所问，不具有相关性。

2.3.4 完整性

完整性用于描述答复意见是否对群众反映的问题进行了系统全面的回答，是否涵盖了所有群众反映问题的主题。因此，我们在这使用 LDA 主题模型对完整性进行描述。

LDA 主题模型是自然语言处理领域中熟知的模型，本次实验中，通过 LDA 主题模型对留言详情和答复意见之间进行一个主题对比，留言详情和答复意见之间相同主题数目越多说明答复意见越完整，由于 LDA 主题模型需要预先设定主题数，我们通过问题 1 中的类别数目来确定主题数，确定主题数为 15。LDA 的具体步骤主要如下：

1. 将留言详情和答复意见进行分词、去停用词，用做训练的语料。
2. 进行 LDA 模型训练，我们可以得到留言详情和答复意见属于各个主题的概率。
3. 通过计算留言详情和答复意见之间的主题概率乘积之和来描述完整性。

2.3.5 指标赋权

对于各个指标，我们再次进行归一化操作，并采用熵值法对三个指标体系进行权重的分配。最后，根据各指标的权重计算每条回复意见的得分，进而评估回复意见的质量。

3. 实验结果与分析

3.1 问题 1 实验结果与分析

问题实验结果如表 2 所示。在众多算法中，对于 RNN、CNN 算法，One-hot 算法和 TF-IDF 算法均未达到预期效果，因此我们用 keras 直接嵌入文本来进行实验。由表二可知，多层感知机算法（MLP）的分类效果显著好于其他深度学习算法。同时 TF-IDF 的文本向量化效果略微好于 One-hot 编码算法（因此我们后续的实验中，文本向量化方法都采用 TF-IDF 算法）。

同时，在该次实验中，我们尝试使用主成分分析法（PCA）对文本-词汇矩阵进行降维，但分类结果并没有改善。

表 2：深度学习分类结果表

方法	F1 值
MLP + One-hot	0.90
MLP + TF-IDF	**0.91**
CNN + keras 嵌入	0.83
RNN(双循环 LSTM) + keras 嵌入	0.74
RNN(双循环 GRU) + keras 嵌入	0.67

为了进一步分析各类留言问题的分类效果，我们采用效果最好的 MLP + TF-IDF 分类算法对七个一级分类问题中的每一类问题的分类效果进行求解，结果如表 3 所示。由表 3 可知，MLP + TF-IDF 对各分类问题的分类效果相对比较均衡，都维持在 0.9 左右的 F1 值，其中教育文体类的分类效果最好。由此可知，对于用户留言的分类问题，运用有监督学习对部分标注数据进行训练，能得到一个性能优异的分类器，从而弥补了人工标注数据量大、时间消耗过长的问题。因此，有关部门可以考虑使用此分类算法解决用户留言的自动分类问题。

表 3：MLP + TF-IDF 分类算法对各类留言的分类结果表

一级分类	F1 值
城乡建设	0.89
卫生计生	0.88
教育文体	**0.95**
环境保护	0.91
交通运输	0.89

劳动和社会保障	0.90
商贸旅游	0.89

3.2 问题 2 实验结果与分析

3.2.1 一级标签标注

由问题 1 的实验结果可知，MLP + TF-IDF 分类模型能十分准确的按一级标签对用户留言的主题和留言详情进行分类，因此，对于问题 2 所涉及的附件 3 中的数据，我们运用该模型对其进行一级标签的标注，数据的预处理过程与问题 1 相同。分类结束后，我们通过对部分数据人工检验证实：分类的结果良好，绝大多数数据都已正确地标记。图 4，5 为其中部分数据的标记结果。

	A	B	C	D	E	F	G	H	I	J	K	L
1	留言编号	留言用户	留言主题	留言主题（去除停用词）	留言时间	留言详情	留言详情	反对数	点赞数	date	datevalue	label
2	188006	A0001029	A3区一米阳光婚纱	A3区一米阳光婚纱艺术	2019/2/28	座落在A市A3	A市A3区联	0	0	2019/2/28	0.666667	城乡建设
3	188031	A0004006	反映A7县春华镇金	反映A7县春华镇金鼎村	2019/7/19	本人系春华镇	系春华镇金	0	1	2019/7/19	0.815873	城乡建设
4	188039	A0008137	A2区黄兴路步行街	A2区黄兴路步行街古道	2019/8/19	靠近黄兴路步	靠近黄兴路	0	1	2019/8/19	0.848677	城乡建设
5	188059	A0002857	A市A3区中海国际	A市A3区中海社区三期四	2019/11/2	A市A3区中海	A市A3区中	0	0	########	0.949206	城乡建设
6	188073	A909164	A3区麓泉社区单方	A3区麓泉社区单方面改	2019/3/11	作为麓泉社区	麓泉社区麓	0	0	2019/3/11	0.678307	城乡建设
7	188074	A909092	A2区富绿新村房产	A2区富绿新村房产性质	2019/1/31	“二高一部”发	“二高一部	0	0	2019/1/31	0.637037	城乡建设
8	188119	A0003502	对A市地铁违规用工	A市地铁违规用工质疑	2019/5/27	我是一名在A市	一名A市地	0	0	2019/5/27	0.759788	城乡建设
9	188170	A8801132	A市6路公交车随意	A市路公交车随意变道通	2019/12/2	12月21日下午	12月21日	0	0	########	0.982011	城乡建设
10	188249	A0008408	A3区保利麓谷林语	A3区保利麓谷林语桐梓坡	2019/9/17	保利麓谷林语	保利麓谷林	0	0	2019/9/17	0.879365	城乡建设
11	188251	A0001309	A7县特立路与东四	A7县路东四路口晚高峰	2019/10/1	近来下午晚高	下午晚高峰	0	0	########	0.913228	城乡建设
12	188260	A0005348	A3区青青家园小区	A3区青青家园小区乐果	2019/5/31	还我宁静我要	宁静我要	0	0	2019/5/31	0.764021	城乡建设
13	188396	A0004758	关于拆除聚美龙楚	拆除聚美龙楚西地省商	2019/4/15	桐梓坡589号	桐梓坡589	2	1	2019/4/15	0.715344	城乡建设
14	188399	A0009793	A市利保壹号公馆	A市利保壹号公馆夜间噪	2019/7/3	您好我想举报	您好想举报	0	0	2019/7/3	0.798942	城乡建设
15	188409	A0003274	A市地铁3号线星沙	A市地铁号线星沙大道站	2019/6/19	尊敬的领导您	尊敬您好A	0	4	2019/6/19	0.784127	城乡建设
16	188414	A0009684	A4区北辰小区非法	A4区北辰小区非法住改	2019/8/1	您好我是北辰	您好北辰C	0	0	2019/8/1	0.82963	城乡建设
17	188451	A0001300	A7县春华镇石塘铺	A7县春华镇石塘铺村党	2019/4/11	我是春华镇一	春华镇一名	0	2	2019/4/11	0.711111	城乡建设
18	188475	A0005581	A6区乾源国际广场	A6区乾源停车场违章乱	2019/12/3	A市A6区月亮	A市A6区月	0	0	2019/12/3	0.960847	城乡建设
19	188535	A0006177	A7县时代星城4幢	A7县星城幢非法经营家	2019/6/13	尊敬的各位领	尊敬A7县	0	0	2019/6/13	0.777778	城乡建设
20	188546	A0006817	A2区佳兆业水新都	A2区佳兆业水新小区垃	2019/1/23	敬爱的领导你	敬爱你好	0	0	2019/1/23	0.628571	城乡建设
21	188592	A0003945	A市长房云时代小	A市长房云小区三期后面	2019/6/18	长房云时代小	长房云小区	0	0	2019/6/18	0.783069	城乡建设
22	188774	A0004879	A2区政府东门至万	A2区政府东门万芙路段	2019/6/18	多年来A2区迎	A2区迎新路	0	0	2019/6/18	0.783069	城乡建设
23	188780	A0009475	请依法解决A7县黄	请依法A7县黄花镇梁坪	2019/10/2	尊敬的县领导	尊敬县您好	0	0	########	0.914286	城乡建设
24	188801	A909180	投诉滨河苑针对广	投诉滨河苑广铁职工购房	2019/8/1	尊敬的张市长	尊敬张市长	0	0	2019/8/1	0.82963	城乡建设
25	188809	A909139	A市万家丽南路丽发	A市万家丽南路丽发新城	2019/11/1	A市万家丽南路	A市万家丽	0	1	########	0.946032	城乡建设

图 4：城乡建设标记示例

	A	B	C	D	E	F	G	H	I	J	K	L
1	留言编号	留言用户	留言主题	留言主题	留言时间	留言详情	留言详情（去除停用	反对数	点赞数	date	datevalue	label
2	190802	A0007263	A市丽发小	A市丽发小	2019/11/2	发同投资有限公司在未	发同投资未经同意情	0	0	########	0.952381	环境保护
3	191154	A0004410	举报A市A3	举报A市A3	2019/7/13	尊敬的胡书记您好作为	尊敬胡书记您好一名	0	16	2019/7/13	0.809524	环境保护
4	191327	A0007369	A3区中海国	A3区中海社	2019/11/1	我是中海国际社区四期	中海社区四期居民几	0	0	########	0.944974	环境保护
5	191440	A0006374	A4区兴汉门	A4区兴汉门	2019/7/15	住宅一楼餐馆厨房设在	住宅一楼餐馆厨房住	0	0	2019/7/15	0.81164	环境保护
6	192214	A0004287	A6区月亮岛	A6区月亮岛	2019/4/23	本人居住月亮岛街道楚	居住月亮岛街道楚江	0	0	2019/4/23	0.72381	环境保护
7	192427	A0009562	A3区雨敞坪	A3区雨敞坪	2019/6/18	本人家住A市A3区雨敞	家住A市A3区雨敞坪	0	0	2019/6/18	0.783069	环境保护
8	192790	A0004537	A3区凯特梅	A3区凯特梅	2019/5/16	我是A市A3区凯特梅溪	A市A3区凯特梅溪紫	0	0	2019/5/16	0.748148	环境保护
9	193385	A909144	请A市坚决	请A市坚决	2019/10/3	从全国各地看校园贷已	校园贷遍死许多大学	0	6	########	0.924868	环境保护
10	193841	A0004874	A6区新华联	A6区新华联	2019/5/31	胡书记你好位于A6区银	胡书记你好A6区银星	0	0	2019/5/31	0.764021	环境保护
11	194133	A0003215	A9市镇头镇	A9市镇头镇	2019/11/1	我们是西地省A9市镇头	西地省A9市镇头镇金	0	0	########	0.938624	环境保护
12	195252	A0001095	A1区汇一城	A1区汇一城	2019/10/2	最近在我们A1区汇一城	最近A1区汇一城小区	0	0	########	0.92381	环境保护
13	195701	A0004832	A3区新民小	A3区小区	2019/5/14	A市A3区新民小区系老	A市A3区小区系老旧	0	0	2019/5/14	0.746032	环境保护
14	196268	A0008650	A市黄花机场	A市黄花机	2019/1/17	想到机场附近买房但考	想到机场附近买房考	0	0	2019/1/17	0.622222	环境保护
15	196530	A000674	春节燃放烟花	春节燃放烟	2019/1/31	听说除夕夜和大年初一	听说除夕夜大年初一	0	0	2019/1/31	0.637037	环境保护
16	196659	A0004410	举报A市A3	举报A市A3	2019/8/7 1	被举报企业A市A3区柏	举报A市A3区柏家塘	0	2	2019/8/7	0.835979	环境保护
17	196797	A0001103	A市A5区林	A市A5区林	2019/6/26	尊敬的领导A市A5区树	尊敬A市A5区树木岭	0	0	2019/6/26	0.791534	环境保护
18	197619	A909107	A7县松雅湖	A7县松雅湖	2019/10/2	每到晚上10点左右松雅	每到晚上10点左右松	0	5	########	0.915344	环境保护
19	197630	A0007296	A2区瑞都路	A2区瑞商街	2019/10/2	深夜经常歌声不断严重	深夜歌声严重噪音扰	0	0	########	0.92381	环境保护
20	198003	A0001132	A3区天顶街	A3区天顶街	2019/5/15	A3区天顶街道尖山安置	A3区天顶街道尖山安	0	0	2019/5/15	0.74709	环境保护
21	198642	A0005485	A8县市天下	A8县市天下	2019/7/25	A市A8县市经开区的天	A市A8县市经开区天	0	0	2019/7/25	0.822222	环境保护
22	198644	A0009665	请将A4区社	请A4区福湾	2019/8/10	尊敬领导你们好A市捷	尊敬你们好A市捷运	0	0	2019/8/10	0.839153	环境保护
23	198673	A0007722	A7县星沙凯	A7县星沙凯	2019/10/6	在A7县碧桂园翘楚棠生	A7县碧桂园翘楚棠生	0	1	2019/10/6	0.899471	环境保护
24	199379	A0009224	A2区丽发新	A2区丽发新	2019/11/2	A市A2区丽发新城小区	A市A2区丽发新城小	0	0	########	0.952381	环境保护
25	199386	A0005193	A2区南托街	A2区南托街	2019/6/21	领导您好万不得已上书	您好万不得已上书维	0	0	2019/6/21	0.786243	环境保护
26	199605	A0009121	A4区楚雅路	A4区楚雅路	2019/7/15	A4区楚雅路办事处新楚	A4区楚雅路办事处新	0	0	2019/7/15	0.81164	环境保护

图 5：环境标记示例

随后进行热度过滤后进行文本聚类，结果如图 6

	A	B	C	D	E	F	G	H	I
1	review	review_id	hot		cate		cate_num	time	一级标签
2	尊敬胡书记您好A4区p2p58车贷非法经营四年受害人下	220711	2344	1523	0		3	2019/2/25	城乡建设
3	胡市长您好西地省展星投资58车贷httpsbaidu亿2018年	217032	2344		0		3	2019/3/1	城乡建设
4	尊敬胡书记您好A4区p2p58车贷非法经营四年受害人下	220711	2344		1		3	2019/2/21	城乡建设
5	胡书记您好58车贷案发引发受害人举报投诉引起市公布	194343	2344		0		3	2019/9/5	城乡建设
6	A市A2区暮云街道丽发新城2014年首次交房已入住几万	193091	242		0		1	2019/8/19	城乡建设
7	您好近日看到渝长厦高铁红线征地走向北三环紧绿地滩	263672	669		0		1	2019/6/19	城乡建设
8	联名信坚决A市润一城三润城润紫郡润长郡润美郡润星城	262052	78		0		1	2019/3/26	城乡建设
9	前消息传出克服银盆岭桥东西两侧四方坪路段堵点将三	226723	66		0		1	2019/9/15	城乡建设
10	尊敬A市委书记听说A9市高铁站很久下来许多人关心渺	203187	63		1		1	2019/8/1	城乡建设
11	您好长房云栋18年10月底交房交房时听闻栋楼板开裂以	281898	60		1		1	2019/2/25	城乡建设
12	尊敬胡书记您好区政府街道办电力公司回复无非两条建	272089	57		2		1	2019/4/9	城乡建设
13	A市经开区东六线以西泉塘昌商业中心以南 新蕾品阁居	239670	94		2		3	2019/1/11	城乡建设
14	A市经开区东六线以西泉塘昌商业中心以南 新蕾品阁居	256358	94		2		3	2019/1/2	城乡建设
15	A市经开区东六线以西泉塘昌商业中心以南 新蕾品阁居	233542	94		2		3	2019/1/2	城乡建设
16	请问A市经开区东四线以西新安路以南闲置土地	261625	37		5		2	2019/2/11	城乡建设
17	请问A市经开区东四线以西新安路以南闲置土地	275990	37		6		2	2019/2/2	城乡建设
18	尊敬A市暮云街道丽发新城一名最近遇到意见烦心事情	208285	35		9		2	2019/12/15	环境保护
19	投诉A市暮云街道丽发新城附近搅拌站水泥厂噪音严重扰	261072	35		9		2	2019/11/23	环境保护
20	局长　　你好A5区东路魅力城小区一楼开几家夜宵快餐	272122	18		1		4	2019/8/1	环境保护
21	局长　　你好A5区东路魅力城小区一楼开几家夜宵快餐	360108	18		1		4	2019/1/3	环境保护
22	局长　　你好A5区东路魅力城小区一楼开几家夜宵快餐	284147	18		1		4	2019/11/13	环境保护
23	局长　　你好A5区东路魅力城小区一楼开几家夜宵快餐	360107	18		1		4	2019/7/21	环境保护
24	京港澳高速城区g41996年西地省建成高速公路原来远郊	284571	80		8		1	2019/1/10	交通运输
25	尊敬A市南横线A市地区东西大通道南横线A9市段浏A市	257376	39		6		1	2019/5/10	交通运输
26	书记先生您好梅溪湖金毛湾一名当初金毛湾承金毛建果	223297	1767		0		1	2019/4/11	教育文体
27	你好基层工作者请问包支付app压给基层正当竞争百姓	200667	78		1		1	2019/1/16	劳动和社

图 5：热度过滤后数据示例

最后利用熵值发分别给归一化后的用户参与度，时间跨度和用户关注度赋予权重以求出最终得分。

表 4：各指标权重表

指标	权重
用户参与度	0.47
时间跨度	0.35
用户关注度	0.18

	B	C	D	E	F
1	问题id	热度指数	时间范围	地点/人群	问题描述
2	1	0.6365135	2019/2/21至2019/9/5	A4区p2p58车贷案受害人	A4区受害人举报p2p58车贷案
3	2	0.5311107	2017/6/8至2019/11/22	A市经济学院学生	强制学术去实习
4	3	0.3528236	2019/4/11至2019/4/11	梅溪湖金毛湾楼盘	学区房问题
5	4	0.2603772	2019/1/3至2019/11/13	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
6	5	0.1115338	2019/1/2至2019/1/11	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

图 6：热点问题表

	A	B	C	D	E	F	G	H
1	问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
2	1	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	切盼望有案情消息总是失望，四处诉	0	821
3	1	217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	卡纳和小股东、苏纳弟弟苏吕是挂名	0	790
4	1	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	市A4区经侦并没有跟进市领导的留言	0	733
5	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	多难过！虽说不是强制性的，但不去	0	0
6	2	360111	A1204455	A市经济学院组织学生外出打工合理吗?	2019-11-05 10:31:38	十个小时以上，（晚班时间是20: 3	1	0
7	2	360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	求学生必须去学校安排的几个点实习	0	0
8	2	360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	们都不知道！学校很小但是这几年只	3	0
9	2	360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	公司签了合同，并且公司也要和我们	9	0
10	3	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	月31日，A市教育局暂未将金毛湾楼	5	1762
11	4	272122	A909113	路魅力之城小区一楼的夜宵摊严重污染附近的空气，	2019/08/01 16:20:02	觉得要维护社会和谐稳定，合法维权	0	6
12	4	284147	A909113	劳动东路魅力之城小区一楼的夜宵摊严重污染附近的	2019/07/21 10:29:36	觉得要维护社会和谐稳定，合法维权	0	3
13	4	360107	A0283523	劳动东路魅力之城小区一楼的夜宵摊严重污染附近的	2019-07-21 10:29:36	觉得要维护社会和谐稳定，合法维权	3	0
14	4	360108	A0283523	路魅力之城小区一楼的夜宵摊严重污染附近的空气，	2019-08-01 16:20:02	觉得要维护社会和谐稳定，合法维权	6	0
15	5	239670	A00080329	A市经开区东六线以西泉塘昌和商业中心以南的有关	2019/1/11 15:46:04	地省达源置业有限公司厂房，有什么	0	41
16	5	256358	A00080329	A市经开区东六线以西泉塘昌和商业中心以南的有关	2019/1/2 20:27:07	恒天九五重工原厂房和闲置的西地省	0	29
17	5	233542	A00080329	A市经开区东六线以西泉塘昌和商业中心以南的有关	2019/1/2 20:27:26	恒天九五重工原厂房和闲置的西地省	0	24

图 7：热点问题留言明细表

在图 6 中展示了我们这次从大量用户评论中挖掘到的热点问题。在图 7 中对这 5 个热点问题的留言进行了详细的展示，可以看出这个 5 个热点问题要么具有很高的用户参与度（点赞数和反对数）要么具有较高的用户关注度（多人对该问题进行发言）。这说明我们的热点挖掘方法是有效的。

3.3 问题 3 实验结果与分析

3.3.1 回复及时性分析

有关部门的回复及时性（为了更好的展示结果，该得分已进行归一化处理）分布如图 8 所示：由图显示，回复及时性的得分大致上呈正态分布趋势，从中我们能够得出以下结论：

（1）有关部门对绝大多数的用户留言能做到在规定时间内进行回复（90 天最长回复期限内），未能及时回复（90 天内未回复，及时性得分为 0）的留言内容不到总留言的 1%(84 条回复)。说明对于用户的留言信息，有关部门并没有忽视。

（2）虽然对于绝大多数的用户留言内容，有关部门能在规定时间内进行回复，但仍有部分回复及时性得分并不高：由图我们可以发现，及时性得分超过 0.9 分（6 天内回复）的回复仅有 1258 条，仅占所有回复的 44.7%左右（总共 2817 条数据），及时性得分超过 0.5 分（30 天内回复）的回复有 2581 条，占总数的 91.6%。说明仍有 10%左右的留言，有关部门未能在一个月内进行回复，这会大大降低回复的时效性，从而降低回复的有用性，我们认为可能的原因有：1.有关部门仅仅将回复用户留言作为工作任务，并未认真对待；2.部分用户留言所涉及的问题比较复杂，有关部门需要花一定的时间进行调查走访、解决问题，随后才能进行回复。

随着微博、信访等反应问题的渠道愈发宽广，功能愈发完善，用户与有关部门之间信息的双向交互变得更加方便、快捷，回复的及时性也变得越来越重要。而就实验结果来看，虽然对于绝大部分的留言数据，有关部门能在相对合理的时间内进行回复，但是对于部分留言，有关部门的回复还相对不够及时，虽然这部分数据占比较小，但可能会导致少数比较重要的问题被忽视，影响政府施政效率。为了解决此问题，有关部门应该给出一定的解决方案，例如适当增加回复的人员数量，以及完善回复平台的功能等。这样会很大程度上增加用户留言的积极性，有助于有关部门更加及时、全面地了解公众的需求。同时，要注意尽量避免超过规定时间未回复（超过 90 天未回复）的情况发生，因为这会使公众感觉自己被忽视，并导致其对相关部门失去信任。

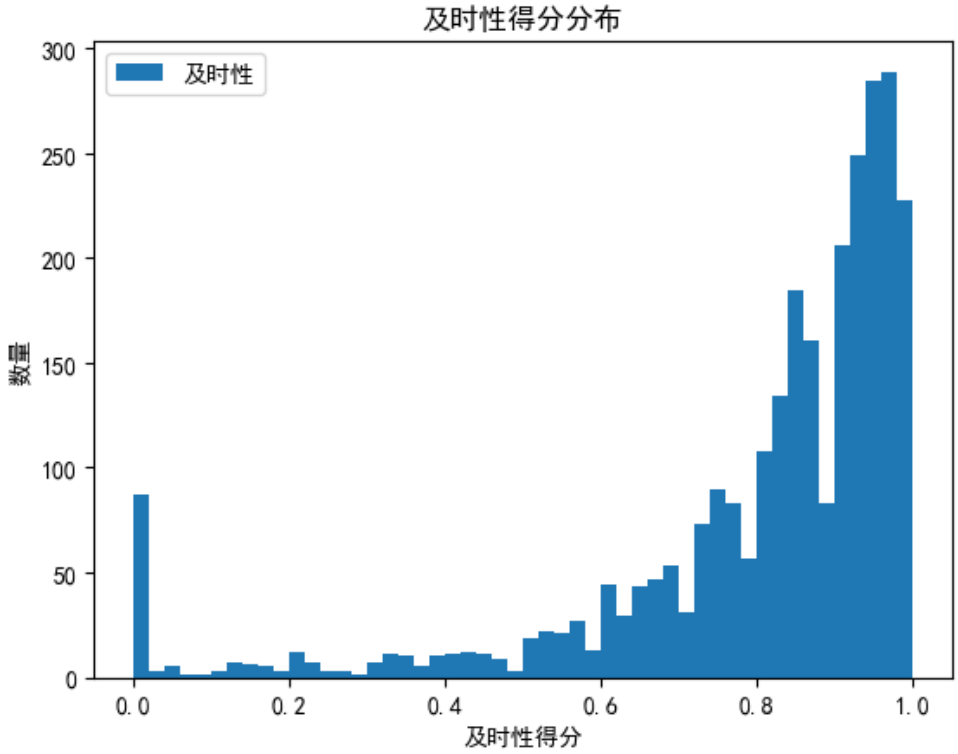


图 8：回复及时性得分分布图

3.3.2 相关性分析

有关部门的回复相关性得分（bm25 输出值）如图 9 所示，由图 9 中我们可以看出，对于绝大多数的留言数据（84.8%），其相关性得分大于 0，这说明有关部门对大多数的用户留言能给出具有针对性的回复。因此，我们推断，有关部门对于用户的留言有进行认真地思考和反馈。为了更进一步了解回复相关性的详细情况，我们同时绘制了回复相关性分布的直方图（图 10），如图 10 所示，我们发现，虽然绝大多数回复具有针对性，但针对性较高的回复仍只占少数，绝大多数的回复的相关性得分并不高。

为了探索不同相关性的回复文本的不同，我们随机筛选了部分高相关性、低相关性（具有相关性，但相关性得分接近 0）和无相关性（相关性得分小于等于 0）的回复文本进行人工检验。经检验发现，高相关性文本大多能够针对用户提出的问题进行逐条解答；而低相关性文本大多只能对用户提出的问题进行概括的说明；而无相关性的回复则大多给出某些条例的内容，甚至仅仅让用户参照某个条例或法规等。

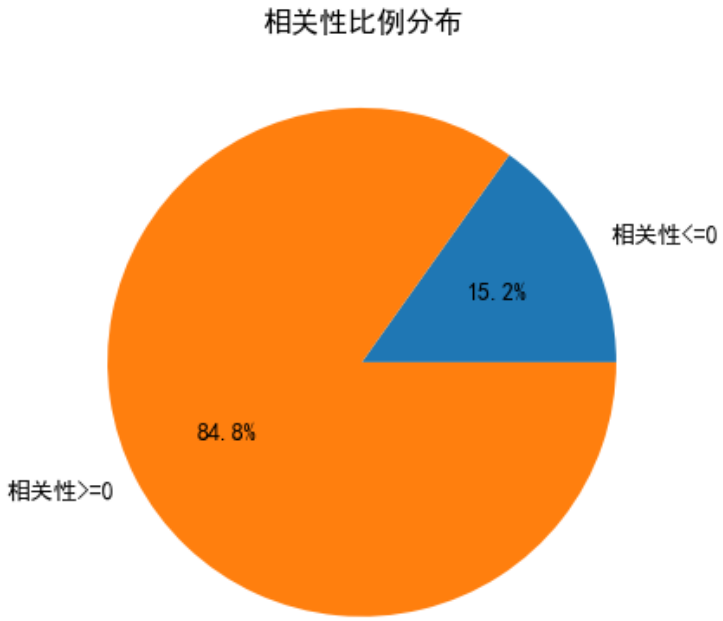


图 9：回复相关性得分（bm25 输出值）分布饼图

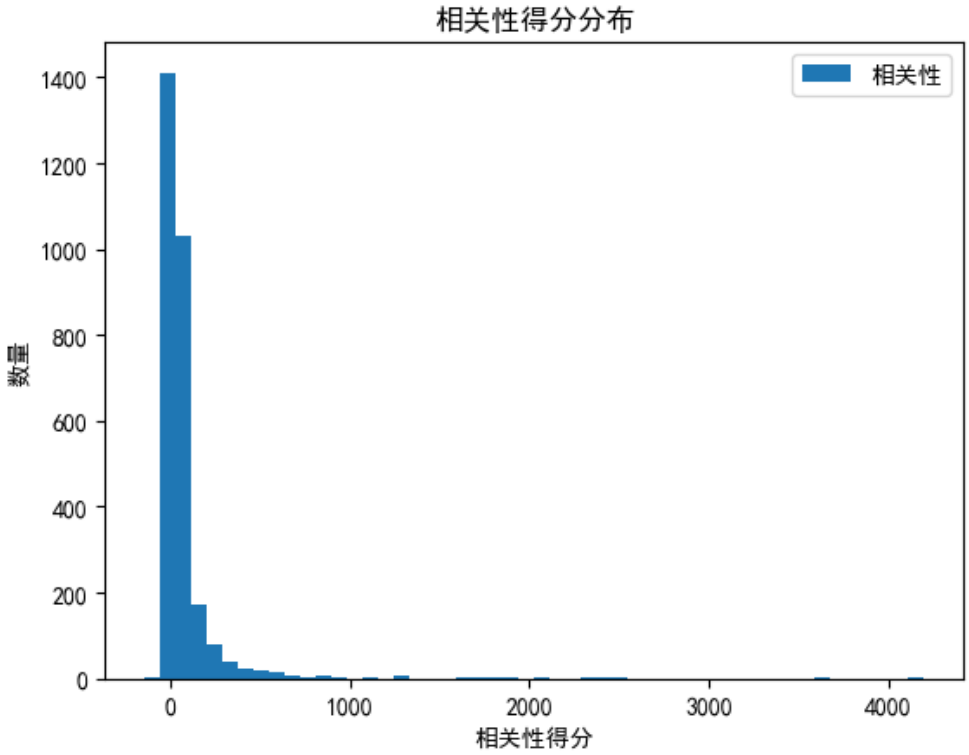


图 10：回复相关性得分（bm25 输出值）分布直方图

3.3.3 完整性分析

有关部门的回复相关性得分如图 11 所示，我们可以看出，有关部门回复意见的完整性并不高，绝大多数回复（92.43%）的完整性未超过 0.5，超过半数的回复（60.05%）的完整性未达到 0.3。这说明有关部门的回复未能很好地涵盖用户所提出的所有问题，从而造成用户的部分问题未能得到很好的解答。

为了进一步了解其完整性较低的原因，我们随机筛选了部分数据进行人工判别。发现大多数低完整性的回复仅对用户提出的问题进行了笼统地解释，因此并不能很好的涵盖所有问题。这说明有关部门的回复还需要更完整和全面。

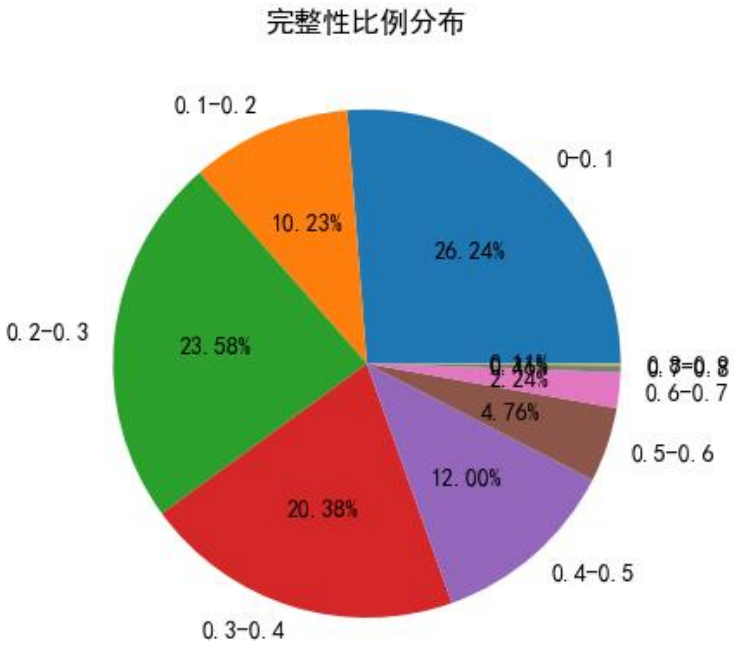


图 11：回复完整性得分分布图

3.3.4 总分分析

首先我们对及时性、相关性和完整性三个指标进行归一化，归一化后，我们采用熵值法对三个指标进行权重分配，赋权的结果如表 4 所示，从中我们发现相关性和完整性的权重相对较高，而及时性的权重较低。因为对于大多数的用户留言，有关部门都能及时给出回复，因此回复性对于区分回复质量的作用相对较低。

表 5：各指标权重表

指标	权重
相关性	0.34
完整性	0.54
及时性	0.12

随后，我们根据各指标的权重对有关部门的回复意见进行了打分，其公式为：

$$回复得分 = 0.34 \times 相关性得分 + 0.54 \times 完整性的分 + 0.12 \times 及时性 \quad （4）$$

回复得分的分布如图 12 所示，从图中我们可以看到，回复的整体性得分普遍偏低，其主要原因是回复的完整性相对较低。其次，多数回复的相关性得分也并不高。说明有关部门的总体回复质量并不高，需要进行相应的改进，改进建议如下：

（1） 对用户的留言要回复的更加完整，不漏掉其中的部分问题，更不该只是笼统的给出解答。

（2） 对用户的留言要更有针对性：有关部门应针对用户的每一点问题，给出具有针对性的建议，而不仅仅是回复某些专业的条例，照本宣科。

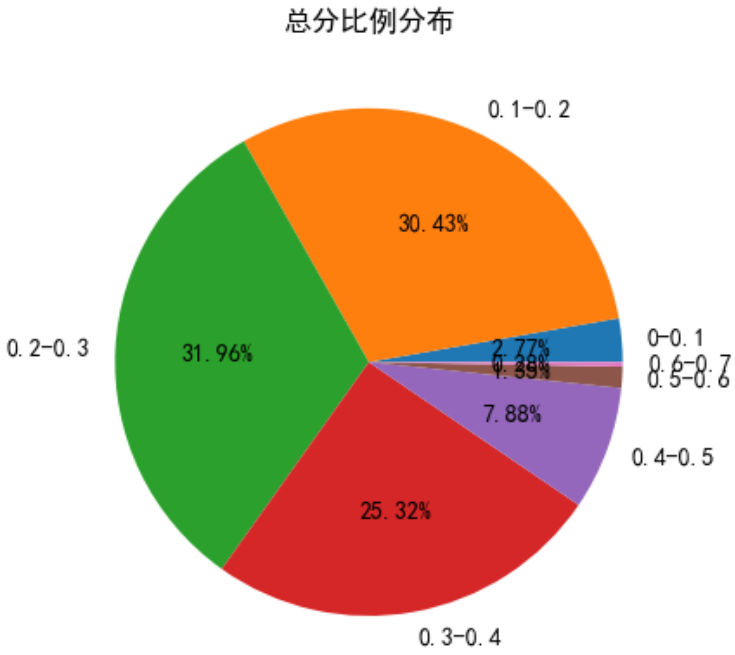


图 12：回复总分得分分布图

4. 结论

随着微博、微信、市长信箱等网络问政平台的发展和功能的完善，群众通过问政平台反应社会问题，表达自己看法的意愿愈发强烈。极速增长的数据量使得传统的人工留言划分、热点整理、政府反馈方式变得愈发难以应对。因而，通过自然语言处理技术、文本分类技术等技术进行政务处理具有重大意义。本文运用了 TF-IDF、MLP、K-means 等多种自然语言处理算法和分类算法，对群众的问政留言记录进行了深入的挖掘和分析。

由分析结果我们可以得出以下结论：（1）对于不同一级标签下的用户留言，其用词特点显著不同，因此采用文本向量化的方法，结合有监督分类算法，能准确的对用户的留言信息进行分类。（2）聚类算法难以针对大量文本数据进行聚类，所以数据的过滤十分重要。训练好的分类模型将全部文本分成七份有效地缓解了聚的压力。热度过滤方法有效地剔除了噪声数据又同时保留住热点评论。（3）对于政府的回复，及时性做的较好，做到了对用户的留言进行及时回复，而对于完整性和相关性两个方面，政府部门仍需要加强。

5. 参考文献

[1]张浩，汪楠. 文本分类技术研究进展[J]. 计算机与信息技术，2007， 23(1)：95—96.

[2]刘树春. 基于支持向量机和深度学习的分类算法研究[D]. 上海：华东师范大学，2015.

[3]侯思耕. 基于主题模型和深度置信网络的文本分类方法研究[D]. 昆明:云南大学，2015.

[4]郑翠翠. 面向领域文本的潜在语义分析研究[D]. 南京理工大学, 2010.

[5]胡学钢, 董学春, 谢飞. 基于词向量空间模型的中文文本分类方法[J]. 合肥工业大学学报:自然科学版, 2007, 30(10):1261-1264.

[6]王美方, 刘培玉, 朱振方. 基于 TFIDF 的特征选择方法[J]. 计算机工程与设计,2007, 28(23):5795-5796.

[7]邬启为. 基于向量空间的文本聚类方法与实现[D]. 北京交通大学, 2014.

[8]黄章益，刘怀亮．一种基于语义的中文文本特征降维技术研究[J]. 情报杂志,2011(S2):123-125.