基于人工智能技术的“智慧政务”文本挖掘研究

摘要：各级网络问政平台公众留言数量迅猛增长，给以往主要依靠人工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，迫切需要新的留言处理方式。人工智能技术在大数据时代取得了突破，在不同领域的应用带来了巨大的社会经济效应。因此本文将人工智能技术引入到智慧政务中的文本信息挖掘，利用附件所提供的数据、数据挖掘与数理统计等相关知识建立分类模型、建立指标和评价模型，解决题中留言信息分类、热点问题挖掘、政府答复评价三个问题。

数据预处理阶段，结合附件留言及答复数据的特点，基于 python、jieba、genism 等工具包和停用词表完成附件文本数据的文本清洗、文本分词、词性标注、去停用词和特征提取工作。将附件留言信息转化为向量形式，为后续文本挖掘工作提供数据基础。

针对问题一，首先构建基于 TF-IDF 和 SVM 分类模型，实现对附件 2 中留言数据的一级分类任务，并将附件 2 数据划分为训练数据集和测试数据集，采用十折交叉验证的 F1 均值作为模型评价指标;从 F1 值和时间复杂度两个角度将基于 TF-IDF 的 SVM 模型与机器学习或深度学习中的 Logistic Regression、Naive Bayes、Xgboost、CNN 等经典分类模型进行对比实验，验证了基于 TF-IDF 的 SVM 模型的优越性，说明该模型能较好地完成留言分类问题。

针对问题二，根据留言数据的表达不规范、不统一、地址信息不具体等特点，首先采用 CRF 模型和自定义位置信息词典对留言信息进行地理位置实体识别，并利用 kmeans 算法对提取到的地理位置特征进行聚类，完成相似性问题归类。然后结合查阅资料和现有数据确定热度评价指标，构造比较矩阵确定指标权重并进行一致性检验，将建立的热度评价模型对问题归类结果进行评价，得出热度排名前五的热点问题和热点问题明细。

针对问题三，首先通过查阅相关文献资料、法律法规等，将答复留言特征划分为文本特征和时序特征，分别从相关性、完整性、可解释性、时序性四个角度确定 TF-IDF 系数、文本长度、标点符号比重、内容词覆盖率、内容词密度、回复时间间隔等 6 个量化指标，然后运用层次分析法构造比较矩阵确定影响答复文本评价指标的权重，并进行一致性检验。最后统计计算附件 4 的 2816 条留言数据的得分，分别给出得分最高的和最低的 3 条答复，完成对政府答复留言的评价工作。

在报告的最后分析了留言文本信息挖掘的核心思路及其价值,浅谈完成实验的收获和感想，同时对人工智能技术在智慧政务中的应用提出新的展望。

关键词：TF-IDF；SVM；CRF；NLP；智慧政务；人工智能；AHP

Research on Text Mining of "Smart Government" Based on Artificial Intelligence Technology

Abstract: The rapid growth of the number of public messages on all levels of the Internet questioning platform has brought great challenges to the work of relevant departments that have traditionally relied on manual message division and hotspot sorting. New message processing methods are urgently needed. Artificial intelligence technology has made breakthroughs in the era of big data, and its application in different fields has brought huge social and economic effects. Therefore, this article introduces artificial intelligence technology to text information mining in smart government affairs. Use the data provided in the attachment, data mining, mathematical statistics and other relevant knowledge to establish classification models, indicators and evaluation models to solve the three problems of message information classification, hot spot problem mining and government response evaluation.

In the data pre-processing stage, combined with the characteristics of the attachment message and reply data, based on the Python language, jieba, genism and other toolkits and the Harbin Institute of Technology stop word list, complete the text cleaning of the attachment text data, text segmentation, part-of-speech tagging, and stop word removal and feature extraction. Convert the attachment message text information into a vector form to provide a data basis for subsequent text mining.

To solve the first problem, first build a classification model based on TF-IDF and SVM to implement the first-level classification task for the message data in Annex 2, and divide the Annex 2 data into a training data set and a test data set, using F1 with 10-fold cross-validation. The mean value is used as the model evaluation index. Then, from the perspective of F1 value and time complexity, the SVM model based on TF-IDF is compared with classic classification models such as Logistic Regression, Naive Bayes, Xgboost, CNN in machine learning or deep learning. In addition, We also tried the combination of word2cec and svm. The superiority of the SVM model

based on TF-IDF is verified, indicating that this model can better complete the message classification problem.

To solve the second problem, according to the characteristics of message data such as irregular expression, inconsistency, and unspecific address information, the CRF model and a custom location information dictionary are first used to identify the geographic location of the message information, and the kmeans algorithm is used to extract the geographic location. The location features are clustered to complete the classification of similarity problems. Then combine the reference data and the existing data to determine the heat evaluation index, construct a comparison matrix to determine the index weights and perform consistency test, and evaluate the classification results of the established heat evaluation model to obtain the top five hot issues and hot spots problem details.

To solve the third problem, first of all, by referring to relevant documents, laws and regulations, etc., the characteristics of the reply message are divided into text features and time series features. Six quantification indexes, such as TF-IDF coefficient, text length, punctuation weight, content word coverage, content word density, and response time interval, were determined from four angles of relevance, completeness, interpretability, and timing. Then use AHP to construct a comparison matrix to determine the weights that affect the evaluation index of the response text, and conduct a consistency check. Finally, statistically calculate the score of the 2816 message data in Annex 4, and give the three answers with the highest score and the lowest score respectively, completing the evaluation of the government's reply message.

At the end of the report, it analyzes the core ideas and values of message text information mining, talks about the harvest and feelings of completing the experiment, and at the same time puts forward new prospects for the application of artificial intelligence technology in smart government affairs.

Keywords: TF-IDF; SVM; CRF; NLP; smart government affairs; artificial intelligence; AHP

目 录

一、引言.....2
1.1 研究背景.....2
1.2 文献综述.....3
1.3 研究目的及意义.....4
二、实验方案.....6
2.1 研究思路.....6
2.2 技术路线图.....7
三、实验过程.....8
3.1 数据预处理.....8
3.1.1 文本清洗.....8
3.1.2 文本分词.....8
3.1.3 词性标注.....8
3.1.4 去停用词.....10
3.1.5 特征提取.....11
3.2 问题一 群众留言分类.....11
3.2.1 基于 TF-IDF 和 SVM 的留言分类.....11
3.2.2 方法对比实验.....13
3.3 问题二 热点问题挖掘.....14
3.3.1 相似问题归类.....14
3.3.2 问题热度评价.....15
3.4 问题三 答复评价.....18
3.4.1 答复评价因素研究.....18
3.4.2 答复评价指标及权重.....21
3.4.3 答复综合评价模型.....23
四、实验结果.....25
4.1 留言分类结果分析.....25
4.2 热点问题挖掘分析.....25
4.3 回复意见评价分析.....28
五、总结与展望.....31
5.1 实验结论.....31
5.2 感想与展望.....32
致谢.....33
参考文献.....34

一、引言

1.1 研究背景

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了政府了解民意、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。据人民网 2019 年 12 月 12 日报道，“截至 2019 年 11 月 30 日，网民通过《领导留言板》给省市县三级领导干部留言累计突破 200 万件，140 万件诉求得到各级政府妥善解决，整体答复率达到 70%”。综合大数据分析，《领导留言板》工作表现出“移动端日益成为网民留言的主要渠道、网上群众工作基层参与度稳步提升”等一系列新的趋势和特点。

各级网络问政平台公众留言数量迅猛增长，人工处理任务繁重，如某省会级城市的问政平台留言板，每日受理群众诉求数千条，每日人均处理新增留言 400~700 条。现有人工处理方式会带来留言审核不及时、职能部门回复慢等问题，势必会导致公众满意度降低，不利于塑造良好的政府形象，因此留言的处理效率亟待提高。因此目前政媒融合问政平台的公众留言数量激增，给传统人工人类处理带来巨大压力，且无法保证分类正确性和一致性，迫切需要新的方式处理留言。

针对上述问题，探讨通过机器学习等方法来自动实现留言数据的快速分类、信息挖掘等具有重要意义。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。人工智能领域核心技术主要包括计算机视觉、机器学习、自然语言处理、机器学习和语音识别等五大类。政务留言文本挖掘主要运用到的是机器学习和自然语言处理技术。机器学习是使用机器来模拟人类学习活动，以获取新的知识或技能，并重新组织已有的知识结构使之不断改善自身的性能，如通过训练样本学习文本相似度潜在规律，实现对留言的自动分类等。而自然语言处理是指计算机拥有的人类般的文本处理的能力，从可读的、风格自然、语法正确的文本中自主解读出含义，如对文本信息的主题识别、命名识别等任务。将人工智能等技术引入智慧政务领域意义十分重大，其对缓解人力资源匮乏、提升政府服务管理效能、提高政府决策质量等方面都具有重要促进作用，必将带来政务服务部门的重大变革。

1.2 文献综述

国内外学者对网络留言文本挖掘开展了大量相关研究，目前常用的文本分类和预测方法有：（1）决策树归纳分类法，其原理是自顶向下递归的分治方法，从训练元组集和它们的相关联的类标号开始构造决策树，从而训练集递归地划分成较小的子集。（2）基于规则的分类，其原理是使用一组 IF-THEN 规则进行分类，根据规则质量的度量，如准确率、覆盖率或者根据领域专家的建议，将规则组织成一个优先权列表，以此来计获具有最高优先权的分类预测。（3）基于智能统计和学习的方法。基于统计的方法有 KNN、SVM、Bayes Rocchio 等分类方法，在文本分类中应用广泛。（4）基于深度学习的方法。近年来随着数据的爆发式增长和计算机存储计算能力的提升，深度学习方法在自然语言处理和图像处理等领域都有突出表现，但深度学习较依赖于训练数据样本量，且训练过程和输出结果具有可不解释性。

而在电子政务文本分析领域，依据政务文本的规范化程度，可以将现有政务领域关于文本的研究分为正式文本、半正式文本和非正式文本。其中正式文本一般指发布主体为政府或权威机构，如法律法规、政策文本、权威报告等。非正式文本指口语化程度很高的文本，如公众留言。半正式文本为经过初步处理的非正式文本，如经过培训的话务员记录的留言。近年来，基于机器学习的文本自动分类算法引起了很多学者的关注[1]。自动文本分类是指机器按照一定的标准或分类体系对文本集进行自动分类标记的行为[2]。常用的文本分类的方法有朴素贝叶斯（NB）、逻辑回归、决策树分类（DT），k 近邻算法（KNN）[3]，支持向量机（SVM）[4]和卷积神经网络（CNN）[5]等。其中，支持向量机在解决小样本、非线性的文本分类问题时表现出了许多特有的优势[6]。研究表明，支持向量机的分类性能尤其是泛化能力优于传统的分类方法[7]。政务领域关于正式文本的研究多为政策文本，研究方法主要有内容分析法[8]、话语分析[9]、文本挖掘方法[10]。除政策文本外，也有学者对权威机构发布的文本如诉讼文本[11]、审计报告[12]等进行文本分类研究。上述文本一般具有表述规范、且具有一定的长度的特点。关于政务领域非正式文本的现有研究，学者主要关注问政平台中的公众留言文本。研究所利用问政平台数据较为丰富，包括留言的时间数据[13]，留言的数量数据[14]、浏览量、满意度等结构化数据以及公众留言的文本数据[15]等非结构化数据。关于留言数据

的利用类型，国内外研究利用留言中的结构化数据如性别、时间、空间数据等较多，而对工中留言的文本数据的研究较为缺乏。半正式文本的研究中，有关于市长公开电话文本自动分类技术的比较研究，其分类的文本数据是接话员在满足记录的规范性要求前提下，将市民意见以文本形式录入并提交[16]，其文本的规范程度远高于工中留言。

综上所述，目前针对网络留言文本挖掘已经成为国内外学者的研究热点问题，研究方法和研究成果较为成熟。但现有研究多集中于留言信息过滤、社区问答系统质量评价、文本相似性度量等领域，其中针对电子政务、智慧政务的非正式文本分类和评价研究较少，由于电子政务留言信息具有格式不规范、表达不统一、文本较短且题中所给数据集样本量较少等特点，自然语言领域中表现较好的深度学习方法不适用于题中所给问题的处理与挖掘，需要综合考虑准确度和效率，探索更加实用的机器学习方法。

1.3 研究目的及意义

当前我们正在逐步进入“互联网+”、大数据以及人工智能三位一体的人类社会新时代[17]。在政务服务领域，“互联网+”热潮的到来改变了原有的政务服务管理模式，办事群众可以足不出户便享受到全方位的立体服务。因此如何分析处理这一模式下产生和积累的海量政务服务数据对于提升服务质量和办事效率有着机器显著的作用。而要实现对海量数据的处理，就自然要应用到人工技能技术，人工智能进入政务服务领域是“互联网+政务服务”的自然选择，并将会在其中有着广泛的应用前景。人工智能已经在社会生活各个领域得到了广泛的应用，促进了经济、社会的不断发展，而在政务服务领域，引入人工智能，其对缓解人力资源匮乏、提升政务服务管理效能、提高政府决策质量等方面具有重要促进作用。虽然目前人工智能在政务服务领域的应用尚处于起步阶段，但已经开始在一定程度上产生了积极的带动效应。根据工作的复杂性以及自动化程度，人工智能可以从分别从解放、分解、取代、增强[18]等四个方面在政府服务领域中发挥作用。

运用人工智能技术对网络问政留言信息的进行挖掘与分析，可以将政务服务人员从繁重的留言信息处理工作中解放出来，并通过分类、热点问题挖掘等自动、快速、准确地发现存在问题，反馈给相关管理部门，使部门工作人员有效地解决问题。利用文本抽取技术进行事件抽取，将非结构化数据转化为结构化数据，形

成统一的表达形式。在此基础上根据已有数据的统计，从多个维度进行统计分析，发现热点事件、热点地区，并对智慧政务和城市建设提供有力的支撑。

综上所述，随着数据资源的积累和大数据技术的创新，使得网络问政留言信息的分析挖掘成为可能。它不仅为百姓提供了便捷的服务，而且为城乡管理建设、政务服务工作提供帮助。由于信息化建设对百姓民生、城市发展有着重要的推进作用，对城市投诉的民生问题，政府部门在解决问题上效率更高、更及时。因此，政务文本挖掘具有重要的研究价值和现实意义。

二、实验方案

2.1 研究思路

随着网络问政平台留言文本信息量的迅速积累，给以往主要依靠人工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。本文将人工智能方法应用到电子政务留言处理领域，主要利用附件所提供的留言数据，结合三个题目要求，最终构建出留言分类、热点问题挖掘、答复质量评价等模型，实验过程包括以下四个步骤：。

（1）数据预处理

结合附件留言及答复数据的特点，基于 python 语言、jieba、genism 等工具包和哈工大停用词表完成附件文本数据进行文本清洗、文本分词、词性标注、去停用词、和特征提取。

（2）群众留言分类

通过控制变量法分别对比向量特征提取方法（TF-IDF 与 Word2vec）和分类模型（SVM、Logistic Regression、Navie Bayes、Xgboost、CNN）在附件数据中的表现，最终选择 F1 最高的 TF-IDF 和 SVM 组合模型对附件 2 留言数据进行训练分类，采用十折交叉验证得到分类模型的 F1 值。

（3）热点问题挖掘

由于附件信息中关于地理位置的描述存在“A 区 A5 市”这种字母代替具体位置的情况，所以根据描述构成规则建立自定义地理位置词典，识别出该类指代模糊的地理位置描述，同时结合 CRF 模型实现对常用地理位置信息的识别完成位置信息实体识别，然后采用 kmeans 算法对识别结果进行聚类，实现相似性问题归类。最后借助判断矩阵定义问题热度评价指标及权重，对问题归类结果进行热度计算，得到热度排名前五的热点问题。

（4）答复意见的评价

结合查阅文献资料和相关法律法规，从答复信息文本特征和时序特征两个角度确定 TF-IDF 系数、文本长度、标点符号比重、内容词覆盖率、内容词密度、回复时间间隔等 6 个量化指标，在数据预处理的基础上，利用 python 编程语言统计每条答复对应的各个指标数值，通过层次分析法确定指标权重，并对每条答复的综合评分进行计算，得到附件 4 中 2816 条答复的综合评分。

2.2 技术路线图

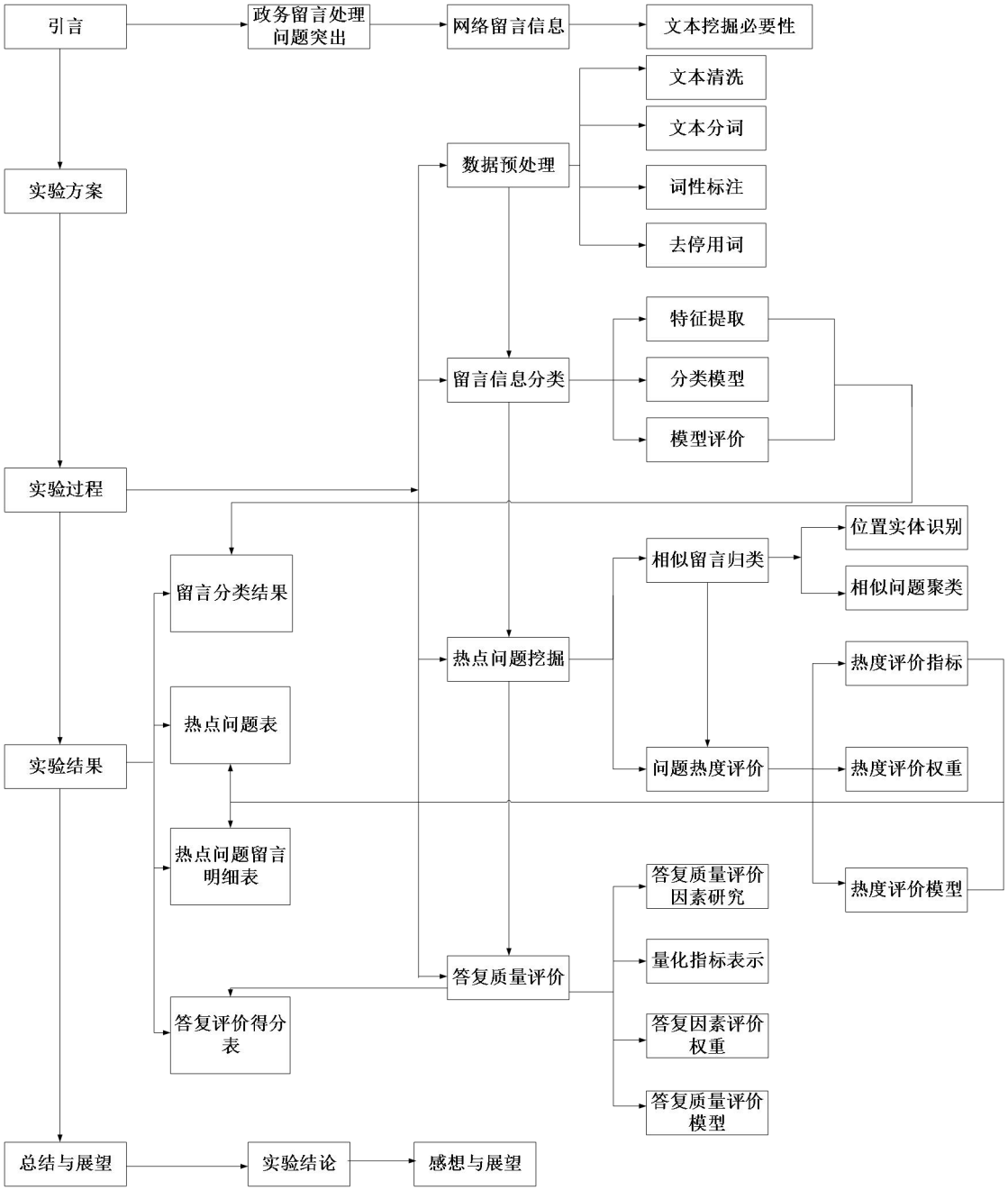

图 2.1 技术路线图

三、实验过程

3.1 数据预处理

3.1.1 文本清洗

题中所给数据是较为整齐的结构化表格，无需太多的数据清洗工作。但附加 2 中的留言详情是文本分类和后续挖掘工作的数据基础，详情信息语句前后存在大量的空字符串，影响后续数据分析。所以利用正则表达式进行文本内容清洗，去除文本中的空白字符和无意义字符，只保留中英文和数字。正则表达式中只保留中英文和数字的模式表达式为‘[^\u4e00-\u9fa5^a-z^A-Z^0-9]’。

3.1.2 文本分词

在完成文本清洗工作后，需要进一步对留言文本进行文本分词。文本分词工作是后续将文本信息转化为向量信息的基础，准确的文本分词为后续文本特征提取提供保障。文本分词指的是将句子中的字序列根据一定的规则组合成词序列的过程。与英文句子不同，中文句子的词语之间缺少自然分界符，而是以字序列的形式出现，因此中文分词比英文分词更为复杂困难。

Jieba 分词是最常用的中文分词工具之一，涉及的核心分词算法主要包括基于树结构实现高效的词图扫描、采用动态规划法获取最大概率路径、基于隐马尔可夫模型识别未涵盖词等。本文选用 Jieba 工具用于留言文本的分 词工作。例如表 3.1 为附件二中某条留言的分词结果。

表 3.1 运用 Jieba 对留言信息进行分词

留言信息	A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。
分词结果	A3，大道，西行，未管，路口，加油站，路段，人行道，包括，路灯，西湖，建筑，集团，燕子，安置，项目，施工，围墙，每天，尤其，上下班，期间，路上，人流，车流，安全隐患，非常，强烈，请求，文明城市，尽快，整改，这个，文明，路段

3.1.3 词性标注

词性标注是指将词语的词性标注出来的过程，即名词、动词、形容词等。

除了语法关系，句中词语的词性也蕴含着信息，定义了它的用途和功能。问题三中答复意见评价指标中涉及内容词密度、内容词覆盖率等指标，因此需要预先对分词结果进行词性标注。此过程同样利用 Jieba 工具完成，它可以对分词的词语进行词性标注，词性类别如表 3.2 所示：

表 3.2 Jieba 词性标注类别及含义

Ag	形语素	形容词性语素。形容词代码为 a，语素代码 g 前面置以 A。
a	形容词	取英语形容词 adjective 的第 1 个字母。
ad	副形词	直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起。
an	名形词	具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起。
b	区别词	取汉字“别”的声母。
c	连词	取英语连词 conjunction 的第 1 个字母。
dg	副语素	副词性语素。副词代码为 d，语素代码 g 前面置以 D。
d	副词	取 adverb 的第 2 个字母，因其第 1 个字母已用于形容词。
e	叹词	取英语叹词 exclamation 的第 1 个字母。
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。
h	前接成分	取英语 head 的第 1 个字母。
i	成语	取英语成语 idiom 的第 1 个字母。
j	简称略语	取汉字“简”的声母。
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。
m	数词	取英语 numeral 的第 3 个字母，n，u 已有他用。
Ng	名语素	名词性语素。名词代码为 n，语素代码 g 前面置以 N。
n	名词	取英语名词 noun 的第 1 个字母。
nr	人名	名词代码 n 和“人(ren)”的声母并在一起。
ns	地名	名词代码 n 和处所词代码 s 并在一起。
nt	机构团体	“团”的声母为 t，名词代码 n 和 t 并在一起。
nz	其他专名	“专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。
o	拟声词	取英语拟声词 onomatopoeia 的第 1 个字母。
p	介词	取英语介词 prepositional 的第 1 个字母。
q	量词	取英语 quantity 的第 1 个字母。
r	代词	取英语代词 pronoun 的第 2 个字母,因 p 已用于介词。

s	处所词	取英语 space 的第 1 个字母。
tg	时语素	时间词性语素。时间词代码为 t,在语素的代码 g 前面置以 T。
t	时间词	取英语 time 的第 1 个字母。
u	助词	取英语助词 auxiliary
vg	动语素	动词性语素。动词代码为 v。在语素的代码 g 前面置以 V。
v	动词	取英语动词 verb 的第一个字母。
vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。
w	标点符号	
x	非语素字	非语素字只是一个符号，字母 x 通常用于代表未知数、符号。
y	语气词	取汉字“语”的声母。
z	状态词	取汉字“状”的声母的前一个字母。
un	未知词	不可识别词及用户自定义词组。

3.1.4 去停用词

停用词一般指对文本特征没有任何贡献作用的字词，经过分词步骤解析后的文本中有很多无效的词，比如“着”,“和”等，需要去掉这类引入对文本分析工作没有任何意义的停用词。本文选用哈工大停用词表对留言数据进行处理，得到隐含信息量较高的简洁文本信息。例如表 3.3 为附件二中某条留言分词后去停用词的结果。

表 3.3 对分词结果去停用词

分词结果	A3，大道，西行，未管，路口，加油站，路段，人行道，包括，路灯，西湖，建筑，集团，燕子，安置，项目，施工，围墙，每天，尤其，上下班，期间，路上，人流，车流，安全隐患，非常，强烈，请求，文明城市，尽快，整改，这个，文明，路段
去停用词结果	A3，大道，西行，未管，路口，加油站，路段，人行道，包括，路灯，西湖，建筑，集团，燕子，安置，项目，施工，围墙，上下班，期间，路上，人流，车流，安全隐患，请求，文明城市，整改，文明，路段

3.1.5 特征提取

经过上述步骤对数据进行清洗处理后，需要将文本符号信息转化为计算机可以理解的向量信息，从而把自然语言问题转化为机器学习问题。TF-IDF 和 Word2vec 是两种常用的词向量模型，其中 TF-IDF 是一种基于统计方法的加权技术，根据字词在文本中出现的次数和在整个语料中出现的文档频率来计算一个字词在整个语料中的重要程度。Word2vec 为计算向量词提供了一种有效的连续词袋和 skip-gram 架构，通过转换将对文本内容的处理简化为空间向量中的向量运算，计算出向量空间上的相似度，来表示文本语义上的相似度，其中词的顺序是不重要的。

根据后续文本挖掘工作需要，问题一群众留言分类中需要比较不同向量化方法和分类模型之间的准确度与效率，因此数据预处理阶段主要运用 TF-IDF 和 Word2vec 这两种常用方法完成文本向量化过程。调用 TfidfVectorizer 类完成 TF-IDF 向量化工作，考虑词语的先后顺序，将参数设置为 ngram=2，max_feature=3000。调用 gensim 模块中的 Word2vec，将词向量的维度设为 3000，且只考虑最小出现频率为 10 的词，取 Word2vec 中的对应词向量，并分别对每一维相加，最后取平均生成句子的 Vector。

3.2 问题一 群众留言分类

3.2.1 基于 TF-IDF 和 SVM 的留言分类

（1）关键技术介绍

①TF-IDF:

TFIDF(term frequency-inverse document frequency)方法由 Salton 在 1988 年提出，主要用于提取特征和计算特征权重。其主要思想是如果一个词语在某类文档中出现频率高，且在其它类文档中很少出现，则认为该词对于这类文档具有较好的区分能力，适合作为特征用于分类。TFIDF 方法因其算法简单易行，准确率较高，常用于关键词和特征提取的研究中。TFIDF 函数主要由 TF (term frequency)和 IDF (inverse document frequency)两部分构成。其中 TF 表示词频，反映词语在某文档中的出现频率。IDF 表示逆文档频率，反映有多少个文档出现了特定词语，出现的文档越多。

②SVM

利用 SVM(Support Vector Machine) 做分类是机器学习比较成熟的算法。非线性映射是 SVM 方法的理论基础，SVM 利用内积核函数代替向高维空间的非线性映射。对特征空间划分的最优超平面是 SVM 的目标，最大化分类边际的思想是 SVM 方法的核心。支持向量是 SVM 的训练结果，在 SVM 分类决策中起决定作用的是支持向量。SVM 是一种有坚实理论基础的小样本学习方法。它基本上不涉及概率测度及大数定律等，因此不同于现有的统计方法。从本质上看,它避开了从归纳到演绎的传统过程，实现了高效的从训练样本到预报样本的“转导推理”，大大简化了通常的分类和回归等问题。SVM 的最终决策函数只由少数的支持向量所确定,计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾难”。少数支持向量决定了最终结果，这不但可以帮助我们抓住关键样本、“剔除”大量冗余样本，而且注定了该方法不但算法简单，而且具有较好的“鲁棒”性。

③K 折交叉验证

K 折交叉验证(k-fold cross-validation)首先将所有数据分割成 K 个子样本，不重复的选取其中一个子样本作为测试集，其他 K-1 个样本用来训练。共重复 K 次，平均 K 次的结果或者使用其它指标，最终得到一个单一估测。 K 折交叉验证使用了无重复抽样技术的好处：每次迭代过程中每个样本点只有一次被划入训练集或测试集的机会。本研究使用了 10 折交叉验证，其过程如下图。

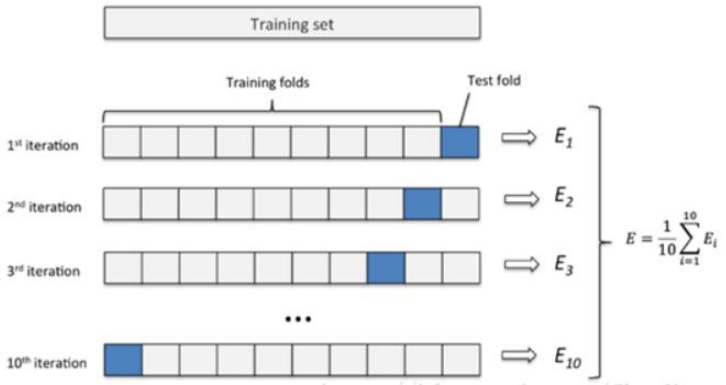

图 3.1 10 折交叉验证过程

（2）留言分类模型

附件 2 所给的文本信息数据量有限，经试验发现 CNN 与 RNN 等深度学习方法并未取得较好的效果。而 SVM 是一种有效的小样本学习方法，经过试验和调参发现 SVM 在几种算法中取得的效果最好。所以本文首先选用 TF-IDF 来实

现文本信息到空间向量的映射，然后将附件 2 数据划分为训练数据集和测试数据集，运用 SVM 对文本空间向量进行模型训练和预测，最后采用十折交叉验证的 F1 平均值来度量模型性能。具体算法流程如下：

算法 2 基于 TF-IDF 和 SVM 的分类模型
输入：留言文本信息
过程：
1. 文本预处理。去停用词，去空行，然后将处理后的文本放入空的列表，这个列表最终由一个个元组组成，元组前面是由空格分割的文本，后面是类别。
2. 将文本信息与标签分开，x 是文本，y 是标签。
3. 采用 TF-IDF 方法训练词向量表。考虑词的先后顺序，故 N-gram 取 2。
4. 随机打乱数据顺序，进行 10 折交叉验证，训练 SVM 分类器进行测试，分别求每轮的 F1 值。
输出：交叉验证 F1 平均值。

3.2.2 方法对比实验

为了直观比较度量上文 SVM 模型的有效性，本文同样运用了 Logist Rregression、Navie Bayes、CNN 等方法在附件二数据上进行分类实验。

（1）Logist Regression

逻辑回归（Logist Regression, LR）是一种广义的线性回归分析模型，常用于解决机器学习中的分类问题，用于估计某种事物的可能性。调用 python 的 scikit learn 工具包实现 LR 模型对附件 2 的分类任务。

（2）Xgboost

Xgboost 是 boosting 算法的一种，boosting 算法的思想是将许多弱分类器集成在一起形成一个强分类器。因为 Xgboost 是一种提升树模型，所以它是将许多树模型集成在一起，形成一个强分类器。

（3）Navie Bayes

朴素贝叶斯（Navia Bayes, NB）是基于概率论的经典机器学习分类算法之一。其中朴素是指对于模型中各个特征有强独立性的假设，并未将特征间的相关性考虑其中，在垃圾邮件分类中广泛应用。调用 python 的 scikit learn 工具包实现 NB 模型对附件 2 的分类任务。

（4）CNN

卷积神经网络（Convolutional Neural Network，CNN）是一种前馈型的神经网络，其在大型图像处理方面有出色的表现，目前已经被大范围使用到图像分类、定位等领域中。相比于其他神经网络结构，卷积神经网络需要的参数相对较少，

使的其能够广泛应用。本文中 CNN 实现基于 tensorflow 深度学习平台，搭建两层 CNN 模型，参数设置如下：文档最大长度 100，最小词频数 2，词嵌入维度 20，filter 个数 50，感知野大小 20，激活函数为 relu，形成词向量-->第一个卷积层-->第一个池化层-->第二个卷积层-->第二个池化层结构。如图 3.1 所示。

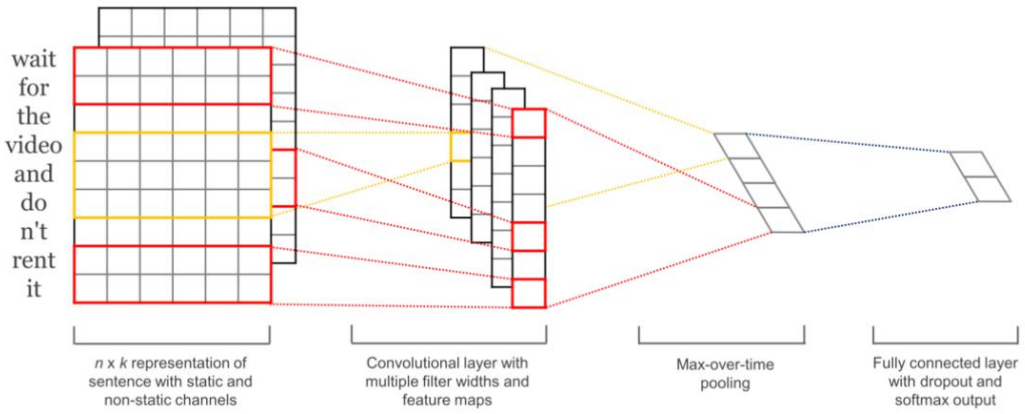

图 3.2 CNN 神经网络

3.3 问题二 热点问题挖掘

3.3.1 相似问题归类

网络媒体平台上的热点问题是指某一时段内群众集中反映的某一问题，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。所以热点问题发现可以转化为特定地点或特定人群集中反映的问题相似性度量和归类问题。

目前已有的地理位置识别研究多针对于新闻等格式与规范严格的规范语料，中文地名特征比较明显，如“北京市”等宽泛的地名。但仔细分析附件 3 中数据，发现存在大量如“K3 市”、“A 市”、“A 市 A5 区”等指代模糊、格式不规范、表达不统一的地名，同时留言中的地理位置实体具有结构复杂、描述的地理位置较详细等特点，使得传统地名识别方法不适用于此类语料中的地理位置实体识别。针对网络媒体留言的特殊性，本文选用 CRF 模型训练与测试语料特征提取，并人工自定义位置词典对用字母表示的区市位置信息进行字符匹配，两种方法结合实现对具体地点和模糊地点的提取，从而完成特定地点聚类工作。

条件随机场（CRFs，Conditional Random Fields）是一个序列标注模型，结合了最大熵和隐马尔可夫模型的有点，克服了标注偏置问题，是目前最优秀的机器学习模型之一。在给定一个 token 序列 x 下，其标注序列 y 的概率如公式 3-1

所示：

$$P(\mathrm{y}\mid \mathrm{x})=\frac{1}{Z(x)}\exp(\sum_{i,k}\lambda_k f_k(y_{i-1},y_i,x)+\sum_{i,k}\mu_k g_k(y_i,x)) \tag{3-1}$$

其中，Z(x)表示一个在所有状态序列上的归一化因子；λ_k 和 u_k 是针对每个特征函数学习得到的权重，可以通过 L-BFGS 算法求解；f_k 和 g_k 是在 CRF 图中边和节点的状态函数。最终得到 P(y|x)最大化的参数模型。测试阶段，标注序列 y 是满足 maxP(y|x)的标注序列。

地理位置实体的特征选取是识别地理位置实体的基础，分析留言文本中的地理位置实体，以及中文地点的特点，本文提取了词形、词性、地理位置词典、后缀词四个特征，并对其进行特征标注。

(1) 词形（Word）：通过 NLPIR 词语切分后单个词语；

(2) 词性（POS）：通过 NLPIR 词语切分后的词语的词性；

(3) 地理位置词典（Dic）：本文的地理位置词典特征主要有两类，一类数于常用地理位置词典库中的词语；另一类是自定义“*市”或“*区”的英文字母表示的模糊位置信息。

(4) 后缀词：本文中的后缀词主要分为两类，一类是属于尾词词典中的词语；另一类是属于尾字词典中以尾字结束的词语。

针对地理位置实体的特征，构建尾字词典、尾词词典以及常用地理位置词典库和自定义词典库，并作为特征资源库对留言文本进行特征提取，运用 CRF 模型实现对留言信息基于地理位置的识别，并运用 k-means 算法对提取的地理位置进行聚类分析，从而实现对相似问题的归类。

3.3.2 问题热度评价

（1）问题热度评价指标

热点话题的评价指标有很多，由于留言数据的限制，有留言用户、留言主体、留言时间、留言详情、点赞数和反对数几种数据，在此基础上选取可利用的留言量、留言用户数、平均点赞及反对量、单条留言最高点赞量及反对量等评价指标，构建问题热度的量化方法。

每天的留言量。留言所提出的问题，有的可能持续时间较长，但反映情况并不强烈，以问题的持续时间为指标不太准确。在某个问题持续反映时间内，每天

的留言数量能够更合理反映问题的热门程度。

留言的用户数。考虑到同一个用户可能会留言多次，这样的情况会影响评价结果，因此本文选取某个问题参与留言的用户数量作为评价问题热度的评价指标，这一指标能够反映问题的影响广度。

平均点赞量及反对量。用户的点赞量及反对量反映了用户的感兴趣程度，点赞和反对都能反映问题的影响范围，点赞数和反对数越高，表明留言收到的关注越广泛。因此，本文选取了留言平均点赞数和反对数以合理评价话题热度。

单条留言最高点赞量及反对量。考虑到某个问题在某个时间点可能受到的关注度较高，点赞量和反对量很多，因此只用平均点赞量和反对量不足以反映话题的热度。

（2）问题热度计算

由于数据的限制，本文根据热度量化的指标即每天的留言量，留言的用户数，留言的平均点赞量和反对量以及单条留言最高点赞量及反对量四个指标构建问题热度的计算方法，公式如下:

$$heat = \alpha 1 * message + \alpha 2 * user + \alpha 3 * (likes + dislikes) + \alpha 4 * m\_like \tag{3-2}$$

上式中，$\alpha 1+\alpha 2+\alpha 3+\alpha 4=1$，heat 表示问题的热度，message 是问题对应的留言数量，user 是问题对应留言的用户数量，likes 和 dislikes 分别表示问题对应留言的平均点赞量和反对量，m_like 是单条留言的点赞量和反对量之和的最大值，$\alpha 1$，$\alpha 2$，$\alpha 3$，$\alpha 4$ 表示各个指标的权重。

构建判断矩阵。通过构建判断矩阵来计算 $\alpha 1$，$\alpha 2$，$\alpha 3$，$\alpha 4$ 的值，利用两个因素之间两两比较的方法，即表 3.4 标度法，计算得到各个指标的权重。

表 3.4 重要性标度表

定义（B_{ij}）	标度
B_i 指标比 B_j 指标一样重要	1
B_i 指标比 B_j 指标稍微重要	3
B_i 指标比 B_j 指标明显重要	5

B_i 指标比 B_j 指标重要的多	7
B_i 指标比 B_j 指标极端重要	9
B_i 指标比 B_j 指标重要性在两个判断尺度中间	2, 4, 6, 8

构造判断矩阵 A：

$$A=\begin{bmatrix}1 & 1 & 1/3 & 1/3\\ 1 & 1 & 1/3 & 1/3\\ 3 & 3 & 1 & 1/2\\ 3 & 3 & 2 & 1\end{bmatrix}$$

利用规范列平均法求权重，首先对列向量归一化得到对应矩阵 C，再对其按行求和并归一化，得到特征向量 w。计算得到特征向量 w=[0.123,0.123,0.313，0.439]。

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 的元素为同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 确定不一致的允许范围。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 $\lambda \geq n$，当且仅当 $\lambda = n$ 时，A 为一致矩阵。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，引起的判断误差越大。因而可以用 $\lambda - n$ 数值的大小来衡量 A 的不一致程度。定义一致性指标为：

$$CI=\frac{\lambda-n}{n-1} \tag{3-3}$$

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。计算得到 CI=0.022<0.1，A 的一致性程度在容许范围之内，有满意的一致性，通过一致性检验。即 $\alpha 1=0.123$，$\alpha 2=0.123$，$\alpha 3=0.313$，$\alpha 4=0.439$。问题热度公式如下所示：

$$heat=0.123*message+0.123*user+0.313*(likes+dislikes)+0.439*m\_lik$$

(3-4)

3.4 问题三 答复评价

3.4.1 答复评价因素研究

相关部门对留言的答复意见是政府部门对群众提出问题的反馈，不仅体现了政府的为解决问题的工作投入，也直接关系到公众满意度。因此建立科学有效的答复质量评价模型能够定量衡量政府工作效率，为合理评价政府工作和政府服务职能的改进具有重要意义。回复留言中包含的信息结构复杂，如何有效挖掘和利用这些信息对回复进行评价是本问题关注的重点。根据 Zhu[19]等提出的影响在线问答社区答案质量评价模型如表 3.5 所示，结合题中所给数据和题目要求，本文将特征划分为文本特征、时序特征两大类，每一类下均有详细的特征表示如表 3.6。

表 3.5 信息质量评价影响因素模型

维度	概述
信息量（Informativeness）	答案提供了适当的信息量
礼貌性（Politeness）	尊重别人的观点和情感
完整性（Completeness）	完整的答案
易读性（Readability）	答案清晰
相关性（Relevance）	答案和问题属于同一个主题
简明性（Conciseness）	答案结构紧凑
可信度（Truthfulness）	答案值得信赖
细节性（Level of Detail）	合适的粒度
原创性（Originality）	答案不是从其他来源复制的
客观性（Objectivity）	答案公正，没有偏见
创新性（Novelty）	答案新颖，具有创新性
实用性（Usefulness）	答案具有实用价值
权威性（Expertise）	答案来源与专家

表 3.6 两类特征

特征类型	特征	概述
文本特征	主题相关性	问题与答案主题相关性
	文本长度	回答的句子长度
	标点符号比重	回答中是否大量使用表情或者省略号等
	内容词密度	问题或回答中实词如名词、动词、代词等出现的比例
	内容词覆盖率	问题中的实词和回答中的实词的重复比例
时序特征	答复及时性	答复在问题提出后及时

		作答

（1）主题相关性

主题相关性描述的是问题文本和答复文本之间主题的相关性，一般问题主题和答案主题越相关，则该答复质量越高，评价得分也越高。本文中度量主题相似性主要用到 TF-IDF 和余弦相似度算法。首先利用 TF-IDF 将问题和答复转化为两个空间向量，将文本主题相关性度量问题转化为空间向量相似性度量问题。假如一个词语在一个文章中出现了 n 次，文章中的词语总数位 N，则：

$$词频(TF)=\frac{某个词在文章中的出现次数n}{文章的总词数N}.$$	(3-5)
$$逆文档频率(IDF)=\log\left(\frac{语料库的文档总数}{包含该词的文档数+1}\right)$$	(3-6)
$$TF-IDF=词频(TF)\times 逆文档频率(IDF)$$	(3-7)

可以看出，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语料库中的出现次数成反比，通过计算 TF-IDF 值将问题和答复转化位空格间向量。而余弦相似度算法是指一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近 1，夹角趋近 0，表明两个向量越相似，余弦值接近于 0，夹角趋近于 90 度，表明两个向量越不相似。若 a 和 b 分别为问题和答复向量化后的向量，则问题和答复的相关性为：

$$similarity(a,b)=\cos\theta=\frac{a\bullet b}{|a|\times|b|} \tag{3-8}$$

计算得到问题和答案的相关性值处于[0,1]之间，值越接近于 1，说明问题和答复的相关性越大；值越接近于 0，说明问题和答复越不相关。

（2）文本长度

本文统计的文本长度是答复的句子长度。答复越长包含的内容信息越丰富，通过 python 语言统计发现，答复文本字符长度最小值为 3，最大值为 7883。为了避免个别极端数值对后续评价模型的影响，根据下列公式对句子长度 L 计算得分 Score_L。

$$Score\_L=\begin{cases}1 & L\geq 1000\\ 0.8 & 700\leq L<1000\\ 0.6 & 400\leq L<700\\ 0.4 & 100\leq L<400\\ 0.2 & 0\leq L<100\end{cases} \tag{3-9}$$

（3）标点符号比重

标点符号比重指答复中标点符号长度与文本长度的比例。该项指标反映了回答中是否大量使用省略号或其他无意义字符。统计发现标点符号比重 S 位于[0.07,0.51]之间，标点符号占比越小说明文本内容信息越丰富，因此该项指标为负向指标，通过下列公式实现将标点符号比重 S 向标点符号比重得分 Score_S 的标准化映射。

$$Score\_S = \frac{\max(S) - S}{\max(S) - \min(S)} \tag{3-10}$$

（4）内容词密度

内容词密度指回答中实词如名词、动词、代词等出现的比例。根据 jieba 词性标注结果含义（表 3.2），去掉下表中非实词则剩下的词语为实词。

表 3.7 Jibe 词性标注非实词类别及含义

c	连词	取英语连词 conjunction 的第 1 个字母。
e	叹词	取英语叹词 exclamation 的第 1 个字母。
k	后接成分	
o	拟声词	取英语拟声词 onomatopoeia 的第 1 个字母。
u	助词	取英语助词 auxiliary
w	标点符号	
y	语气词	取汉字“语”的声母。
z	状态词	取汉字“状”的声母的前一个字母。
un	未知词	不可识别词及用户自定义词组。取英文 Unkonwn 首两个字母。(非北大标准，CSW 分词中定义)

假设答复中词语总数为 n，其中实词的数量为 n_r，则内容词密度得分 Score_R 为：

$$Score\_R = \frac{n_r}{n} \tag{3-11}$$

（5）内容词覆盖率

内从此重复率指问题中的实词和回答中的实词的重复比例。假设问题和答复中的词语总数分别为 n_1 和 n_2，其中 n_1 和 n_2 重复的词语个数为 n_c，则内容词覆盖率 r_c 如公式 3-8 所示，内容词覆盖率为正向指标，将其按照公式 3-9 映射为内容词覆盖率得分 Score_C

$$r_c = \frac{n_c}{\max(n_1, n_2)} \tag{3-12}$$

$$Score\_C = \frac{n_c - \min(n_c)}{\max(n_c) - \min(n_c)} \tag{3-13}$$

（6）答复及时性

答复及时性指相关部门答复问题和群众提出问题的时间间隔。若间隔时间较短说明问题处理效率较高，公众更可能有更高的满意度。统计发现间隔时间 T（天）处于[0.02,1160.44]之间，参考相关政务处理时间，将间隔时间数据 T(天)按照下列公式转化为答复及时性得分 Score_T

$$Score\_T=\begin{cases}1 & T<3\\ 0.8 & 3\le T<7\\ 0.6 & 7\le T<15\\ 0.4 & 15\le T<30\\ 0.2 & T>30\end{cases} \tag{3-14}$$

3.4.2 答复评价指标及权重

（1）答复评价指标体系

根据问题三的要求，针对附件 4 相关部门留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见给出一套评价方案。根据评价准则和上节影响因素研究，建立指标体系（如图 3.3），结合查阅资料和层次分析法构造比较判断矩阵并求出相应权重。

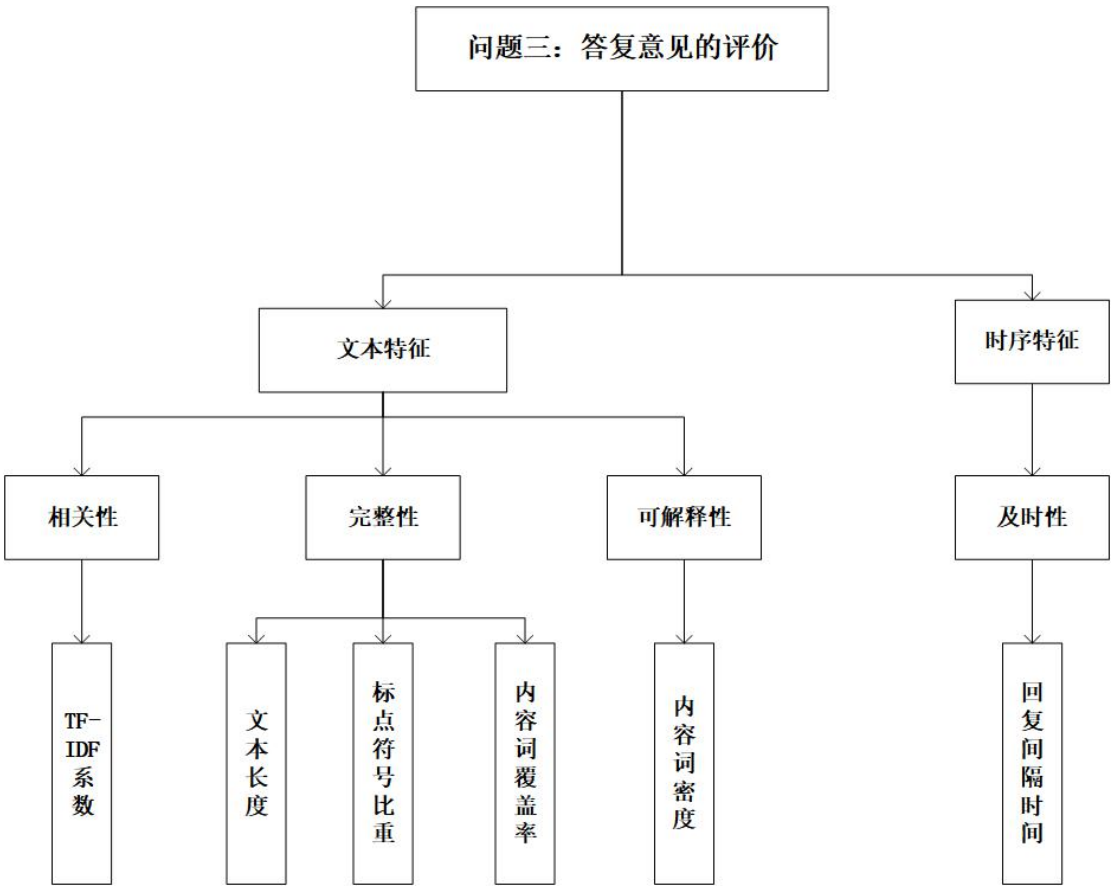

图 3.3 答复评价指标体系

（2）构造比较判断矩阵并求权重

综合考虑指标数据情况和相关资料，将文本特征和时序特征的权重定为 0.7 和 0.3。

1）指标层 1 判断矩阵 A

重要性标度表如下所示：

表 3.8 重要性标度表

重要性标度	含义
1	表示两个元素相比，具有同等重要性
3	表示两个元素相比，前者比后者稍重要
5	表示两个元素相比．前者比后者明显重要
7	表示两个元素相比，前者比后者强烈重要
9	表示两个元素相比，前者比后者极端重要
2,4,6,8	表示上述判断的中间值
倒数	若元素 i 与元素 j 的重要性之比为 a_{ij}，则元素 j 与元素 i 的重要性之比 $a_{ji}=1/a_{ij}$

指标层 1（完整性、相关性、可解释性）判断矩阵 A 如下所示：

$$A=\begin{bmatrix}1 & 1/2 & 1\\ 2 & 1 & 3\\ 1 & 1/3 & 1\end{bmatrix} \tag{3-15}$$

2）指标层 2（完整性、相关性、可解释性）判断矩阵 B

时序特征只有一个指标，因此不用构建判断矩阵，文本特征判断矩阵构造如下。文本特征（相关性、完整性、可解释性）如下所示：

$$B=\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \tag{3-16}$$

3）利用规范列平均法求权重

首先对列向量归一化得到对应矩阵，再对其按行求和并归一化，得到特征向量 w_i。

$$w_A = [0.333 \quad 0.333 \quad 0.333]^T \tag{3-17}$$

$$w_B = [0.241 \quad 0.548 \quad 0.210]^T \tag{3-18}$$

（3）一致性检验

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，其对应关系如表 3.9：

表 3.9 平均随机一致性指标 RI 标准值

（不同的标准不同，RI 的值也会有微小的差异）

矩阵阶数	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

考虑到一致性的偏离可能是由于随机原因造成的，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 3-19。计算得到矩阵 A 的 CR 等于 0<0.1，满足检验；矩阵 B 的 CR 等于 0.016<0.1 满足检验；

$$CR = \frac{CI}{RI} \tag{3-19}$$

因此得到指标 1 权重（完整性、相关性、可解释性）的权重为 w 为（0.548,0.241,0.210）。

指标 2 完整性权重（文本长度、标点符号比重、内容词覆盖率）的权重为（0.1278，0.1278，0.1278）

3.4.3 答复综合评价模型

基于以上研究结果，构建答复意见综合评价模型如下：

表 3.10 答复意见综合评价模型

目标层	准则层	指标层 1	指标层 2	权重
答复意见综合评价模型	文本特征（0.7）	完整性（0.548）	文本长度（0.084）	0.1278
			标点符号比重（0.172）	0.1278
			内容词覆盖率（0.220）	0.1278
		相关性（0.241）	TF-IDF 系数（0.25）	0.1687

		可解释性（0.210）	内容词密度（1）	0.147
	时序特征（0.3）	及时性（0.3）	回复间隔时间（0.3）	0.3

根据评价模型和相应权重，可得到最终答复质量综合评价得分

$$y = \sum_{j=1}^{n} (\sum_{i=1}^{n} x_i * w_i) * w_j \tag{3-20}$$

其中 x_i 是指标层 1 中指标 j 的 2 级指标中各指标得分，w_i 为其对应权重，w_j 为指标层 1 中各指标权重。

四、实验结果

4.1 留言分类结果分析

本文所有的实验都是在普通 PC(Intel(R) Core(TM) i5-8300H CPU,8GB RAM)上完成，实验软件环境为 Python3.6、tensorflow1.14.0、gensim3.8.2。

分类模型的性能一般从准确度和复杂度两个方面来度量。文本分类中一般用 F1 值来衡量分类模型的有效性，同时本文选用运行时间来衡量时间复杂度。

（1）对于模型准确度，从表 4.1 和 4.2 可以看出，针对本文的网络留言数据，相比于 Word2vec，TF-IDF 是一种更加高效准确的特征提取方法，而本文选用的 SVM 模型在训练样本有限时表现出色，F1 值达到 0.884，由于其他所有分类模型。而深度学习模型 CNN 太过依赖于训练样本量，所以在给定的有限样本量下表现较差；

（2）对于模型时间复杂度，结合表 4.3 可以看出，SVM 模型时间复杂度尚在可接受范围内，在本文分类任务中，时间复杂度和准确度都远优于 CNN 和 Xgboost 模型。若对 F1 值没有严格要求，Navie Bayes、Logistic Regression 也基本能完成网络留言分类任务。

表 4.1 分类模型 F1 值比较

模型	**TF-IDF+ CNN**	**Word2Vec +SVM**	**TF-IDF+Logistic Regression**	**TF-IDF+ Xgboost**	**TF-IDF+Naive Bayes**	**TF-IDF +SVM**
F1	0.835	0.871	0.877	0.866	0.846	**0.884**

表 4.2 特征提取模型运行时间

特征提取方法	**TF-IDF**	**Word2vec**
运行时间（s）	7.958	50.954

表 4.3 分类模型运行时间

分类模型	**CNN**	**SVM**	**Logistic Regression**	**Xgboost**	**Naive Bayes**
运行时间（s）	48.695	10.184	1.261	39.080	**0.134**

4.2 热点问题挖掘分析

（1）相似问题归类结果

根据 3.2 节中基于 CRF 模型的地理位置识别和 k-means 聚类分析后，得到 55 类热点问题，详细数据见附件问题二。例如表 4.4 为 40 位用户关于“伊景园捆绑销售车位”问题在网络平台的留言，说明该问题对多数用户生活造成了严重影响，有关部门应该对问题进行调查和处理。

表 4.4　40 位用户关于“伊景园捆绑销售车位”问题的留言

留言编号	留言主题
190337	关于伊景园滨河苑捆绑销售车位的维权投诉
191001	A 市伊景园滨河苑协商要求购房同时必须购买车位
195995	关于广铁集团铁路职工定向商品房伊景园滨河苑项目的问题
196264	投诉 A 市伊景园滨河苑捆绑车位销售
205277	伊景园滨河苑捆绑车位销售合法吗？！
205982	坚决反对伊景园滨河苑强制捆绑销售车位
207243	伊景园滨河苑强行捆绑车位销售给业主
209571	伊景园滨河苑项目绑定车位出售是否合法合规
213584	投诉 A 市伊景园滨河苑定向限价商品房违规涨价
214975	关于房伊景园滨河苑销售若干问题的投诉
218709	A 市伊景园滨河苑捆绑销售车位
218739	A 市伊景园·滨河苑欺诈消费者
220534	投诉武广新城伊景园滨河苑为广铁集团的定向商品房
222209	A 市伊景园滨河苑定向限价商品房项目违规捆绑销售车位
223247	投诉 A 市伊景园滨河苑捆绑销售车位
224767	伊景园滨河苑车位捆绑销售！广铁集团做个人吧！
230554	投诉 A 市伊景园滨河苑捆绑车位销售
234633	无视消费者权益的 A 市伊景园滨河苑车位捆绑销售行为
236301	和谐社会背景下的 A 市伊景园滨河苑车位捆绑销售
239032	请维护铁路职工权益取消伊景园滨河苑捆绑销售车位的要求
244243	关于伊景园滨河苑捆绑销售车位的投诉
244342	投诉 A 市伊景园滨河苑定向限价商品房违规涨价

244528	伊景园滨河苑开发商强买强卖！
246407	举报广铁集团在伊景园滨河苑项目非法绑定车位出售
251601	A 市伊景园滨河苑诈骗钱财
251844	投诉伊景园滨河苑项目违法捆绑车位销售
255507	违反自由买卖的 A 市伊景园滨河苑车位捆绑销售行为
258037	投诉伊景园滨河苑捆绑销售车位问题
258386	A 市伊景园滨河苑欺压百姓
260254	投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位
268299	惊!! A 市伊景园滨河苑商品房竟然捆绑销售车位
268626	A 市伊景园滨河苑坑害购房者
268920	武广新城伊景园滨河苑商品房霸王购房规定
276460	A 市伊景园滨河苑捆绑销售车位是否合理？
279070	投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位
283879	A 市伊景园滨河苑项目捆绑销售车位
285897	武广新城伊景园滨河苑违法捆绑销售车位,求解决
286304	无视职工意愿、职工权益的 A 市伊景园滨河苑车位捆绑销售行为
289588	投诉 A 市伊景园.滨河苑开发商
289950	投诉 A 市伊景园滨河院捆绑销售车位

（2）问题热度评价结果

根据 3.2 节中热度计算模型，针对问题 2 的留言明细表，计算得到各热点问题的热度指数，其中热度排名前五的问题如下表 4.5 所示，关于热点问题热度指数详细数据见附件问题二“热点问题表.xls”和“热点问题留言明细表.xls”。从该表中可以看出，“A 市金毛湾入学问题”、“A5 区群租房泛滥”、“A 市 58 车贷特大集资诈骗案”等问题是公众关心的重点问题，针对该问题的留言人数、点赞数都较多，也是有关部门需要迫切解决的问题。

表 4.5 热度排名前五的热点问题

热度排名	问题 ID	热度指数	地点人群	时间范围	留言描述
1	55	0.744	梅溪湖金毛湾的一名业主	2019/4/11	反映 A 市金毛湾配套入学的问题一直没有

					得到解决
2	28	0.627	A 市 A5 区汇金路五矿万境 K9 县小区业主	2019/1/13 至 2019/8/19	小区群租房泛滥，路口没有人行天桥和地下通道，物业未尽责任，交通不安全等
3	22	0.232	A 市 58 车贷受害人	2019/1/11 至 2020/7/8	A 市 58 车贷特大集资诈骗案保护伞
4	9	0.186	A 市 A2 区丽发新城小区业主	2019/4/10 至 2019/2019/12/27	A2 区丽发新城附近建搅拌站噪音扰民，污染环境
5	27	0.166	A4 区绿地海外滩小区业主	2019/1/30 至 2019/9/4	当前的高铁规划，A 市绿地海外滩小区会饱受噪音困扰

4.3 回复意见评价分析

以附表 4 中 2816 条答复留言为例，代入 3.4 节中答复综合评价模型进行计算，可得到每条答复评分。评分结果分布情况如图 4.1 所示，其中评分最高的三条答复和评分最低的三条答复如表 4.6 和 4.7 所示。更多详细评分数据见附件问题三“评分结果.xlsx”。

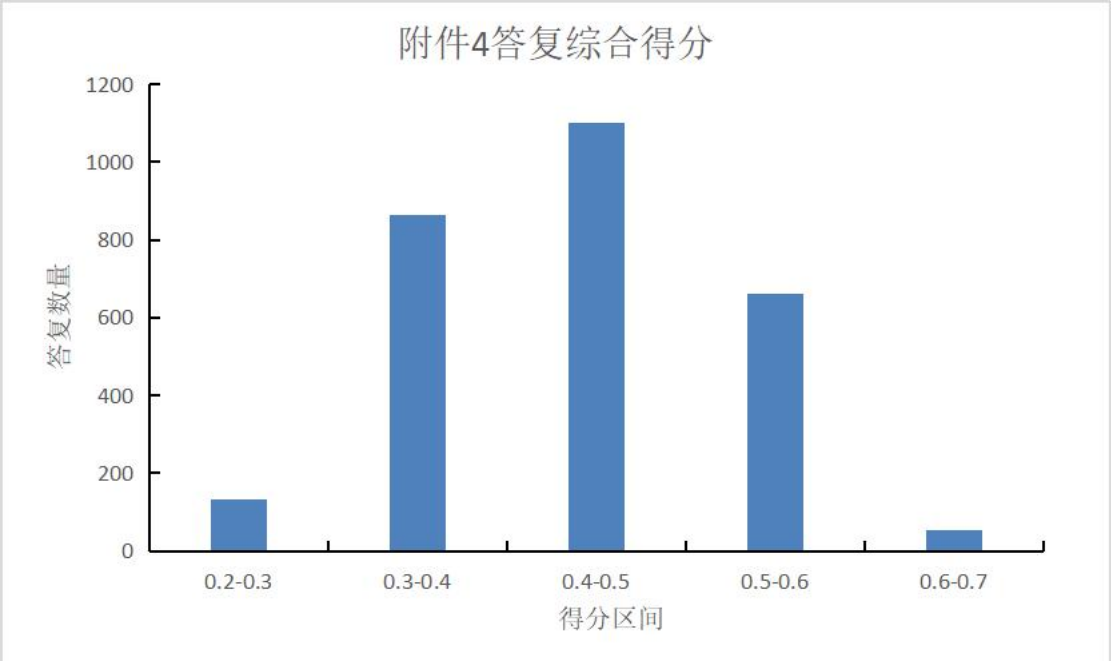

图 4.1 附件 4 回复数据评分结果

表 4.6 评分最高的三条回复意见

网友“幸福 A7 县”您好！来信收悉。现回复如下：感谢关注 A7 县网络电视台的星沙新闻，最近因为 A7 县电视台相关网络出现断网现象，可能导致上传新闻有断网现象，

所以不能正常播放。A7 县网络电视台正在努力升级改造中，敬请持续关注！A7 县电视台 2015 年 5 月 29 日
关于《J11 市康顺液化气站罐装气缺斤少两》的回复　　首先，非常感谢市民在《问政西地省》反映的《 关于 J11 市康顺液化气站罐装气缺斤少两问题》，感谢对我局工作的关注、关心和支持！　　针对《 关于 J11 市康顺液化气站罐装气缺斤少两问题》，我局积极行动，接到投诉当天上午立即组织检定人员前往 J11 市康顺液化气站对其电子罐装秤和电子公平秤进行现场检定，共检定电子罐装秤 10 台/件，电子公平秤 2 台/件，共计 12 台/件，均检定合格。J11 市康顺液化气站充装时由电子罐装秤自动计量控制，气瓶重量不同，充气后的重量也会有所不同。如果用户在 J11 市康顺液化气站购买液化气时对重量有质疑，可在气瓶充气后利用公平秤对气瓶进行现场复秤；若发现有违规行为，可保存好物品原状进行投诉，一经查实将严肃处理。在下阶段，我局将继续对液化气站罐装气缺斤少两问题进行专项检查，同时，欢迎广大市民发现问题及时向我局举报，举报电话：0735---3323506。也欢迎广大市民对我局工作的监督。 2016 年 2 月 26 日　　我所于 2016 年 2 月 23 日接到由办公室转发的平台上关于 J11 市康顺液化气站罐装气缺斤少两的投诉，当天上午立即组织检定人员前往 J11 市康顺液化气站对其电子罐装秤和电子公平秤进行现场检定，共检定电子罐装秤 10 台/件，电子公平秤 2 台/件，共计 12 台/件，均检定合格。　　据悉，J11 市康顺液化气站充装时由电子罐装秤自动计量控制，气瓶重量不同,充气后的重量也会有所不同。如果用户在 J11 市康顺液化气站购买液化气时对重量有质疑，可在气瓶充气后利用公平秤对气瓶进行现场复秤；若发现有违规行为，可保存好物品原状进行投诉，一经查实将严肃处理。 2016 年 2 月 25 日
雪峰山水库是我市五座中型水库之一，集雨面积 16.57 平方公里，总库容 1160 万方，承担农田灌溉和农村供水功能。网友反映水库剩下不到小半湖黄泥巴水情况基本属实。一、主要原因：在安全生产检查中，发现雪峰山水库隧洞进口控制闸门出现锈蚀、破损和钢丝索锈蚀老化现象，启闭机不能正常启闭，造成出口闸阀长期处于中高水位运行状态，存在严重安全隐患，危及大坝安全。根据省水利厅和 B 市水利局相关文件精神，2018 年 10 月我市编制了《B9 市公益性水利工程维修养护项目（雪峰山水库）实施方案》。2018 年 10 月 15 日，B 市水利局下达了《关于 2018 年全市公益性水利工程维修养护项目实施方案的批复》（株水函【2018】154 号），批复同意我市雪峰山水库闸门进行维修。闸门维修计划 2018 年底开工建设，考虑临近春节，如果放空库水，将会影响雪峰山水厂春节供水，带来诸多影响，故闸门维修工程推迟到春节后实施。拟定在库内围堰方案，围堰高度不低于 2 米，对应库内水位（高程）121.0 米，相应库容 75.5 万方。工程工期为 2019 年 3 月 5 日--3 月 25 日。为给工程施工创造条件，雪峰山水库从 2019 年 2 月 11 日起开始放水，直至 3 月 2 日，雪峰山水库蓄水距离低涵高程 3.5 米，水面开始有黄泥巴水现象。由于近段时间长时间降雨，入库水流冲刷淤泥，加上水库水位不断下降，自净能力减弱，浑水现象不断扩大，出现了网民所反映的现象。二、采取的处理措施由于雪峰山水厂供水范围广，影响大，3 月 7 日下午 13 点我局主要负责人召集白兔潭、浦口镇、雪峰山水厂、抢险施工单位负责人进行调度。会议一致确定采取以下措施：1、市水利局负责向社会公布抢险通告；2、雪峰山水厂负责向用户发出“致用户的一封信”；3、要求雪峰山水厂加强管理，优化制水工艺，加大检测密度，提升供水水质，4 月 5 日恢复正常供水；4、镇政府负责对涉及供水的村组和企事业单位做好信息通报和安全用水引导工作；5、抢险施工单位负责科学组织施工，抢抓工期，确保在 3 月 25 日前完成抢险任务；6、责成雪峰山水管所加强日常巡查，确保入库水质正常稳定。非常感谢网友对水利工作的关心，我们将进一步改进工作，提升工作水平，确保农村居

民饮水安全！2019 年 3 月 7 日网友您好：您反映的问题已收悉并交有关单位调查核处，如有情况将及时回复！2019 年 3 月 6 日我院在收到 B9 市浦口镇雪峰山水厂的情况反馈后，指派我院民行科同志至 B9 市水利局、B9 市河长办、B9 市环境保护局、雪峰山水库管理所等相关部门了解情况，通过询问上述行政机关相关负责人后，核实了相关情况，现将调查情况回复如下：一、B9 市浦口镇雪峰山水库确实存在黄泥水的情况，举报人当时的举报客观情况属实。二、根据我院调查了解，B9 市水利局下属雪峰山水库管理所在进行例行安全检查过程中，发现雪峰山水库隧洞进口控制阀门出现锈蚀、破损及钢丝索锈蚀老化现象，启闭机不能正常启闭，造成出口闸阀长期处于中高水位运行状态，存在严重安全隐患，危及大坝安全。B9 市雪峰山水库管理所随即依法按照省市两级文件要求，对安全隐患处进行维修。为给维修工程施工创造条件，雪峰山水库自 2019 年 2 月 11 日开始放水，直至 3 月 2 日，由于水位下降，水面出现黄泥水现象，加之当时降雨较多，雨水将黄泥水冲刷入库，造成雪峰山水库黄泥水，群众反映的现象。该维修工程持续到 2019 年 3 月 25 日结束。B9 市水利局为此专门在平台百姓呼声 B9 市站对此事进行了官方回复。三、我院民行干警再次到雪峰山水库实地查看，水库维修工程已经结束，水库水位已经恢复，雪峰山水库黄泥水现象已经消除。2019 年 4 月 22 日

表 4.7 评分最低的三条回复意见

“UU0081893”您好！您所反映的问题已收悉，并转交教育部门办理。谢谢！
已收悉
您的留言已经收到！并已将您反映的情况转交相关部门进行处理，处理情况会向您反馈，感谢您的留言！

分析图表结果可以看出：

（1）得分结果整体呈正态分布。得分高于 0.4 即可认为答复良好，大部分回复得分情况良好，说明网络留言平台对公众参与问政和政府服务效率提高存在积极的促进作用。为了区别不同回答的评价得分，定制评分标准时较为严格，所以导致答复平均分较低。但存在少量的回复得分极低，得分极低的回复在完整性、相关性、及时性等方面都表现较差，对于解决群众提出问题帮助不大，需要引起相关部门关注，核实公众提出问题是否得到解决。

（2）得分高的留言都具有回答及时、答复完整、与留言中提出问题相关性较大等特征，能较好地解决留言中提出的问题。而评分较低的回复意见蕴含有效信息较少、回复时间间隔太久，不利于留言问题的解决。这也从侧面印证了文中所建立的综合评价模型的有效性，基本能实现对答复的完整性、相关性、可解释性等方面的定量度量。

五、总结与展望

5.1 实验结论

本文基于网络平台公众留言数据，通过文本清洗、文本分词、词性标注等预处理过程，选用 TF-IDF 完成文本信息的特征提取并建立了政务留言文本分类模型、留言热点问题挖掘及热度评价模型和政务留言答复评价模型并进行评分计算，主要结论有：

（1）本文在对文本信息预处理时，主要基于 python 语言使用了 jieba 分词和词性标注、TF-IDF、Word2Vec 等文本特征提取模型。实验结果表明，基于 python 语言和各类开源工具包对中文自然语言处理十分友好，能够高效准确地完成各类数据挖掘任务。且通过对比计算复杂度和准确率发现，TF-IDF 比 Word2Vec 更适用于本题所给附件留言文本，保证较小时间复杂度的同时能够更好地提取文本潜在特征。

（2）针对问题一，本文综合考虑了分类模型准确度、时间复杂度和附件数据的特点，建立了基于 TF-IDF 和 SVM 的文本分类模型，其 F1 值为 0.884，优于同类机器学习方法和 CNN 模型，且时间复杂度优于 CNN 和 Xgboost。最大程度地保持了算法自动化、智能化，具有很强的适用性和自我学习性。通过对比实验发现，针对经过 TF-IDF 特征提取后的网络留言数据，Naive Bayes、Logistic Regression 在分类准确度和时间复杂度上都表现良好，基本能满足文本分类要求。

（3）针对问题二，首先基于 CRF 和自定义地理位置词典对留言信息中地理位置实体进行识别，然后根据 kmeans 算法实现特定地点问题归类实现热点问题挖掘，最后通过查阅资料和已有数据，确定可利用的留言量、留言用户数、平均点赞及反对量、单条留言最高点赞量及反对量等评价指标，权重分别为 0.123，0.123，0.313，0.439，计算热度得到“梅溪湖金毛湾配套入学”、“A 使 A5 区群租房泛滥”、“A 市 58 车贷特大集资诈骗”、“A 市 A2 区噪音扰民”、“A4 区噪音扰民”等公众普遍反映的问题。希望政府部门能尽快给予关注并采取相关措施落实解决问题，更好地服务于公众生活。同时可以发现单条留言最高点赞量及反对量权重最大，较能反映问题热度信息，政府相关部门可根据单条留言最高点赞量及反对量排序，粗略筛选出热度信息，对公众反映热点问题相关内容进行了解。

（4）针对问题三，综合考虑问题中相关性、完整性、可解释性要求，并通

过查阅相关资料在评价准则中加入及时性，通过层次分析法确定四者的权重分别为 0.241、0.548、0.210、0.3，并转化成可量化指标文本长度 0.1278，标点符号比重 0.1278，内容词覆盖率 0.1278，TF-IDF 系数 0.1678，内容词密度 0.147，回复间隔时间 0.3，据此对附件 4 中 2816 条数据进行评价。全部答复评价得分位于[0.2326，0.6666]之间，通过比较得分较高和得分较低答复留言发现，得分价高的答复留言及时清楚完整地对反映问题的处理情况和结果进行了解释和回答，而得分较低的留言回复较笼统模糊，对公众满意度和积极性有一定的负面影响。

基于以上结论，建议公众在网络问政平台留言时尽量规范、准确地描述反映问题的时间、地点和问题内容。建议政府服务部门可以用文中建立的地理位置实体识别、热点问题发现和问题热度评价模型对大量的网络留言信息进行归类，及时集中同一处理。同时政府答复留言评价可以作为公众监督政府工作、政府工作自我评价的依据，针对评分较低的答复，政府部门应该重新落实核查反映问题是否解决，并给予公众更加积极的反馈，以提高政府部门服务效率和公众参与满意度。

5.2 感想与展望

通过本次对网络问政平台留言文本信息挖掘和完成此份报告的过程中，小组成员思考总结了所作工作、面临问题和未来研究方向，主要可以概括为以下几点：

（1）通过竞赛期间对 C 题的研究，小组成员在阅读了大量文献及亲自实践中了解、完善并建立了政务留言文本分类模型、留言热点问题挖掘及热度评价模型、政务留言答复评价模型。解决该问题的过程中经历了完整的数据预处理、规律发现、建立模型、评分分析的过程，不仅是完成 C 题三个要求，也是对自身能力的一次提升。通过对网络问政平台的留言文本信息挖掘，我们可以快速发现公众反映的普遍问题并进行归类，有助于政府服务部门集中及时处理相关问题，对于政府回复意见的评价也从侧面反映了政府服务部门工作效率。比赛结束后也会持续关注相关研究，并如果有机会将继续开展该方面课题的研究。

（2）由于硬件设备、比赛时间和所给数据限制等原因，针对深度学习方法，本文仅尝试两层 CNN 神经网络，未对其他深度学习分类方法和神将网络内部结构、参数设置等方面进行深入探究，该部分可作为未来研究方向，比如可尝试 Xgboost 与神经网络的结合，通过对比实验研究是否能够取得更高的 F1 值。同

时网络问政平台可以考虑接入更多源的数据，比如用户注册信息（年龄、居住地址、职业）、在用户留言界面将反馈问题的时间地点和问题类别设置为可选择的结构化信息，可以显著提高文本相似性分类模型性能、简化建模过程，从而有助于政府服务部门更加准确高效地发现问题并解决问题。

（3）本次研究充分体现了团队合作、集思广益、优势互补的重要性，当今社会是寻求合作共赢的时代，一个人的思路与能力始终是有局限性的，在合作中实现创新，在过程中寻找到快乐，在感悟中得到升华。小组成员在面对问题时集体思考、各展所长、互帮互助，最终使报告得以保质保量地完成，同时通过此次比赛，个人能力和团队友谊均有所增长。

人工智能在大数据时代取得了巨大突破，并日益在社会生活的各个领域展露头角。随着技术的发展，人工智能的应用领域还将进一步扩大，所带来的经济社会效益也将进一步提升。对于政务服务领域而言，人工智能具有广阔的应用前景，同时也存在系列问题有待解决。总之从长远来看，随着人工智能技术本身的不断成熟，人工智能技术将在智慧政务领域发挥出应有的作用，为政府部门提供辅助决策、态势分析等作用，提高政府部门的公共服务效率和公众办事满意度。

致谢

值此报告完成之际，首先感谢我们小组的导师。本次实验的成果离不开导师耐心的指导和监督，感谢导师为我们研究思路和研究方法提供的帮助。

同时感谢大赛组委会，给予我们这次前进创新的机会和展示的平台，我们将会持续关注智慧政务、文本挖掘、语义分析等方面的研究，注意收集相关数据，为后续更深入的研究做好准备工作。

感谢自然语言处理中各种开源数据包，包括 jieba 分词、word2vec、tensorflow 平台、scikit learn 等报告中使用到的工具包，高效简洁的 python 语言和封装好的工具包使我们在实现文本挖掘目标的同时，省去了大量重复枯燥的基本功能实现，让我们更能专注于问题解决和算法逻辑。

最后感谢小组的每一位成员，尤其是疫情期间沟通诸多不便，所有人在繁忙的研究所科研工作之余，牺牲节假日和平时休息的时间坚持线上交流讨论来开展相关工作，才让此次实验和这份报告得以完成，是我们努力的付出换回了今日的成果。

参考文献

[1] CAMBRIA E, WHITE B. Jumping NLP curves: a review of natural language processing research[J]. IEEE Computational Intelligence Magazine, 2014, 9(2):48-57.

[2] 张坤,王文韬,谢阳群.机器学习在图书情报领域的应用研究[J].图书馆学研究,2018(1):47-52.

[3] ALGHOBIRI M. A comparative analysis of classification algorithms on diverse datasets[J]. Engineering, Technology & Applied Science Research,2018,8(2):2790-2795.

[4] LIU Y, BI J W , FAN Z P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm[J]. Information Sciences, 2017, 394:38-52.

[5] 张海涛,王丹,徐海玲，等.基于卷积神经网络的微博舆情情感分类研究[J].情报学报,2018,37(7):695-702.

[6] 汪海燕,黎建辉,杨风雷.支持向量机理论及算法研究综述[J].计算机应用研究,2014,31(5):1281-1286.

[7] 刘晓亮,丁世飞,朱红，等.SVM 用于文本分类的适用性[J].计算机工程与科学,2010,32(6):106-108.

[8] 田志龙,陈丽玲,顾佳林.我国政府创新政策的内涵与作用机制:基于政策文本的内容分析[J].中国软科学,2019(2):11-22.

[9] 于良芝,李亚设,权昕.我国乡镇图书馆建设中的话语与话语性实践——基于政策文本和建设案例的分析[J].中国图书馆学报,2016,42(4):4-19.

[10] 魏伟,郭崇慧,陈静锋.国务院政府工作报告(1954—2017)文本挖掘及社会变迁研究[J].情报学报,2018,37(4):406-421.

[11] 朱青,卫柯臻,丁兰琳等.基于文本挖掘和自动分类的法院裁判决策支持系统设计[J].中国管理科学,2018,26(1):170-178.

[12] 伍洋,钟鸣,姜艳，等.面向审计领域的短文本分类技术研究[J].微电子学与计算机,2015,32(1):5-10.

[13] 孟天广,李锋.网络空间的政治互动:公民诉求与政府回应性——基于全国性网络问政平台的大数据分析[J].清华大学学报：哲学社会科学版,2015,30(3):17-29.

[14] 杨槟泽,李长军.市长公开电话文本自动分类技术比较研究[J].中国海洋大学学报：自然

科学版,2017,47(S1):173-177.

[15] 马宝君,张楠,谭棋天.基于政民互动大数据的公共服务效能影响因素分析[J]. 中国行政管理,2018(10):109-115.

[16] 李莉,孟天广.公众网络反腐败参与研究——以全国网络问政平台的大数据分析为例[J].中国行政管理,2019(1):45-52.

[17] 何哲.面向未来的公共管理体系:基于智能网络时代的探析[J].中国行政管理,2017(11):100-106.

[18] 陈涛,冉龙亚,明承瀚.政务服务的人工智能应用研究[J].电子政务,2018(03):22-30.

[19] 张仰森,郑佳,唐安杰.基于多特征融合的微博用户权威度定量评价方法[J].电子学报,2017,45(11):2800-2809.