**选题报告**

**论文题目：“智慧政务”中的文本挖掘**

**组 长** 郑鸣宇2021111874

**组 员** 叶尔古力2021110131

努尔比耶姆2021112010

周松烨2021111733

**学 院**  信息管理与工程学院

**专业班级** 信息管理与信息系统

**日 期** 2023.12.2

|  |
| --- |
| 【研究背景与意义】 |
| 近年来，随着“互联网+政务”服务的深入推动，网络问政平台如市长信箱、民意留言板、阳光热线等已逐渐成为政府了解民情、倾听民声、感知民意、汇集民智的重要渠道。同时，伴随着大数据、云计算、人工智能尤其是语言智能等技术的持续突破，构建基于自然语言处理（Natural Language Processing, NLP）技术的智慧政务系统已经成为社会治理创新发展的紧迫需求和新趋势。运用NLP技术，批量、智能、高效地处理大量的政务文本，进而构建智能化的电子政务系统，已经成为服务型政府建设的一个重要课题。这对于提高政府的施政效率和治理水平，增强人民群的福感，促进社会和谐，都具有极大的积极影响。 |
| 【研究问题】 |
| “智慧政务”文本挖掘的目标主要包含两个部分，分别是（1）群众留言的自动分类；（2）群众留言的热点话题发现。   1. **群众留言的自动分类**   在处理网络问政平台的群众留言时，工作人员首先按照预设的分类体系对留言进行归类，这样才能将大量的留言分派给相应的职能部门进行处理。然而，目前大部分的电子政务系统仍然依赖于人工经验进行处理，这不仅工作量大，效率低，而且错误率高。因此，利用自然语言处理中的文本分类技术建立一个能够对群众留言进行分类的一级标签分类模型实现留言自动分类，不仅能极大地减轻政务工作人员的负担，也能提高政务服务的质量和效率。  针对问题一：BERT模型是一种预训练的深度神经网络模型，由Google团队于2018年提出，基于Transformer架构，旨在解决NLP中的语言表示问题。该模型采用了自编码器（AutoEncoder）的预训练思想，先在大规模语料库上进行无监督的预训练，再在具体任务上进行微调，从而达到很好的效果。因此问题一基于bert-base-chinese的预训练模型，解决文本分类问题。   1. **群众留言的热点话题发现**   在海量的群众留言中，存在着许多反映共同问题、表达共同诉求的留言，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。因此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。  针对问题二：留言的点赞数和反对数可以再一定程度上反映该话题的关注程度。因此该问题初计划将一条留言的点赞数和反对数之和作为留言的关注度量化评分。除此之外，留言的内容也是很重要的因素，重复出现内容相近的留言也能说明话题具有热点。故计划综合两方面指标计算热度指数。数据的预处理采用jieba分词，词频统计等方法，聚类则考虑采用K-means，DBSCAN，凝聚层次聚类，高斯混合聚类这些常见的聚类方法。 |
| 【数据描述】 |
| 本文共有三个留言数据集，每一个问题对应一个数据集，另有一个数据集提供了留言的分类标签。数据集共有 9210 条留言。数据集内共有 7种一级标签，其中城乡建设共 2009 条，环境保护共 938 条，交通运输共 613 条，商贸旅游共 1215 条，卫生计生共 877 条，教育文体共 1589 条，劳动和社会保障共 1969 条。  该数据包含了如下字段  留言编号：是每条留言的编号，一条留言有一个编号  留言用户：每位网友的账号，一位网友一个账号  留言主题：记录用户所反映的现象  留言时间：用户留言的具体时间  留言详情：记录用户所反映的现象  答复意见：工作人员对相应的留言做出的回应  答复时间：工作人员对留言做出答复的时间  留言点赞数  留言反对数  数据样例如下：   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **留言编号** | **留言用户** | **留言主题** | **留言时间** | **留言详情** | **一级分类** | | 744 | A089211 | 建议增加 A 小区快递柜 | 2019/10/18 14:44 | 我们是 A 小区居民... | 交通运输 |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **留言编号** | **留言用户** | **留言主题** | **留言时间** | **留言详情** | **点赞数** | **反对数** | | 744 | A08921  1 | 建议增加 A 小区快递  柜 | 2019/10/18  14:44 | 我们是 A 小区居  民... | 100 | 2 | |
| 【数据来源】 |
| 2020年“泰迪杯”数据挖掘挑战赛（C题）数据 |
| 【技术路线】 |
| 针对群众留言的自动分类:  设计基于BERT的文本分类模型，模型搭建采用torch,transformers,sklearn 的框架，预训练采用bert-base-chinese模型，数据预处理包括划分数据集，选取（或拼接）数据列，将数据格式转成BERT模型的输入格式等。模型训练则包括参数设置，调节过程。最后计算得分，并可视化。  针对热点问题的发现：  首先基于jieba分词，正则，去除停用词，去除低频词，提取关键词，计算关键词词频等中文文本预处理，然后进行文本特征提取，再采用如kmeans,DBSCAN等聚类方法聚类。也可考虑利用LDA实现主题建模与文本聚类。 |
|  |