项目编号：

**吉林大学“大学生创新创业训练计划”**

**创新训练项目申请书**

项目名称

基于深度学习/大模型的通讯平台信息聚合模型

项目负责人 杨胜宇

所在学院、年级、专业 计算机科学与技术学院

2023级 物联网工程

联系电话 13177802360

电子邮箱 Elicy04@outlook.com

指导教师姓名 郭东伟 职称 教授

填表日期 2025 年 4 月 19 日

**吉林大学教务处、吉林大学创新创业教育学院制表**

**填表须知**

一、**本表适用于创新训练项目**。本科生个人或团队，在校内导师指导下，自主完成创新性实验方法的设计、设备和材料的准备、实验的实施、数据处理与分析、总结报告撰写等工作。

二、申报书请按顺序逐项填写，实事求是，表达明确严谨。空缺项要填“无”。

三、申请参加大学生创新训练项目团队的人数为3—5人。

四、申请项目，必须聘请教师作为指导老师，并请指导教师在申请书中的指导教师意见栏中签署意见。

五、填写时可以改变字体大小等，但要确保表格的样式不变；不得随意涂改；A4纸正反面打印，左侧装订。

六、“项目编号”由教务处、创新创业教育学院填写。

七、申报过程有不明事宜，请与创新创业教育学院双创竞赛办公室联系，电话85166413。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **项目名称** | | |  | | | | | | | |
| 项目起止时间 | | | **年 月 至 年 月** | | | | | | | |
| 负责人 | 姓名 | | 学院 | 专业 | 教学号 | 联系电话 | | E-mail | QQ | 各类实验班 |
| 杨胜宇 | | 计算机科学与技术学院 | 物联网工程 | 21231431 | 1317780  2360 | | Elicy04@outlook.com | 3428790280 | 是□ 否☑ |
| 项目组成员 | 杨启航 | | 计算机科学与技术学院 | 物联网工程 | 21231423 | 1957369  2761 | | 2700539510@qq.com | 2700539510 | 是□ 否☑ |
| 宁欣怡 | | 计算机科学与技术学院 | 物联网工程 | 21231403 | 1593299  4842 | | 2040893162@qq.com | 2040893162 | 是□ 否☑ |
| 郭川嘉 | | 计算机科学与技术学院 | 物联网工程 | 21231408 | 1863563  0496 | | Smithmmo1@outlook.com | 1299364279 | 是□ 否☑ |
|  | |  |  |  |  | |  |  | 是□ 否□ |
| 指导教师1 | 姓名 | | 郭东伟 | | | | | 职务/职称 | | 教授 |
| 所在单位 | | 吉林大学公共计算机教学与研究中心 | | | | | | | |
| 联系电话 | | 13009006800 | | | | E-mail | |  | |
| 对本课题相关领域研究情况 | | 熟悉 | | | | | | | |
| **项目性质** | | 1.小发明、小创作、小设计（）  2.开放实验室或实习基地中的创新性实验或新实验开发（ ）  3.基础性研究（ ）  4.应用性研究（ ☑ ）  5.社会调研 （ ） | | | | | | | | |
| **项目选题来源** | | 1.自主立题（☑）  2.教师科研课题的子项目（ ） | | | | | | | | |
| **项目学科类别** | | 计算机科学与技术 | | | | | | | | |
| **项目受其他渠道资助情况**（填“无”或具体资助来源和经费，包括获奖情况） | | 无 | | | | | | | | |
| **一、立项背景和依据**（包括研究目的、国内外研究现状分析与评价、研究意义，应附主要  **研究目的**  随着互联网技术、大数据等数字化科技的迅速发展，涌现了大量具有通讯功能的网络平台，为人们提供自由交流和便于管理的虚拟通信空间。通讯平台为用户工作和学习生活提供了极大便利，基于通讯平台的信息交流已呈现出越来越不可或缺的趋势。但也存在诸如平台孤岛化，信息过载等弊端，主要表现在用户常难以依赖单一通讯平台获得全面的通讯信息，需要人工对多个平台内的碎片化通讯信息进行选择并分类整合，引起通讯交流和消息收取效率的降低和通讯发生错误的概率提高。在跨平台通讯中，对不同来源的信息进行聚合是提高智能化通讯服务的重要手段。为此，构建一个有效的讯息聚合模型可以具体描述讯息聚合的流程和应用方法，从而实现更加智能和高效的通讯服务。  **国内外研究现状分析与评价**  随着信息时代的到来，信息的快速增长和多样化使得信息管理和处理成为一个重要课题。国内外众多学者对相关领域进行了深入研究，并取得了一定的成果。  **1.跨平台信息聚合的研究分析**  在文本挖掘和消息分类方面，孙佳提出了基于文本挖掘的跨平台网络社区知识聚合模型，通过实证研究验证了该模型的有效性。高洪设计并实现了一种跨平台多媒体信息聚合发布系统，有效地整合了不同平台的多媒体信息，为用户提供了便捷的信息获取方式。该系统通过统一的接口和协议，实现了对不同平台信息的采集、处理和发布，具有良好的跨平台性和扩展性。类似地，高洪也对跨平台多媒体信息聚合发布系统进行了研究，进一步优化了系统的性能和功能。高李政结合类型域识别技术，提出了改进的消息分类方法，提高了跨平台信息聚合的效率和准确性。王红则从通信系统的角度，对消息分类输出方法进行了深入研究，为跨平台信息聚合提供了技术支持。  此外，冉启对跨平台多媒体信息系统进行了详细的分析和设计，提出了一种高效的智能信息平台系统架构，为本项目的开发提供了重要的参考。  但是当前研究所针对的内容更多地聚焦于信息链的下层，即知识情感等更深层次，反而很大程度上忽略了那些虽然简短，信息密度较低，却与所有人日常息息相关的一般通讯信息的聚合和处理，而部分系统的设计也由于市场竞争而无法落实使用。  **2.针对短文本的分类模型研究现状**  冉启所设计的智能社交平台中针对短信息文本引入了深度学习技术和BERT语言模型作为数据处理方法。史文通过融合BERT和自注意力机制的张量图卷积网络来实现文本分类。蔡九则致力于中文推送短消息文本分类技术的研究与实现，为消息分类提供了新的方法和思路。罗基于朴素贝叶斯算法，研究了QQ群信息分类方法，有效地提高了QQ群信息的分类效率和准确性。高李政提出了基于类型域识别的消息分类方法，通过识别消息的类型域特征，实现了对消息的准确分类。王红研究了通信系统消息分类输出方法及控制器，为消息分类提供了硬件和技术支持。  然而，现有研究仍存在一些不足之处。一方面，部分研究侧重于单一平台或特定类型消息的处理，缺乏对多平台消息的综合筛选和整合。另一方面，对于消息的时间段分类、关联整合等功能的研究相对较少，无法满足用户对信息时效性和关联性的需求。同时，如何在保证信息准确性的前提下，实现高效的信息抓取和处理，仍然是一个具有挑战性的问题。  **研究意义**  本项目的研究具有重要的理论和实际意义。理论上，通过对多平台消息的筛选整合和分类标记的研究，将进一步丰富和完善信息管理和处理的理论体系。项目的成果将为跨平台信息聚合和深度学习在信息处理中的应用提供新的方法和思路，推动相关领域的理论发展。  实践上，本项目开发的工具将为用户提供更加便捷、高效的信息管理解决方案，有助于提高用户的信息处理效率和决策能力。通过整合多平台消息，用户可以更全面，高效地了解信息，避免重要消息的遗漏。此外，项目的成果还可以为相关企业和机构提供技术参考，促进信息管理系统的优化和升级，具有广阔的应用前景和市场潜力。  **参考文献**  [1] 孙佳佳. 基于文本挖掘的跨平台网络社区知识聚合模型及实证研究[D]. 东北师范大学,2023.  [2] 蔡九鸣. 中文推送短消息文本分类技术研究与实现[D]. 西安电子科技大学,2019.  [3] 罗杰. 基于朴素贝叶斯的QQ群信息分类方法[J]. 中国新通信,2018,20(10):137-139.  [4] 高李政,等. 基于类型域识别的消息分类方法[J]. 信息工程大学学报,2018,19(01):66-69.  [5] 王红春. 通信系统消息分类输出方法及控制器[P]. 陕西省,2016.  [6] 高洪,等. 跨平台多媒体信息聚合发布系统的设计与实现[J]. 广播与电视技术,2012,39(12):33-37.  [7] 王红春,通信系统消息分类输出方法及控制器.陕西省,中国航空工业集团公司第631研究所,2016-11-01.  [8] 高洪,黄志强,莫晓山,等.跨平台多媒体信息聚合发布系统的设计与实现[J].广播与电视技术,2012,39(12):33-37.DOI:10.16171/j.cnki.rtbe.2012.12.024.  [9] 高李政,罗军勇,刘琰,等.基于类型域识别的消息分类方法[J].信息工程大学学报,2018,19(01):66-69.  [10] 王红春,通信系统消息分类输出方法及控制器.陕西省,中国航空工业集团公司第631研究所,2016-11-01.  [11] 高洪,黄志强,莫晓山,等.跨平台多媒体信息聚合发布系统的设计与实现[J].广播与电视技术,2012,39(12):33-37.DOI:10.16171/j.cnki.rtbe.2012.12.024.  [12]冉启睿.基于DeepFM的智能社交平台系统的设计与实现[D].北京交通大学,2023.DOI:10.26944/d.cnki.gbfju.2023.002125.  [13]史文艺,朱欣娟.融合BERT和自注意力机制的张量图卷积网络文本分类[J].计算机系统应用,2025,34(03):152-160.DOI:10.15888/j.cnki.csa.009831 | | | | | | | | | | |
| **二、项目研究内容**（项目主要研究内容；拟解决的关键问题、重点和难点）  **主要研究内容**   1. **文本主题模型**   现有的主题模型包括BTM（Biterm Topic Model）、HDP（Hierarchical Dirichlet Process）、CTM（Correlated Topic Model）、ATM（Author-Topic Model）、RTM（Relational Topic Model）、ProdLDA（Product of Experts LDA）、ETM（Embedded Topic Model）、GSDMM（Gibbs Sampling Dirichlet Multinomial Mixture）等等。  而当前的常用主题模型是IDN主题模型（经典概率主题模型）和bertopic主题模型（结合深度学习的新兴模型）  lda主题模型出现较早，衍生出了许多变体，通过 Dirichlet 分布建模文档的主题分布及每个主题的词分布。  bertopic主题模型的基本流程是词向量嵌入，umap降维，HDBSCAN聚类，但这并不是一成不变的，也可以用其他代替，例如使用bert进行词向量嵌入，pca主成分分析进行降维，k-means聚类。  主题模型可以作为文本分类的前向工作，有着轻量级部署的特点，如果需要手动标注数据的话，可以根据主题模型和关键词提取的结果作为参考来辅助进行数据的人工标注。  **2.BERT预训练模型**  BERT（Bidirectional Encoder Representations from Transformers）是一种基于变换器（Transformer）的预训练语言模型，由 Google 于 2018 年发布。它是迄今为止最先进的预训练语言模型之一，具有双向编码器和预训练任务的特点，可以用于各种下游 NLP 任务。  BERT 的训练集来自英文维基百科和书籍语料库，其预训练的目标是通过双向语言模型学习自然语言表示，使其在借助相对较少的标注数据下，能够更好地完成不同的 NLP 任务。BERT虽然计算复杂度较高，但 相比于以前的模型具有以下优点：  (1)双向性：传统的 NLP 模型如 LSTM 和 GRU 通常是单向的，但 BERT 是双向的，它同时考虑前面和后面的词语，因此它可以更好地捕捉上下文信息。  (2)预训练方式：BERT 使用了两个预训练任务，分别是：1)掩码语言建模（Masked Language Modeling，MLM）：将语句中一定比例的词语掩盖，让模型预测出被掩盖的词语。2)下一句预测（Next Sentence Prediction，NSP）：模型需要成功预测两个语句是否在同一语境下。  (3)可微性：BERT 基于变换器结构，这种结构可以很好地解决传统 NLP 中的梯度消失问题，因此 BERT 可以训练更深的神经网络，更好地学习嵌入式表示。BERT 语言模型已经在多种任务上取得了出色的结果，例如语言推断、问答、文本分类、命名实体识别、关系抽取等。此外，BERT 的成功也激发了其他模型的发展，例如升级版的 RoBERTa 和 ALBERT 等。  **3. DeepSeek API集成**  DeepSeek API的调用模式主要基于其开放性和灵活性，支持多种技术架构与应用场景的整合。  API配置流程：通过注册获取API Key并配置至目标平台（如Zotero、Autogen Studio等），实现模型能力的调用。例如，在文献管理工具Zotero中，用户需安装插件并填入API地址、模型名称（如`deepseek-chat`）及密钥，结合向量模型（如Qwen的文本嵌入服务）完成PDF解析、摘要生成等任务。  RAG架构整合：支持检索增强生成（Retrieval-Augmented Generation, RAG），通过将外部知识库与模型推理结合，提升答案的准确性和专业性。例如，用户可预先存储领域文献至数据库，模型在回答时动态检索相关知识并生成结论。  联网搜索功能：部分平台（如腾讯云）支持联网搜索模式，突破模型预训练数据的时间限制，实时获取最新信息，增强问答的时效性。  在中文语境理解（如成语、方言）及数学推理（MATH 500测试得分90.2%）、编程任务（Codeforces测试）中表现优于GPT-4o等闭源模型，更适合本土研究需求。  API调用成本仅为同类闭源模型的1.8%-3.7%，且提供免费额度，降低学术研究的经费门槛。  支持文本生成、代码生成、图像理解等任务，并可整合私域知识库（通过RAG技术），提升领域研究的深度。    **拟解决的关键问题**  **1. 信息处理的准确性与效率平衡**  - 问题：轻量模型精度不足，大模型API响应延迟高。  - 解决方案：  - 混合推理：本地模型处理高频简单任务，复杂任务异步调用大模型API。  - 缓存机制：对常见查询结果缓存，减少重复计算。  **2. 多平台数据同步的实时性**  - 问题：依赖第一接收端更新导致信息延迟（如未启动QQ时无法获取新消息）。  - 解决方案：  - 后台监听服务：在用户设备部署轻量级后台进程，实时监听平台数据更新（需解决移动端权限问题）。  **重点与难点**  模型泛化能力：适应不同平台的消息多样性（如邮件正文、群聊碎片化文本）。  用户体验一致性：实现多终端界面交互逻辑统一，降低用户学习成本。  多模态信息处理：整合文本、图片、附件（如PDF课件）的关联分析，需设计跨模态模型。  资源限制下的性能优化：在移动端有限算力下平衡模型精度与响应速度，可能需量化压缩本地模型。  项目开发中需持续关注开源社区（如GitHub）工具更新，优先选择成熟方案降低实现风险。 | | | | | | | | | | |
| **三、项目特色及创新点**  **1.跨平台聚合** 收集微信、QQ、邮箱等主流平台的数据，通过API混合方案实现多源异构数据整合（如微信  QQNT导出、IMAP/POP3邮箱协议适配），打造跨平台信息聚合工具。  **2.ai辅助个人日常信息管理**  不仅仅是将ai用于企业级任务或个人的规模工作场景，让ai助力个人的日常信息方便接受，高效处理。 | | | | | | | | | | |
| **四、申请理由**（1、团队条件——自身/团队具备的知识、素质、能力、特长、兴趣；2、前期准备基础等）  团队成员均来自计算机科学与技术学院，具备较强的编程能力与项目开发能力，数理 基础扎实，在老师的指导下系统性的学习了深度学习理论并精读前沿文献数篇。  负责人杨胜宇：有较多的项目开发和团队协作经验。具有良好的学习态度和探究精神，对深度学习理论有一定了解，对数据结构有扎实的基础，能较熟练地使用各种语言如Java，Go语言，Python等等，精通算法，对其有较为深入的理解，有较强的解决问题的能力。学习热情高，且有强大的学习新东西的能力。  成员杨启航：具有较强的项目开发与模型建立能力，算法编程能力较强，对计算机数据库如SQLite，MySQL 等相关知识有较多了解，在数理基础方面有扎实基础，计算机相关理论基础 扎实，学习积极度高，具有良好的学习态度和探究精神。  成员郭川嘉：计算机科学与技术专业背景，具备扎实的编程能力与算法基础，深入了解Python等技术栈。具备系统设计、性能优化能力。积极参与开源社区，保持高强度自学，擅长快速掌握新技术并解决复杂问题。良好的团队协作意识，追求以代码驱动价值。  成员宁欣怡：具备扎实的算法设计与优化能力，能够高效解决复杂问题。在数据可视化方面，我熟练使用Matplotlib等工具进行专业级图表绘制。凭借深厚的数理基础，我能够开展大数据处理工作。同时，具备一定的美术基础。此外，我始终保持快速学习新技术的能力。决复杂问题。在数据可视化方面，我熟练使用Matplotlib、Seaborn和LaTeX等工具、 | | | | | | | | | | |
| **五、项目实施方案**（研究思路和方法，实施计划、技术路线、人员分工等） ****一、研究思路与方法****1. ****技术路线设计**** 基于分层架构，将项目拆解为以下核心模块，结合深度学习与大模型技术，实现多平台消息的整合与分类：   * **数据采集层**：通过开源项目,IMAP协议,API调用获取多平台数据（微信、QQ、邮箱，教务网站等），需解决数据加密和获取问题 * **预处理模块**：统一数据格式（如JSON或CSV），清洗冗余信息（如重复消息、无效链接）。 * **数据存储层**：采用SQLite或MySQL存储结构化数据，支持快速查询与更新。 * **ai处理层**：   **a.本地轻量模型**：使用NLP模型（如BERT、LDA等）进行基础分类（如通知、闲聊、广告）。  **b.Deepseek Api调用**：  **接口规范分析**  需明确接口终结点、认证机制（如API Key鉴权）、请求负载格式（JSON/XML）及参数约束（如`max\_tokens`控制输出长度，`temperature`调节生成随机性）。验证输入文本长度限制，确保符合接口承载能力，对于超长文本需预先设计分块处理方案。  **语义模糊性消解**  • 上下文增强技术  在请求负载中注入历史对话记录或领域背景知识，构建多层次语义框架。例如，在会议安排场景中，将前序讨论中的时间提议、地点备选项作为上下文附加，辅助模型识别歧义表述中的潜在意图。  • 指令显性化设计  在用户输入前添加结构化任务指令，明确输出目标与格式要求。例如：“任务：从下列模糊对话中提取实体，按{时间、地点、参与方}字段返回JSON。对话内容：[...]”。  **长程上下文关联建模**  1.对话状态维护机制  采用对话树（Dialogue Tree）或会话记忆池（Session Memory Pool）技术，持久化存储多轮交互状态。每次接口调用时传递完整会话序列，利用模型的注意力机制实现跨轮次信息关联。  2.实体一致性校验  对输出结果实施实体消歧（Entity Disambiguation）与共指消解（Coreference Resolution），通过外部知识库或自定义规则库确保跨上下文的实体指代一致性。  **信息结构化输出规范**  1. 模式驱动生成（Schema-Driven Generation）  定义输出模式（如JSON Schema），通过提示工程约束模型生成结构。例如，强制要求按“摘要”、“关键实体”、“置信度”三级结构组织结果。采用受限解码（Constrained Decoding）技术，在接口层启用输出格式校验，避免非结构化文本导致的解析复杂度。  2. 混合式后处理流程  对模型原生输出实施正则表达式匹配、基于规则的语义解析（Semantic Parsing）或轻量级机器学习模型（如CRF）进行信息标准化。建立异常输出熔断机制，当置信度低于阈值时触发人工审核流程。  **系统鲁棒性保障**  1.容错与重试策略  设计分级错误码映射体系，针对网络抖动、接口限流等临时性故障，实施指数退避（Exponential Backoff）重试算法，最大重试次数需结合业务SLA设定。  2.性能优化维度  实施异步批处理调用（Async Batch Processing），通过请求聚合降低接口调用频次。动态调节生成参数：在信息提取场景降低`temperature`（建议0.1-0.3）以提升确定性，在摘要生成场景适当提高`top\_p`值（建议0.7-0.9）以保留语义多样性。  **效果评估与持续迭代**  1. 量化评估体系  构建领域特定的评估指标：采用ROUGE-L衡量摘要覆盖度，F1值衡量实体抽取准确率，人工标注验证上下文关联合理性。实施A/B测试，对比不同提示模板、参数组合的业务指标差异。  2. 闭环优化机制  建立反馈知识库，收集处理失败的典型案例，用于提示工程迭代与规则库扩充。监控接口响应延迟、token消耗等运维指标，动态调整分块策略与缓存策略。   * **后处理模块**：结合时效性分析（如时间戳排序）与关联整合（如合并相似通知）。 * **用户交互层**：基于React Native开发跨平台GUI，实现消息分类展示、时间轴视图与全局搜索功能。  2. ****关键技术难点与解决方案****  * **数据加密与提取**：   + **微信/QQ**：利用开源工具（如PyWxDump、QQDecrypt）导出本地数据库，提取聊天记录。   + **邮箱/校园OA**：优先通过IMAP协议或官方API获取数据 。 * **模型泛化能力**：   + 采用混合标注策略：本地模型处理高频场景（如课程通知），大模型补充复杂场景（如多语义修正消息）。 * **实时性保障**：   + 通过定时轮询（邮箱）或监听本地文件变化（微信/QQ）更新数据，避免依赖平台推送。 | | | | | | | | | | |
| **六、项目进度安排**（文献查阅、社会调查、方案设计、开题报告、实验研究、数据处理与分析、研制开发、填写结题表、撰写论文和研究报告、结题答辩和成果推广等时间安排）  1.开题报告：2025年4月  2.实验研究：2025年4月-6月  3.研制开发：2025年6月-10月  4.数据处理与分析：2025年10月-11月  5.功能完善与系统测试：2025年11月-2026年1月  6.填写结题表：2026年2月  7.撰写论文和研究报告：2026年1月-3月  8.结题答辩：2026年3月-5月  9.成果推广：2026年3月后 | | | | | | | | | | |
| **七、项目研究所需资源**（实验室、仪器设备、实验材料、资料等）  1.吉林大学图书馆主页的网络数据库（CNKI、SCI、EI等）。  2.机器学习所需高性能GPU，大容量存储设备 | | | | | | | | | | |
| **八、项目经费预算与用途**（购置实验消耗材料、低值品、资料、加工测试、打字复印、调研、市内公交、论文发表、专利申请等经费开支）  软件著作权申请：1000元  GPU算力租用费用 ：3000元  耗材（纸张，鼠标，硬盘等）：1000元 | | | | | | | | | | |
| **九、项目完成预期成果**（成果形式：研究论文、专利、设计、产品、软件、研究或调研报告等）  1. .软件著作权一份  2．论文一篇  3.安卓端app或window上的一个工具软件 | | | | | | | | | | |

