浙江大学

2021大学生科研训练计划申报书

项目编号：

项目名称：　基于自然语言处理的新冠相关推特属性判断

项目负责人：　段皞一　　学　　号：　3190105359

院（系）：　计算机科学与技术学院

联系电话：　19883161889　　电子邮件：　1031004722@qq.com

指导教师：　汤斯亮　　职　　称：　副教授

是否依托学校科教协同基地项目：　否

申报级别：　校级SRTP、　院级SRTP

浙江大学本科生院教务处

填 写 说 明

一、申报书要按照要求，逐项认真填写，填写内容必须实事求是，表达明确、严谨，首页只填负责人，“项目编号”一栏不填。

二、格式要求：申报书中各项内容以Word文档格式填写，表格中的字体为小四号楷体，行距为最小值20磅；表格空间不足的，可以扩展或另附纸张。

一、项目简介

<table>
<tr><td rowspan="6">项目概括</td><td>项目名称</td><td colspan="7">基于自然语言处理的新冠相关推特属性判断</td></tr>
<tr><td>所属一级学科</td><td colspan="7">计算机科学技术</td></tr>
<tr><td>项目性质</td><td colspan="7">（） 基础研究 （√） 应用基础研究</td></tr>
<tr><td>项目来源</td><td colspan="7">（√） 自主立题 （） 教师指导选题 （） 社会企业事业</td></tr>
<tr><td>申请经费</td><td colspan="7">600</td></tr>
<tr><td>起止时间</td><td colspan="7">至</td></tr>
<tr><td colspan="2">项目状况</td><td colspan="7">（√） 研发阶段 （） 中试阶段 （） 批量（规模）生产 （选项打√）</td></tr>
<tr><td rowspan="3">项目申报人</td><td>姓名</td><td>段皞一</td><td>性别</td><td>男</td><td>出生年月</td><td></td><td>入学年月</td><td>2019</td></tr>
<tr><td>学号</td><td colspan="3">3190105359</td><td>联系电话</td><td>19883161889</td><td>电子信箱</td><td>1031004722@qq.com</td></tr>
<tr><td>院系专业</td><td colspan="7">计算机科学与技术学院、混合班</td></tr>
<tr><td rowspan="3">项目组主要成员</td><td>姓名</td><td colspan="2">联系电话</td><td colspan="3">院系专业</td><td>年级</td><td>具体分工</td></tr>
<tr><td>杨浩峰</td><td colspan="2">15779509886</td><td colspan="3">竺可桢学院、混合班</td><td>2019</td><td></td></tr>
<tr><td>翟智超</td><td colspan="2">19883145932</td><td colspan="3">计算机科学与技术学院、计算机科学与技术</td><td>2019</td><td></td></tr>
<tr><td rowspan="4">项目指导老师</td><td>姓名</td><td colspan="2">联系电话</td><td colspan="3">所在单位</td><td>职务/职称</td><td>主要研究方向</td></tr>
<tr><td>汤斯亮</td><td colspan="2">13588196277</td><td colspan="3"></td><td>副教授</td><td>信息抽取、自然语言处理、跨媒体计算</td></tr>
<tr><td colspan="8">近三年成果：国家级_1_等奖 _0_项，省部级_1_等奖_0_项</td></tr>
<tr><td colspan="8">近三年科研经费_300_万元，年均 _100_万元</td></tr>
</table>

项目主要内容简介	本项目基于自然语言处理，任务是预测关于COVID-19的tweet的二进制属性：它是否有害，是否包含可验证的声明，是否可能引起公众的兴趣，是否似乎包含虚假的信息，等等。
项目负责人参与科研情况	1. 大一上学期，在心理学与行为科学系副教授张萌老师的指导下，参与研究课题“大脑语言偏侧化”的延伸课题。 2. 大一下学期，短学期综合实践中，参加了陈建海老师的超算课程，参加了第八届“英特尔”杯全国并行应用挑战赛，课程小组作为优化组参加初赛，入围150多支参赛队伍的16强并进入决赛，同时也是浙江大学进入决赛的两只队伍之一。最终决赛中取得了三等奖，奖金5000元。 3. 大二上学期，物理实验参与课题“锁相放大器的设计与弱声压信号的测定”。
项目组成员参与科研情况	暂无

二、项目背景、目的及意义

（简要说明项目背景、意义和实施必要性，研究现状和发展动态，不超过1100字）

项目研究背景目的

首先，从大数据本身来说明。目前以及未来很长一段时间我们都处于大数据阶段，而大数据想要提现出数据的价值，就离不开机器学习、人工智能技术，同样人工智能想要体现出优势必须基于大数据。目前各大公司都有自己海量的数据，并且运用人工智能技术展现出价值，但从数据量本身来说，或许目前最大的数据量应该是互联网上茫茫多的网页，现阶段对这些茫茫多的网页的利用度还是比较浅的。这些网页对包括google、百度等互联网巨头来说还有很大的价值需要去挖掘，而几乎每个网页都有一定量的文本内容，绝大部分网页完全靠文本来展现其核心内容，而这些文本内容都是自然语言。那么自然语言处理的研究价值就相当明显了，想要深度去挖掘网页的价值，就必须有好的自然语言处理手段。

第二，从人工智能技术本身来说明。人工智能技术目前来说发展较好，但是又在文本处理方向相对发现欠缺，对文本的处理将来预计要登上舞台中心。从人工智能这个词本身来看，机器要想实现智能，如果连人类的语言都不能理解，怎么和人类好好的交流，怎么体现智能之处。

项目的意义

利用现代化的工具（计算机）高效地进行数据处理。

对传统语言学进行正反馈。

第一方面的适用场景非常广泛，如电商平台的推荐系统，搜索引擎，智能对话，都需要相应的NLP技术进行数据处理与分析。尤其是在电商、金融等领域，每天都会产生大量的信息，我们需要对其进行分类与提取，实现产品在用户身上的精准打击。

当然了，除此之外还有机器翻译，输入法等等应用。

第二方面，不是NLP的主要目的。注意这里说的是“正反馈”，也就是说，传统语言学是为了弄清人类语言语法、语义的形式化与非形式化表示与内涵，从语素，词，词组，小句，句组到段落、篇章，语言学全方位无死角覆盖。然而，处理自然语言的背景下，如果不的传统语言系统的内在理路进行深耕和远拓的话，很难在自然语言处理上走得远

综上所述，现在的NLP关注核心是在它的应用上，即如何用更高效的算法指导计算机处理大规模数据，这要求我们设计更高效、更通用的模型，所以我们才会有Transformer，才会有BERT。

三、项目研究方案

（包括项目的主要内容、计划目标和拟解决的问题，思路方法、组织实施及进度安排，不超过1200字）

本项目主要时进行一种二层的词义分类，所以研究步骤大致如下：

1. 先进行调研，观看基于深度学习的自然语言处理基础课程，为后期项目的有效推进作必要的知识储备。

2. 通过调研和学习，了解目前词义分类的方法和模型具体有哪些，它们的优点和缺点分别是什么。

3. 在之前的准备基础之上，再去看看可以做两层的词义分类的模型有哪些，特点是什么，如何使用。

4. 之后搭建自己的测试平台，在平台上进行测试样例的测试。具体方法是，目前打算分割出比赛测试数据集的一部分用来做测试。

5. 测试完成后，分析不足，进行代码上的优化。

6. 优化完成，进行比赛测试集的最终测试。

7. 总结上述的实验过程，撰写相关报告和论文，完成项目的结辩。

时间进度的安排：

2021年4月-6月 相关领域入门，学习基于深度学习的自然语言处理基础课程，了解机器学习基础，深度学习基础，词嵌入，卷积神经网络，词向量，生成模型与语言生成等知识，进一步深化巩固C++、python的语言知识，位置之后编写项目相关代码提供支持。

2021年6月-7月 领域文献综述 读懂相关模型算法

2021年7月-10月 搭建具有两层词义分类功能的平台

2021年10月-11月 通过分割的数据集对搭建的平台进行调试，分析其可靠性

2021年11月-2022年1月 在之前的测试基础上，对平台进行算法上的优化

2022年1月-4月 整理项目成果，进行实验报告、论文的撰写

2022年4月-5月 准备答辩

四、项目研究条件及创新之处

（已有研究基础，包含与项目有关的研究积累、已取得的成绩和已具备的条件，尚缺少的条件及解决办法，项目优势和风险，以及项目创新点等，不超500字）

项目研究条件

汤斯亮老师课题组研究内容涵盖跨媒体内容理解、信息抽取、自然语言处理等人工智能方向，为SRTP项目提供了很好的指导和支持。

创新之处

我们小组作为初来乍到，刚刚接触自然语言处理的新人，我们在自然语言处理的项目上有很多需要准备的，有很多路要走，有很多“硬骨头”要啃，但是，我认为，初来乍到也不完全是一个劣势，在看待自然语言处理的问题上，我们有着自己的一些独特的认识，不受传统模型的禁锢，可以跳脱出一些已有的算法，另辟蹊径，这在项目中算法优化的尝试道路上是有一定帮助的。

此外，我们作为大二学生，在语言这一块是相对较熟悉的，之前也接触过并行计算，想进行多种语言的混合开发，这样能在不同环节利用不同编程语言的优势。

如果时间语允许的情况下，后期的优化过程中，想进行OpenMP\MPI优化，提高计算的速度。

五、项目预期成果

（包括知识产权成果，如论文成果、获奖成果、评议鉴定成果、推广成果、论著成果、专利成果、、研制产品、开发软件，与毕设、学科竞赛等其他学习环节结合情况，或其他成果等，以及经济效益、社会效益等，不超130字）

1. 基于自然语言处理，搭建平台预测关于COVID-19的tweet的二进制属性：它是否有害，是否包含可验证的声明，是否可能引起公众的兴趣，是否似乎包含虚假的信息，等等。

2. 根据实际问题，提出改进算法。

3. 实现平台，语言处理效果的演示。

4. 整理相关结果，撰写论文。

六、项目财务预算

（包括经费预算及经费支出明细等）
专用材料费 400.0 元；
印刷费与资料费200.0 元；
交通与差旅费 0.0 元；
出版费 0.0 元；
邮寄费 0.0 元；
评审费待确定。国创为300，省创 200 元。校级200元，院级自筹

七、项目组承诺

承 诺 书

以上所填内容真实可靠，本项目组承诺：该项目立项后，将严格遵守有关规定、遵守本申报书和预算表中规定的条款和内容，保证按计划进度完成项目任务。

项目组全体成员（签字）： 段皞一　杨浩峰　翟智超

年 月 日

八、指导老师意见

同意参与

指导老师（签字）：汤斯亮

年 月 日

九、院（系）专家组意见

专家组组长（签字）：

年 月 日

十、学校审核意见

（盖章）：

年 月 日