

上海交通大学

SHANGHAI JIAO TONG UNIVERSITY

本科生毕业设计（论文）中期检查报告



论文题目: 舆情文本多分类与敏感性检测技术研究

学生姓名: 周笑

学生学号: 516030910370

专 业: 软件工程

指导教师: 唐新怀

学院 (系): 电子信息与电气工程学院

教务处制表

填表说明

1. 请每位学生根据学校及院（系）检查的要求认真进行自查，及时发现课题研究过程中存在的问题，分析原因，并提出解决思路 and 措施，明确下一阶段任务。
2. 每位学生应根据项目实施情况认真、实事求是填写。填写字体请用宋体小四号，并用 A4 纸打印，于左侧装订成册。
3. 毕业设计（论文）中期检查报告总字数应满足本院（系）要求。
4. 该表填写完毕后，须请指导教师审核，并签署意见。
5. 《上海交通大学本科生毕业设计（论文）中期检查报告》将作为答辩资格审查的主要材料之一。
6. 本表格不够可自行扩页。

课题进展情况：

此课题自开题以来，已经经过了一个多月的时间。在这一个多月里，受限于是硬件条件以及沟通成本，此课题并没有进行地十分顺利。但是这一个月里我也完成了诸多工作。

首先，我按时完成了前期地准备工作。包括理论学习，论文的研读以及对于框架的初步认识。我学习了解了 BERT 框架的工作原理，认识到了 BERT 框架的优势和不足，也认识了 BERT 框架在实际应用中的难点和至今没能解决的一些问题。这让我对于整个项目的难度和工作量有了初步的认识。

接下来，我初步上手了 BERT 框架并依照网络上和 Google 官方的指导进行了初步的尝试。这个过程并不容易。从环境的配置，到各个环节的代码，再到一些具体参数的调整，我都经历了不断地尝试，最终在第一阶段检测的时间点成功运行了测试代码。虽然训练结果没有达到课题的要求，但是是一个好的开始。

自第一阶段检查以来，课题遇到了瓶颈，受限于是缺少服务器的硬件支持，在自己的电脑上运行太慢，而且结果不佳，导致实验的时间成本太高。在学长的介绍下，我找到 Google 提供的 Colab，但是又受限于是网络，经常出现网络连接断开的问题。这些问题导致课题研究的进展并没有预想中的顺利。

课题研究已取得的阶段性成果：

课题进展到现在，我已经取得了以下成果：

第一，拥有了一份可运行的代码，并了解了代码的工作原理和具体的参数调试方法。

第二，得到了一份初步测试的结果。

第三，已经得到了处理好的数据集，可以用于接下来的调优。

存在的问题及解决思路：

课题进行到现在，依然存在以下问题：

第一， 缺少硬件支持，导致调试成本太高。

第二， 在现有条件下，训练模型的准确率不高，而且由于课题要求中有明确的准确率要求，能否达到要求暂时还没有保证。

第三， 受限于 BERT 模型自身的设计，长度超过 256 个字的长文本需要进行第二次数据处理，可能会影响长文本的训练结果。

第四， 由于没有开学，沟通不是很顺畅。课题要求中的“集成到一个平台”的环节还没有开始。

根据以上问题，我提出如下的解决思路：

第一， 希望可以早点开学，利用学校的网络和服务器资源，改善现在的硬件条件。

第二， 积极与学长和导师沟通，向他们吸取经验，交流心得，不断尝试，看能否取得突破。

第三， 我决定采用多个不同的数据处理办法，分别实验，比较测试数据。最后采纳准确率最高的方法。

第四， 与同导师的同学们积极沟通，互相交流。



下一阶段的工作计划和研究内容：

下一阶段我将主要着力于以下任务：

第一， 进行更多的实验和测试，不断地进行模型调优，尽可能地达到课题的要求。

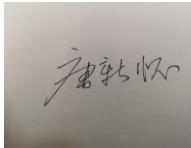
第二， 向导师，学长，同学和网络上的资源学习，寻找解决长文本问题的新的思路并进行实践。

第三， 利用开学的时间，抓紧机会提升自己的代码质量，提高测试效率，降低训练成本。

指导教师意见：

进度符合开题要求，关键算法得到实现，望进一步加强算法优化，抓紧论文写作，同意通过中期考核

指导教师签名：_____



2020 年 5 月 8 日

学院（系）意见：

审核通过

审 查 结 果： ☒ 同 意 ☐ 不 同 意

学院（系）负责人签名：_____

2020 年 5 月 8 日