NLP小白的成长之路

我是朱政烨，目前是一名大三学生，就读于北京大学元培学院数据科学专业。

上个学期，我开始接触到自然语言处理，并且学习了一些比较入门的理论知识，但是苦于没有实践的机会。今年，由于疫情的形势我只能待在家里，于是开始接触一些线上的比赛，而天池就是这样的一个著名的平台。 “公益AI之星”挑战赛第一时间就吸引了我，这是一个比较经典的NLP分类任务（相似句对判定），而且也与目前的疫情、医疗息息相关，具有一定的现实意义。虽然没有丰厚的奖金，但是有助于我入门NLP，也许还能为医疗问答方面的公益事业做一点贡献，何乐而不为呢？

初生牛犊不怕虎，我立即报名了比赛。第一道关卡就难到了我——docker镜像的提交方式。我反复查询资料，通过其他比赛尝试提交，总算在比赛前天弄清了提交方式。接着，我首先尝试了baseline，它调用了苏剑林大神的bert4keras库，封装得很好，对于简单的任务很方便。我使用了bert模型，以及数据竞赛常用的五折交叉验证，就有了95.5左右的成绩，前期也在榜单上名列前茅。初尝甜头的我还不满足，浏览了一些文章以及类似比赛的攻略，尝试了一些技巧，其中模型融合最为有效，三个模型融合后，直接在第六天提到了96.30的高分。在接下来的一两周内，我暂时放缓了节奏，没有过多地尝试，直到96.42大佬的出现让我又有了动力。我在最后一周努力尝试，在融合方式、后处理等上下功夫，总算有所提升，最高达到了96.44，圆满地完成了比赛。

通过这次比赛，我进一步了解了一些NLP知识，提高了实践能力，可以说是受益匪浅。同时，我也颇有感慨，现今是互联网和大数据时代，寻医问诊也有了线上的形式，如果相关的技术能应用到实际中，对于大众可以说是再好不过了。最后，感谢天池官方的组织，希望以后能多多参加这样有意义的比赛。