为期6周的项目结束了，从8 月 22 日晚上连夜训练跑代码到8 月 23 日上午的 Presentation，虽然还有点紧张，但我觉得我们组的表现还是挺好的，毕竟在这之前彩排了差不多 3 次呢，Pre 期间教授一直对我们说” I appreciate your speed” ,总感觉我们说的太快了，很多教授没有太在意。毕竟在 10 分钟之内把那么多东西装进去实在不容易，不管怎么说，我在这次的远程项目中还是收获蛮多东西的。

我们组最开始的想法是通过Bert提取出向量，之后通过提取的向量送入其他分类器进行处理，但在实际操作工程中我发现得到向量的时间非常长，在有限的时间内可能无法得到很好的结果；其次，即使得到了最终的结果，向量值也包括了词向量和句向量，处理起来相对来说比较困难，所以最终我们选择了 Bert 单独一个 Pipeline。我通过一些文章了解到 Glove比 Word2Vec有更高的f1 score，所以利用 Glove 取代了 Word2Vec，并加入相关分类器进行单独的 Pipeline，这项工作由亦辰和赵续负责。之后我和谭滨凯进行沟通，了解到 Elmo 其实和 Bert 的工作原理类似，都是进行动态矩阵的更新，所以滨凯利用了 Elmo 进行了最终的 Pipeline，这就是我们组的 3 个完整的 Pipeline，还是有点复杂的。

为了学习 Bert，我在网上找了很多的资料，因为 Bert 实现的原理是基于前人的经验，所以我从最开始的 Word2Vec 开始，到中间的Elmo，GPT到最终的 Bert，了解了 NLP 大概的历史以及实现原理思想，最终才大概的搞清楚了 Bert 该如何运用到该项目中。在代码方面，我借鉴了别人完成的 Bert 文本二分类的代码，在其基础上进行了修改，因为最开始将文本送入模型就跑出了很不错的成绩(0.980)，所以在如何进行再次提高分数的方面我思考了很多，也看了很多的论文，最终采用了EDA 和删除一些特殊标记的方法以提高分数，最终得到了 0.986 的成绩。

回顾完成的所有的工作，的确成就感是蛮强的，因为最开始我对于 NLP 基础基本为零，到最后我了解了 Bert并用 Python 代码实现了文本分类的功能。在最终的测试过程中，由于时间紧任务重，我在淘宝租用了 1080ti 去完成所有的工作，并细心准备了PPT 供Pre使用，希望能给教授留下不错的印象吧，也希望大家都好好努力，一起加油！

最后附上我的贡献图:

A screenshot of a cell phone

Description automatically generated

GitHub 的链接: <https://github.com/NavePnow/Google-BERT-on-fake_or_real-news-dataset#5-part4-reference>

以及我的博客: <https://nave.work/categories/FakeNews/>

谢谢。

王依凡 2019.08.30