自然语言处理

Project 1 Task 3

院系：人工智能学院
姓名：王蔚昕
学号：211300042

2024 年 4 月 28 日

目录

1 问题一

1.1 实验方法

在这一环节，我使用了 Kmeans 进行新的划分，同时为了能与 MBTI 进行的性格分类进行划分，所以我依然把它分成了 16 种性格，以此来检验不同分类方法对于这一分类任务的影响。

具体的来讲，我先用 Bert-tiny 得到原始数据的特征，随后使用输出的 pooler_output 作为 Kmeans 的分类标准，将样本进行十六分类，得到新的标签。

1.2 实验结果

使用 Kmeans 成功完成了一个十六分类，同时根据这个十六分类的 Kmeans 对验证集也进行标签的更新。

2 问题二

2.1 实验方法

结合上一问题中得到的新标签，和提取后的特征，我设计了一个线性分类网络，由四个线性层，一个 Softmax 层组成，用作分类器。

2.2 实验结果

使用模型	使用分类方法	准确率
BertBase	MBTI	40%
BertTiny	Kmeans	70%

2.3 结果分析

显而易见的是，更复杂，能力更强的模型在 MBTI 上取得了更低的准确率，这显然是因为使用的分类方法的区别。

2.4 一些思考

2.4.1 Kmeans 得到的分类是否有意义

Kmeans 一定可以对任何样本集完成一种分类，但是这种分类是否具有现实意义呢？我们知道 Kmeans 仅仅是对样本特征进行划分，找到一些“深层”联系，但是这些联系可能在实际生活中没有体现到，或者说，不能体现我们想要的信息。就本问题而言，被 Kmeans 分为同一类的很可能不是同一种性格，而是在语义上大意相同的语句，在我进行一些语句展示时证明了这一点，亦或者我本人不具备区分性格的能力。所以 Kmeans 得到的只是一个统计学上的分类，而并非是我们想要的性格分类。

2.4.2 为什么 Kmeans 得到的分类效果会好很多

Kmeans 的分类标准是按照样本特征的距离来分类的，在这种情况下，同一类的样本的特征往往很相似，所以对于分类器而言，他只需要找到这些特征的分界线，也就是说 Kmeans 分类得到的结果几乎都是可分的。然而根据 task-1 中 KNN 的糟糕效果不难看出，对于大部分样本来讲，他在 Bert 编码后几乎是杂糅在一起的，只有只针对某一些类而言可能更容易完成分割。所以这可能降低了线性分类器分类的难度，导致 Kmeans 得到的分类效果更好，但是还是回到第一个问题，Kmeans 的分类或许对应的是语义而非性格。

2.4.3 文本提取器可能的改进

有研究表明，LLM 输出的特征信息越深层可能越接近语言的深层含义，表层特征更容易做错别字检测与语法检测，但是深层特征做这些效果不好，结合此前提出的猜测，语义往往和话题有关，而非性格，所以用来编码的特征应该更浅层，或者浅层和深层的结合，但是绝对不应该是最深层。为此，可能要用一些能够综合考量句子特点的模型来进行特征提取，而不是能够深入语义的模型。

3 对整个实验的回顾

3.1 不足

在早期实验中，因为某些代码的错误，导致实验结果存在偏差，包括小数点位数的错误。

同时，早期选择了错误的模型，导致训练难以在较大数据集上展开。

由于一些技术设备问题，前期实验进行并不充分，数据及利用有限，导致某些结论来源于推测而非直接的实验证明。

以上问题都会在我的 github 账户中进行更新完善。

3.2 收获

体验学习了面对 nlp 问题的处理方法，包括数据处理，模型训练等过程，对开源库 transforms 有了更多的了解和认识。

通过调查文献与实验，对大语言模型的训练过程，推理过程有了更深层次的认识，同时借由此次文本分类任务，对分类器这一日常使用的简单部件有了深层的理解。

这次项目锻炼了我进行问题探究的能力，受益匪浅。