基于 Transformer 的双向编码器表征预训练模型的文本分析

摘要：近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。本文将基于自然语言处理和数据挖掘技术对文本中的政务信息数据进行内在信息的挖掘与分析。在本次挖掘中针对问题 1 即留言分类问题，我们使用了 pytorch_pretrained_bert 库方法，来处理数据，构建模型，并通过使用附件 2 中的数据来进行训练，最终利用 F-Score 方法进行检测，最终成绩约为 0.92。

关键词：文本分类、热点挖掘，bert, pytorch

目录

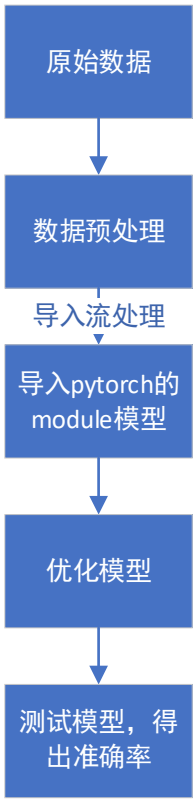
一、 引言

本次建模，利用互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的 答复意见的数据。我们利用 bert 方法我们构建了一个分类模型，并利用以上数据进行了训练，从而达到了对网络问政平台的群众留言数据的一级分类。

二、 分析方法与过程

2.1 群众留言分类

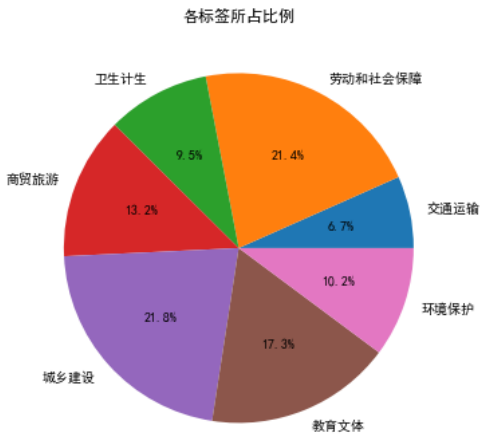
2.1.1 总体流程



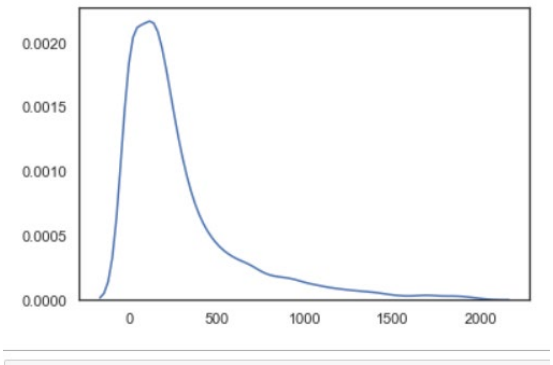
2.1.2 具体步骤

1. 数据分析与处理

（1）数据分析



我们对附件二中各类标签所占百分比进行了分析，如上图，可以看到各类标签的分布并不均匀，城乡建设、劳动社会保障占比较多，而交通运输占比较少，这也符合生活实际的情况。另外数据中只有附件一中出现的部分一级标签，而我们只能基于附件二的数据拟合模型，故最后得到得模型只能对城乡建设、劳动和社会保障、教育文体、商贸旅游、环境保护、卫生计生、交通运输这 7 类进行分类。



我们再对附件二中留言主题与留言内容的文本总长度进行了统计，其中大部分数据都都没有超过 500 的长度。200 左右的文本是最多的。

（2）数据处理

首先我们用 StratifiedKFold 方法将附件二按照标签的的数量成比例的将数据分成 5 份训练集与预测集。

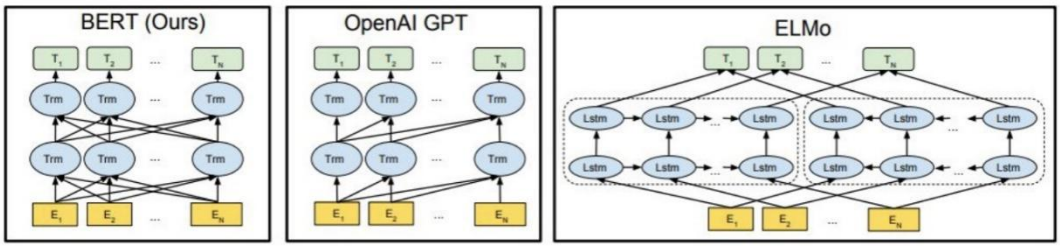
再将统计文本中所有的字符，在排除常规的符号后做成停用词表，并用正则表达式将文本中的停用词进行清洗掉。

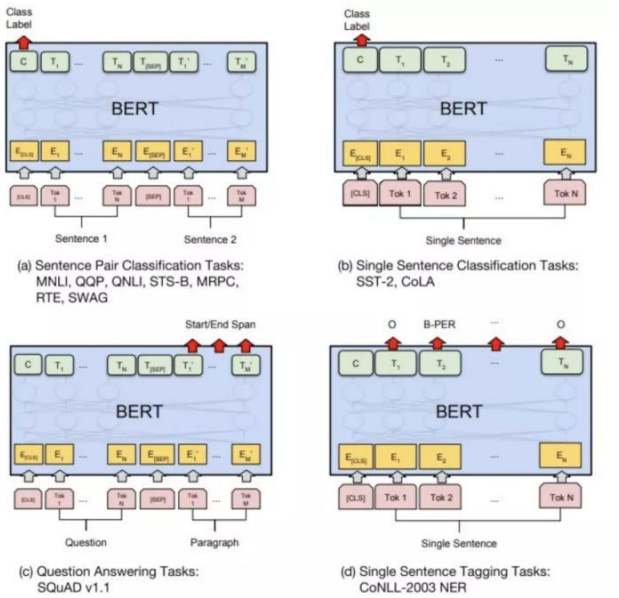
因为模型输入的需要，将数据的留言详情的开头融入留言主题，作为后续模型输入的文本。然后将一级标签分别用 0-6 的数值代替。

2.构建模型

在模型方面，我们采用了基于 BERT 的改进版本的 RoBERTa 模型

BERT 是一个语言表征模型（language representation model ），通过超大数据、巨大模型、和极大的计算开销训练而成，在 11 个自然语言处理的任务中取得了最优（state-of-the-art, SOTA）结果。





BERT 提出一种新的预训练目标：遮蔽语言模型（masked language model，MLM），MLM 随机遮蔽模型输入中的一些 token，目标在于仅基于遮蔽词的语境来预测其原始词汇 id。与从左到右的语言模型预训练不同，MLM 目标允许表征融合左右两侧的语境，从而预训练一个深度双向 Transformer。

因为 BERT 所能接受的最大长度是 512 个字符，所以我们要对数据的长度进行截断，同时由与电脑配置的限制，我们最终将字符的长度限制到了 64 个字符，但是由于文本的主要信息都可以由文本主题提取，因此尚能提取出一定的文本特征。如果提高机器性能，设

置更长的最大文本长度，效果应该会更好。

同时还在文本的两端分别加上"[CLS]" ,"[SEP]"标志句子的开始位置与结束位置。

并用 tokenizer 转化为字符向量。

接着将数据导入 DataLoder 中以 BATCH_SIZE = 16 的大小导入模型。

模型是 chinese_roberta_wwm_ext_pytorch，是基于 pytorch 的 bert 预训练模型。

将处理好的数据对预训练模型进行 fine-tune,融合成为需要的模型。

3.模型测试

将训练好的模型对测试数据集进行测试，计算 acc(正确率)与 F1-score，对模型的训练重复了 3 次，如果 acc 更高则覆盖前一个模型，最后的到模型就是最好的模型。

三、 结论

本次学习我们尝试了使用了基于 BERT 的预训练模型，由于之前并没有怎么接触过深度学习的内容，导致最后的代码完成的十分的简单。都是通过本次学习，我们认识到了深度学习构建模型的大致流程，同时加深了对 BERT 这个先进模型的理解，在后续的比赛中一定可以更好的构建模型。

四、 参考文献

hccccccc：https://zhuanlan.zhihu.com/p/112655246

Naturali 奇点机智：https://zhuanlan.zhihu.com/p/51413773