基于Fine-tuning BERT的情绪诊断辅助系统
结合LIME的解释性模型设计

刘沛卓
西北工业大学附属中学
pzliu@berkeley.edu 181-8920-2653

1. 数据集介绍

本研究使用了Sentiment140数据集，该数据集由Twitter上的用户推文组成，包含了大约160万条数据。

数据包括目标(target)、索引(ids)、日期(date)、标记(flag)、用户(user)、文本(text)共六方面。其中，目标(target)部分表示该推文被标注为积极或消极的情绪，该数据集庞大的数据量和准确的标记使得其成为一个较为理想的训练数据集。

图1:Sentiment140数据集概览

2. 模型选择

我们调取了Hugging Face的Transformers库，对预训练好的BERT模型进行Fine-tuning，并根据二元分类任务设置num_labels=2，并使用稀疏分类交叉熵*(Sparse Categorical Cross Entropy)*损失函数进行模型编译，使用准确率(Precision)作为模型的评估标准(Metrics)。

训练过程中，我们使用*Adam(Adaptive Moment Estimation)*优化器并采用自适应学习率，将批次大小设置为16，并设定4轮训练，在训练过程中通过验证集评估模型表现，防止过拟合。

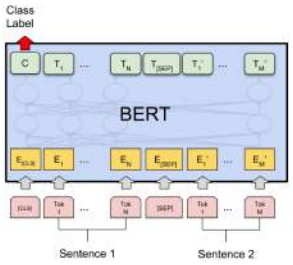


图3（上）：使用BERT进行分类任务的流程图

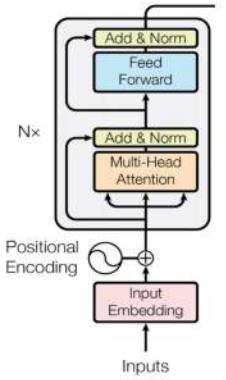


图4（右）：BERT的实现原理，即Transformers编码器

3. 数据预处理

我们将数据集分成了互相不重叠的训练集和测试集两部分。此外，我们还对数据进行了预处理，去除了特殊字符，例如标点符号和网址链接，将所有字母均转换为小写，并去除了停用词，以尽量减小以上因素对模型的干扰， 并且通过tokenization和添加特殊标记（如[CLS]和[SEP]）将推文文本转换为了模型可接受的输入格式。

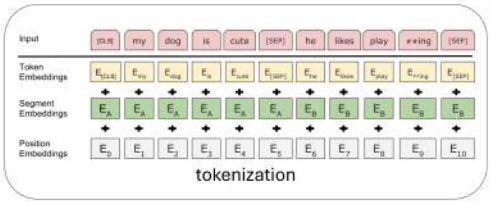


图2：数据预处理流程图，其中：[CLS]表示句子的开始，[SEP]用于分隔句子（多句输入）或表示句子的结尾，标记添加完成后，*BERT(Bidirectional Encoder Representations from Transformers)*将会使用其内置的词汇表生成一一对应的索引，这些索引就是最终输入模型的ID。

4. 模型训练

训练过程中，我们以batch为单位，绘制Loss & Accuracy 曲线，以及学习率的变化情况如下：

Loss: Accuracy: 学习率:

我们可以发现：

1. 损失函数整体呈现下降趋势，在某些学习率调整后，下降更明显，并且最终趋于稳定。
2. 准确率逐渐提高，并在后期趋于平稳，准确率上升代表模型在逐渐学习训练数据中的模式，后期趋于平稳表明模型已经接近训练集的最优性能。
3. 由于我们采用自适应学习率， 学习率呈现分段式下降，初始较高的学习率有助于快速收敛，后续逐步降低学习率后，模型可以更精确地调整权重。

5. LIME 解释

为了解决深度学习模型的“黑箱”问题[2]，本研究引入了*LIME(Local Interpretable Model-agnostic Explanations)*技术，从而帮助用户理解模型的决策过程。LIME通过提供特征的重要性分析，提升了模型的可解释性，帮助用户进一步相信模型并合理判断是否使用模型的决策。

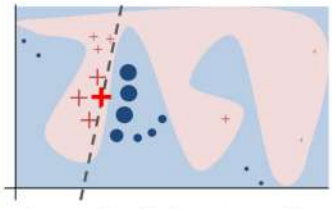


图5：LIME的工作原理，以二元分类问题为例。

$$\varepsilon(x) = \underset{g \in G}{\operatorname{argmin}}(L(g, f, \pi_x(z')) + \Omega(g))$$

$\xi(x)$
局部解释模型，用于解释f在x附近的行为

$\underset{g \in G}{\operatorname{argmin}}$
从候选解释模型集合G中选择最优模型g

$L(g, f, \pi_x(z'))$
损失函数，衡量简单模型g对复杂模型f在点x邻域内的拟合效果

$\Omega(g)$
正则化项，限制模型g的复杂性

6. 模型预测结果分析

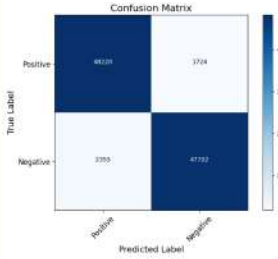
通过K-fold交叉验证的方法，本研究对模型在Sentiment140数据集上的性能进行了评估。具体结果如下：

准确率（Precision）： 对于积极和消极推文的平均准确率分别约为**96.51%**和**95.34%**。

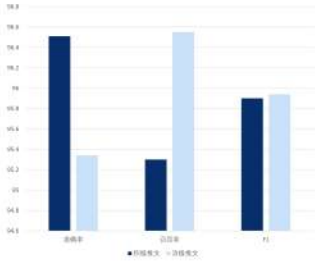
召回率（Recall）：对于积极和消极推文的平均召回率分别约为**95.30%**和**96.55%**。

F1-score：对于积极和消极推文的F1-score约为**95.90%**和**95.94%**。

混淆矩阵表示如下：



结果热力图如下：



7. 用户交互界面

我们采用最简单的控制台交互界面，用户输入想要分析的语句后，模型会输出情感预测，置信程度和LIME解释：

```
请输入文本进行情感分析: what a lovely afternoon!
情感预测: Positive
置信程度: 1.00
LIME解释:
Feature: afternoon, Weight: 0.35192266534772875
Feature: lovely, Weight: 0.3449941127841383
Feature: a, Weight: 0.09364772062183808
Feature: what, Weight: 0.06168221533507195
```

在LIME解释部分，Feature作为输入特征，代表原始模型做出这个决策的判断依据，而Weight则代表该特征对于决策的贡献程度。

图6：用户交互界面