# 基于多BERT集成模型的SMP2020-EWECT分类方法

## 摘要：在SMP2020微博情绪分类技术评测中，本文基于哈工大发布的roberta\_wwm\_ext\_large[4][5]模型，通过迁移学习、k-fold、投票等技术的融合，在通用和疫情测试集上的宏平均F1的平均值为0.7314，在最终提交的42个模型中排名第6。

关键词：情绪分类，BERT模型，迁移学习，集成学习

## 1 引言

SMP2020-EWECT是第九届全国社会媒体处理处理大会的微博情绪分类技术评测，和上一届的中文隐式情感分析评测不同，本次评测，数据情绪由三类（褒义、贬义和无情感）变为六类（积极、愤怒、悲伤、恐惧、惊奇和无情绪），此外还新增了微博疫情数据，相比以往更具挑战性。

微博情绪分类任务旨在识别微博中蕴含的情绪，输入是一条微博，输出该微博所蕴含的情绪类别，是一个多分类任务。在本次评测中，微博按照其蕴含的情绪分为以下六个类别之一：积极、愤怒、悲伤、恐惧、惊奇和无情绪；按照数据所在领域可以分为两类：通用与疫情。通用数据集是随机获取的微博内容，不针对特定的话题，覆盖的范围较广。疫情数据集是在疫情期间使用相关关键字筛选获得的疫情微博，其内容与新冠疫情相关。

表1列出了我们所使用的数据集大小，不难看到virus的规模只有usual的1/3，我们尝试使用伪标签、迁移学习等技术来缓解这一问题；表2展示了每个类别的数据样例，同种情绪下，虽然通用数据和疫情数据内容有差异，但依旧有部分相似性，这是我们使用迁移学习的基础。

**表1 数据集样例数目**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train | Dev | Test |
| 通用数据集 | 27768 | 2000 | 5000 |
| 疫情数据集 | 8606 | 2000 | 5000 |

**表2** **数据样例**

|  |  |  |
| --- | --- | --- |
| 情绪 | 通用数据集 | 疫情数据集 |
| 积极 | 哥，你猜猜看和喜欢的人一起做公益是什么感觉呢。我们的项目已经进入一个新阶段了，现在特别有成就感。加油加油。 | 愿大家平安、健康[心]#致敬疫情前线医护人员# 愿大家都健康平安 |
| 愤怒 | 每个月都有特别气愤的时候。，多少个瞬间想甩手不干了，杂七杂八，当我是什么。 | 整天歌颂医护人员伟大的自我牺牲精神，人家原本不用牺牲好吧！吃野味和隐瞒疫情的估计是同一波人，真的要死自己去死，别拉上无辜的人。 |
| 悲伤 | 回忆起老爸的点点滴滴，心痛...为什么.接受不了 | 救救武汉吧，受不了了泪奔，一群孩子穿上大人衣服学着救人 请官方不要瞒报谎报耽误病情，求求武汉zf了[泪][泪][泪][泪] |
| 恐惧 | 明明是一篇言情小说，看完之后为什么会恐怖的睡不着呢，越想越害怕[吃驚] | 对着这个症状，没病的都害怕[允悲][允悲] |
| 惊奇 | 我竟然不知道kkw是丑女无敌里的那个 | 我特别震惊就是真的很多人上了厕所是不会洗手的。。。。 |
| 无情绪 | 我们做不到选择缘分，却可以珍惜缘分。 | 辟谣，盐水漱口没用。 |

本文接下来将从模型的整体结构、实验结果与分析以及总结完整地介绍我们所使用的方法。

**2. 数据分析**

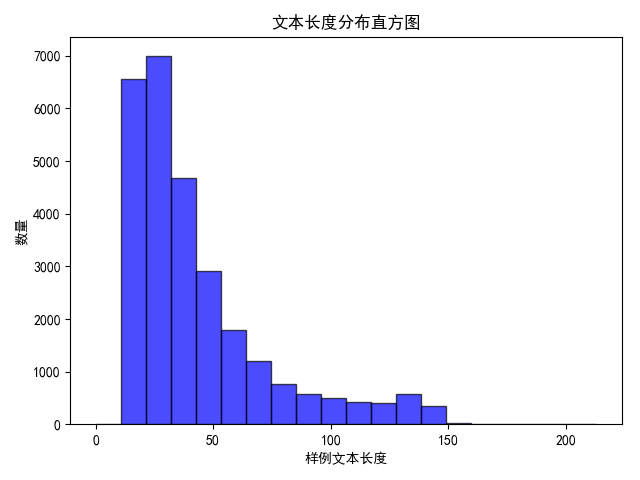
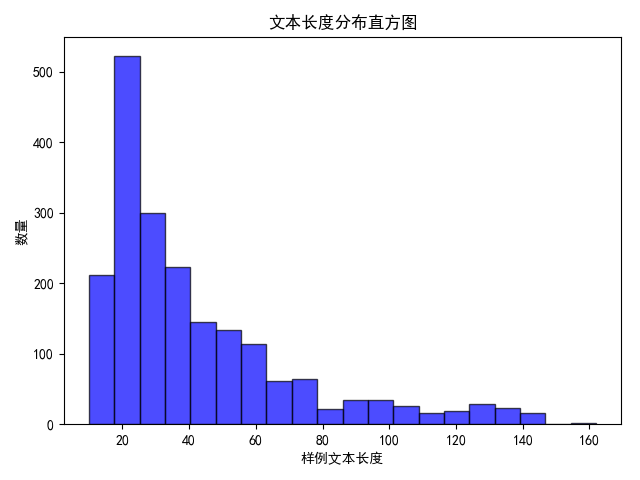
** **

图1 usual train 文本长度分布 图2 usual eval 文本长度分布

从图1和图2可以看出，usual数据集的文本长度基本都在150以下，所以在模型在处理数据时，可以将最大长度设置为150。

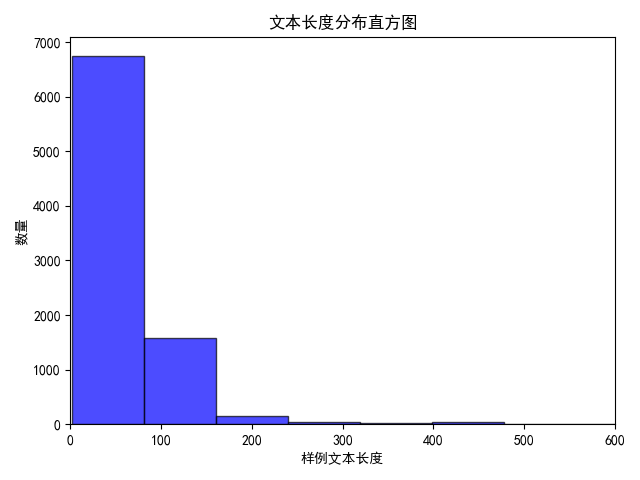
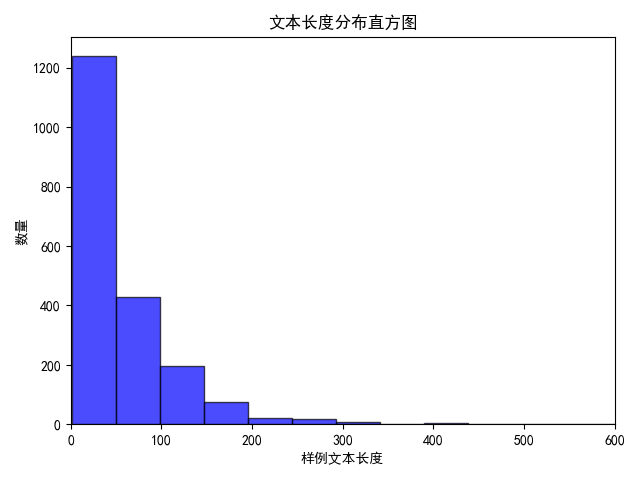
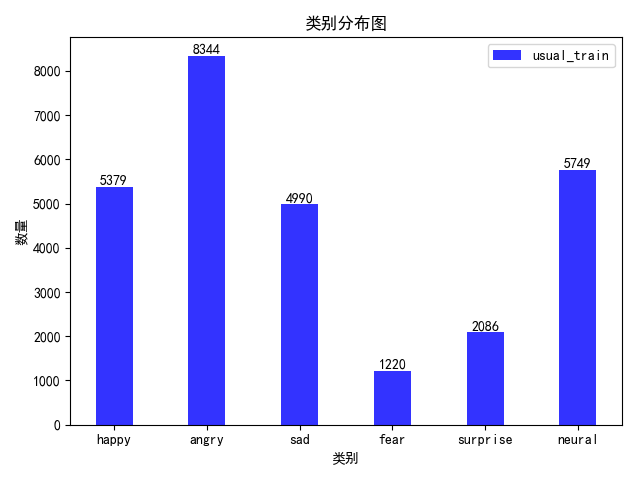
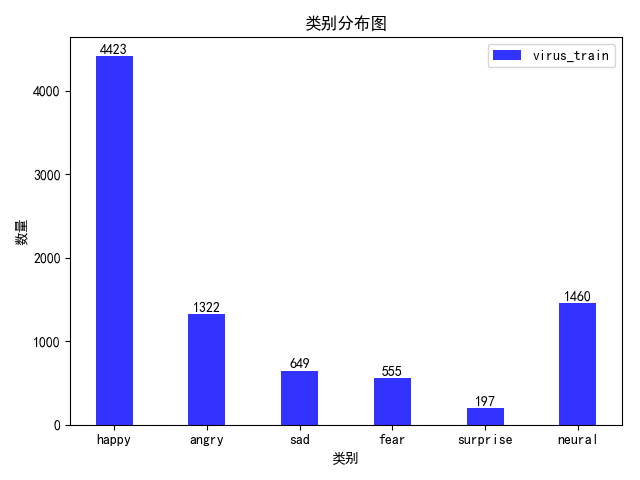
 

图3 virus train文本长度分布 图4 virus eval文本长度分布

从图3和图4可以看出，virus的文本长度分布范围更广，少部分在300~400之间，但大部分都在200以内，所以设置最大长度为200。

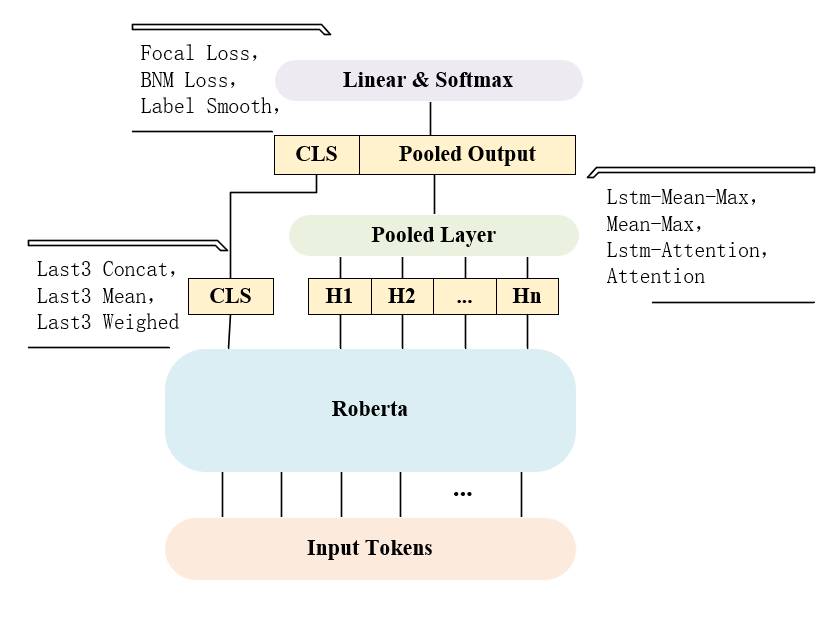
 

**图5 usual train类别分布**  **图6 virus train类别分布**

从图5和图6可以看出，usual和virus数据集都存在类别不平衡的问题，我们尝试使用过采样以及eda数据扩充技术，但模型性能并没有提升。

## 3 集成模型

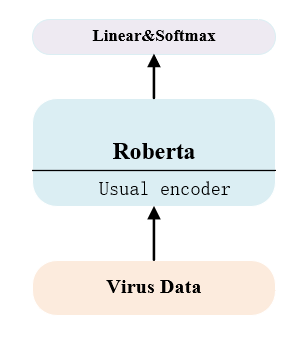
我们在Transformers库中提供的BertForSequenceClassification的基础上，对结构进行了调整，主要分为三部分，分别是对CLS pool方法的调整、对H1-Hn pool方法的调整以及对output layer的调整。模型的通用架构可以看作图7结构，Roberta最终输出的CLS，我们尝试过将其替换为最后三层网络CLS的concat、mean以及weighted，而在最后的linear & softmax层，我们也尝试过focal loss、bnm loss以及label smooth，但模型性能均未提升，所以在这两部分我们最终采用了默认的CLS，以及基础的cross-entropy损失函数。此外我们尝试过四种池化策略处理H1~Hn，分别是mean max、LSTM mean max、attention以及LSTM attention，实验表明后两种池化策略较为有效。



**图7 模型结构图**

### 3.1 迁移学习

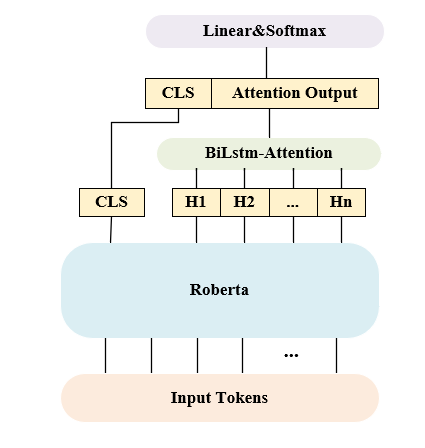
如图8所示，我们用训练好的usual模型的encoder参数初始化virus模型的encoder参数，其它保持不变，有效提升了virus模型的性能，我们认为更大规模的usual训练数据使得encoder能够更好地处理文本的语言结构，另外由于疫情微博数据与通用微博数据的内在相似性，使得encoder能够迁移使用。同时我们也尝试过使用完整的usual模型参数来初始化virus模型，但virus模型性能反而有所下降，显然用usual来初始化virus的深层语义是不适用的。



**图8 迁移学习示意图**

### 3.2 LSTM-Attention

如图9所示，我们尝试在原始Roberta模型后添加BiLstm-Attention，然后将其输出结果与原CLS拼接后输入一个线性层进行softmax分类，在usual数据集中效果较好，在virus上效果欠佳。



**图9 LSTM-Attention示意图**

### 3.3 其他技巧

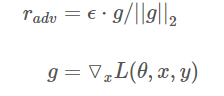
数据预处理方面，我们进行了全角转半角、繁转简、英文大写转小写、去除url、去除email、去除@以及保留emoji等操作，表3展示了部分清洗数据。在模型处理中，我们限制数据的最大长度为140，之前数据分析所使用的数据为原始数据，文本相对清洗后的数据长一点。

表3 数据清洗

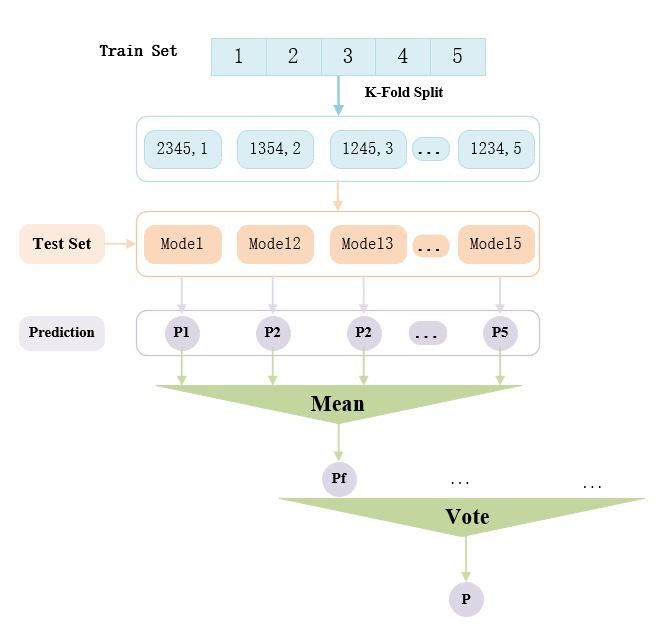
|  |  |  |
| --- | --- | --- |
| 清洗策略 | 清洗前 | 清洗后 |
| 繁简体转化 | 願2015餘下的日子里，美好能夠多一些，快樂能夠如影隨形…Goodmorning! | 愿2015馀下的日子里,美好能够多一些,快乐能够如影随形…goodmorning! |
| 微博@标签 | 保护好他 平平安安//**@朱一龙工作室:**工作室第一时间准备了黑色和蓝色的外科医用口罩、N95口罩、酒精等必备品，彩排和平时也都有戴，请大家放心。也希望大家保护好自己，注意安全。 | 保护好他 平平安安工作室第一时间准备了黑色和蓝色的外科医用口罩,n95口罩,酒精等必备品,彩排和平时也都有戴,请大家放心。也希望大家保护好自己,注意安全。 |
| Email、url | 吃野味的以后看好自己的妈吧。[太开心][太开心][太开心]#全国确诊新型肺炎病例# **http://t.cn/RDUnNFD** ??西安 | 吃野味的以后看好自己的妈吧。《太开心》《太开心》《太开心》#全国确诊新型肺炎病例# ??西安 |

伪标签，使用训练好的virus模型去标注virus eval set，从2000条标注结果中随机选取700条，有效提升了virus在eval set上的性能，我们也尝试使用usual去标注usual eval set，但性能并没有提升。

FGM对抗学习[2]，通过对输入文本的embedding vector添加扰动，提高模型的泛化能力。如下所示，为1，是embedding vector的梯度。



集成方法，我们尝试过stacking集成以及投票集成，stacking集成的效果略逊于投票集成，因此我们最终选用投票集成方法。k折中，我们将model1~model5的输出结果的均值作为最终结果， 随后我们将所有结果进行投票，如图10所示。



**图10 模型集成示意图**

## 4.实验结果及分析

表4都是在fgm对抗学习和清洗后的数据的基础上进行的k折实验，除最后一个模型，其它模型都是5折，macro F指标是k折在随机划分的dev上的均值结果（非系统提交结果）。在virus数据集上，roberta large[3]的效果优于roberta wwm ext，但由于之前的疏忽，我们之后都是使用roberta wwm ext作为baseline进行实验，LSTM attention适用于usual数据集而非virus，transfer learning适用于virus而非usual，其中使用伪标签（倒数第二个实验）的效果最好，但提交到系统上的结果并不如人意。

**表4 实验结果**

|  |  |  |
| --- | --- | --- |
| Model | Usual Macro F | Virus Macro F |
| Roberta base | 0.7677 | 0.6801 |
| Roberta large | 0.7696 | 0.6856 |
| Roberta wwm ext base | 0.7698 | 0.6807 |
| Roberta wwm ext large | **0.7744** | 0.6828 |
| Roberta wwm ext large attention | 0.7714 | 0.6876 |
| Roberta wwm ext large lstm attention | 0.7737 | 0.6830 |
| Roberta wwm ext base transfer learning | 0.7680 | 0.6932 |
| Uer mixed Roberta model[6] | 0.7733 | 0.6849 |
| Roberta wwm ext large attention transfer | \*\*\* | 0.6913 |
| Roberta wwm ext base transfer learning (pseudo) | \*\*\* | **0.6998** |
| Roberta wwm ext base transfer learning (10 fold) | \*\*\* | 0.6932 |

表5是我们投票集成所使用到的模型。

**表4 实验结果**

|  |  |  |
| --- | --- | --- |
| Data | Model | Macro F |
| Usual | Roberta Base  Roberta Large  Roberta wwm ext Large  Roberta wwm ext large Lstm Attention  Roberta wwm ext Transfer Learning  Roberta wwm ext Large Attention  Uer mixed Large | 0.7799 |
| Virus | Roberta Base  Roberta Large  Roberta base transfer learning  Roberta wwm ext transfer learning(2个)  Roberta wwm ext transfer learning(10折)  Roberta wwm ext transfer learning(伪标签) | 0.703 |

## 5 总结

此次比赛，我们参考2019 CCF-BDCI的冠军团队“我们都上哈工深”的roberta模型结构，通过对数据清洗、fgm对抗学习、k折训练、迁移学习得到的多个模型进行投票集成，取得了不错的成绩。但遗憾的是，由于缺少误差分析，我们对模型的改动并没有针对数据特性进行优化。

参考文献

[1] Zhou, Peng, et al. "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification." *meeting of the association for computational linguistics* (2016): 207-212.

[2] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification[J]. 2016.

[3] Liu Y , Ott M , Goyal N , et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2019.

[4] Cui Y , Che W , Liu T , et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. 2019.

[5] Cui Y , Che W , Liu T , et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[J]. 2020.

[6] Zhao Z , Chen H , Zhang J , et al. UER: An Open-Source Toolkit for Pre-training Models[J]. 2019.