清博大数据 for SMP2020—EWECT

作者：夏茂晋，关宇航，马云腾，王屹东

北京清博大数据科技有限公司

摘要

本届微博情绪分类评测任务一共包含两个测试集：第一个为通用微博数据集，其中的微博是随机收集的包含各种话题的数据；第二个为疫情微博数据集，其中的微博数据均与本次疫情相关。需要设计模型对输入的微博文本预测出文本中蕴含的情绪标签。模型采用了交叉验证、训练数据权重、数据增强、模型融合、迁移学习等技术。共使用了三种预训练模型分别是 Roberta - Large 、XLNet - Large 、 XLM - Roberta - Large 。使用三种预训练模型采用交叉验证的方法，分别对疫情数据和通用数据训练了一个或多个模型。采用概率平均和投票的方式对不同模型的结果进行融合。通用微博在验证集的最高分为 0.8040，疫情微博验证集最高得分为 0.6929。最终的测试提交的得分是通用数据 0.7823，疫情数据 0.6964。在所有参赛队伍中排名第二，获得二等奖。

关键词：预训练模型，情绪分类，数据增强，迁移学习

1.引言

1.1 任务

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，输出是该微博所蕴含的情绪类别。在本次评测中，我们将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。

1.2 数据介绍

第一部分为通用微博数据集，该数据集内的微博内容是随机获取到微博内容，不针对特定的话题，覆盖的范围较广。

第二部分为疫情微博数据集，该数据集内的微博内容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。

因此，本次评测训练集包含上述两类数据：通用微博训练数据和疫情微博训练数据，相对应的，测试集也分为通用微博测试集和疫情微博测试集。

两个数据集的各类情绪微博举例如下表所示：

训练数据示例

情绪	通用微博数据集	疫情微博数据集
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[吃惊]	对着这个症状，没病的都害怕[允悲][允悲]

情绪	通用微博数据集	疫情微博数据集
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。

表–1

每条微博被标注为以下六个类别之一：nerual（无情绪）、happy（积极）、angry（愤怒）、sad（悲伤）、fear（恐惧）、surprise（惊奇）。

通用微博训练数据集包括 27,768 条微博，验证集包含 2,000 条微博，测试数据集包含 5,000 条微博。

疫情微博训练数据集包括 8,606 条微博，验证集包含 2,000 条微博，测试数据集包含 3,000 条微博。

1.3 报告整体结构简介

本次报告从使用的技术和针对性技巧入手，分别对通用微博模型和疫情数据模型的具体实现流程做详细介绍。包括调优过程中的各种实验方法和结果的分析，和对于本次算法设计比赛的总结反思。

2.模型及方法介绍

为了解决不同数据集上的分类问题，考虑到数据集的分类标准差异，所以针对不同的数据，本组分开训练了两个模型。除了数据上的差异，两个模型的结构上并没有太大差异，主要差异体现在训练过程和参数调整上。两个模型都采用了三个相同的预训练模型，分别是 Roberta、XLNet、XLM-Roberta，构建了六个 fineturn 结构相同的模型，调参和调整模型输入来获得最好的单模型效果。

2.1 预训练模型增量训练

预训练模型的训练是在大量开源数据集下定义不同的训练目标来实现的，并没有对领域的针对性，对于微博文本来说，从文本的长短和文本的写作风格上都有一定的规律，实验证明，用领域内的文本训练的预训练模型用在同一个领域的效果往往要比拿通用预训练模型做的表现要好。在资源紧张的情况下，在通用模型最终 checkpoint 的基础上，组织领域内的文本用小学习率对预训练模型进行细微调整的策略同样能达到近似的效果。

本组采用了疫情期间网民情绪文本（https://www.datafountain.cn/competitions/423），共 1820606 条，对 Roberta 模型进行了动态掩码(dynamic Masking)任务的训练。得到了针对疫情数据的预训练模型。用我们训练的有针对性的预训练模型在不改变任何参数的情况下，疫情数据验证集的结果有一定提升，详情见下表：

增量训练实验

Model	微博数据 F1	微博数据 ACC
Roberta_large	0.6889	0.8050
Roberta_large_Virus	0.6912	0.8060

表-2

2.2 迁移学习

两个数据集合在标签上有一致性，对于通用的情感词例如‘哈哈哈’，无论在那个数据集中，都更倾向于分类到 happy 标签，基于这样的客观事实，我们可以在训练好的适用范围更广的通用模型的基础上，加入疫情标注数据对模型进行适应性训练，相比于在不足一万的疫情数据集上训练得到的模型，这种方式大大增加了模型的泛化能力。实验证明，这种策略可以显著提升疫情测试集的得分，具体情况见下表:

迁移学习实验

Model	微博数据 F1	微博数据 ACC
Roberta_large	0.6889	0.8050
Roberta_large_From_Usual	0.6912	0.8100

表-3

2.3 数据增强

数据增强的方法有很多，传统的方法有翻译回译方法、同义词替换方法等，

本组采用的数据增强的方式类似同义词替换的思路，只是获取同义词的方式不同于以往的词典或者 Word2vec 模型的方式，而是采用 BERT 模型天生的 MaskedLM 能力，对文本中的 Token 随机进行遮挡预测，并选取可能性最大的两个预测结果替换原文中的 Token，最终从一个文本中获得多个生成文本，并控制总体的数据比例，减少数据不均衡带来的影响。但从实验结果来看，这样的技巧没有为我们带来提升，分析增强的数据可以看到，BERT 预测出来的字不是很符号语言规律，融合在原文本中使原来的句子变得晦涩难懂。无法表达原来的意思，生成的文本见表 4，实验结果见表格 5。

Bert 数据增强示例

编号	文本内容	情绪标签
16-原文	为什么泰国治愈率这么高//@英伦圈:???	surprise
16-BERT 生成	为什麼泰国的愈率这么高//@英伦 bbc:???	surprise
16-BERT 生成	为什么泰国治愈率这么高//@英伦圈:??/	surprise
16-BERT 生成	为甚么泰国治愈率这麼高?/@英語圈:???	surprise
16-BERT 生成	为什么泰的出座率那么高//@英伦圈:/??	surprise
16-BERT 生成	為何麼泰国治愈率这么高//@英伦圈:???	surprise
16-BERT 生成	为什么病症治愈率这么高//@英伦圈:???	surprise
16-BERT 生成	为什么泰国治愈率这么高?/@英伦圈:???	surprise
16-BERT 生成	为什么泰囧治愈率这么高//@英伦 m:/??	surprise
16-BERT 生成	为什么泰国的愈率这麼高？？@英伦圈:)??	surprise

表-4

数据增强实验

Model	微博数据 F1	微博数据 ACC
Roberta_large	0.6889	0.8050
BERT 增强_Roberta_large	0.6701	0.7930

表-5

2.4 其他技巧

除此之外，在用不同预训练模型训练的时候，对原样输入做了不尽相同的预处理，例如微博昵称的去除和文本内网址的去除。具体包括哪些处理手段可参考训练代码。

为了完整的利用所有训练数据，我们采用了 9 折交叉验证的训练方式，同时也减少了预测结果的波动。

在检查数据的时候，发现训练集和验证集之间有数据泄露，对于泄露的数据，我们选取了训练集的标注结果对预测结果进行修正，泄露的数据有一百条左右，但预测不一致的数据仅有四条，但是验证集的分数也带来了一定提升。

对于模型融合，本组认为相同的预训连模型的概率输出具有一致性，应该采用概率加权的方式融合模型，而对于不同预训练模型之间，相同的损失函数带来了预测结果的一致性，应该采用投票的方式进行融合，且考虑每个模型在验证集上的得分当做投票权重。最终融合结果在验证集上取得了远超单模型的得分，具体提升见下表。

模型融合

	通用微博 F1	通用微博 ACC	疫情微博 F1	疫情微博 ACC
Roberta	0.7855	0.8045	0.6891	0.8040
XLNet	0.7914	0.8120	0.6734	0.7975
XLM-Roberta	0.7953	0.8140	0.6795	0.7995
Merge	0.8040 / 1	0.8225 / 1	0.6929 / 7	0.8095 / 3

表-6

3.实验结果及分析

在调参的过程中我们做了大量的对比实验，列举以表格的形式记录的实验过程如下：

实验记录

Number	Batch_size	LR	Class_weight	Loss	Score
1	12	5e-6	1,1,1,1.2,1.2,1.2	0.3064	0.6647
2	12	5e-6	1,1,1,2,1.5,1.5	0.3289	0.7059
3	12	1e-5	1,1,1,2,1.5,1.5	0.3977	0.6878
4	12	5e-6	None	0.2632	0.6758
5	2*6（梯度累计）	5e-6	1,1,1,2,1.5,1.5	0.3029	0.6967

表-7

通过上面的实验可一看出来，最好的参数是 Batch_size：12；LR：5e-6；Class_weight：[1,1,1,2,1.5,1.5]。

在不同的模型参数调整的过程中，要做的实验次数远多于我们可进行的测试

集提交的次数，故以上实验数据的获取都是在保留训练数据上得到的实验结论，我们已经尽可能的保持了保留数据的量和标签比例与验证集的一致性，以此来更好的模拟验证提交。

本次最终提交的测试集分数与验证集分数对比如下表所示：

测试分数分析

	最终指标	通用微博 F1	通用微博 ACC	疫情微博 F1	疫情微博 ACC
验证集	0.7485	0.8040 / 1	0.8225 / 1	0.6929 / 7	0.8095 / 3
测试集	0.7393	0.7823 / 3	0.8076 / 1	0.6964 / 2	0.8100 / 4

表-8

分数波动较大的指标在通用微博，且观察其他队伍的最终提交来看，也是同样的情况，故可以得出结论这种降低和使用的模型没有关系，最可能的情况是两两个数据集的数据分布以及其他方面的差异造成的；对于疫情微博的结果分析来看，各单位的的变化都很大，且有的提升有的下降，完全找不到规律，这也在意料之中，从训练集的数据分布中可以看出疫情数据的分布极不平均，较少的类别的正误对整体分值的影响很大。由于本组的预测结果来自于多个模型的融合结果，且模型之间的差异比较大，所以在泛化能力上有较强的适应性，所以在分布极不平均的数据集上是的最终分数有所提升。

4.总结

本次测评的任务在实际生产中有直接的应用，具有很高的实际价值。通过这样的算法竞赛，不仅促进了相关领域的发展，而且为人工智能在国内的发展起到了积极作用。

参加本次测评的一个多月以来，通过对排名分数的追求，我不断探求模型的极限，在预训练模型的应用上更加娴熟多变，不再局限于工具类的模型结构束缚，通过一次次实验，积累了大量经验可用于今后的生产模型的构建中去。希望以后可以更多的参与到这样的测评中去，精进自己技能的同时，也为人工智能的发展贡献绵薄之力。