拿第一导师请吃肯德基

for SMP2020-EWECT

作者：张辉，于凤娇，殷峥　指导老师：曹玖新 教授

东南大学

摘要

2020 年，新冠肺炎疫情的爆发引起了全国人民的关注，新浪微博成为了广大人民群众发声、表达观点的重要平台，其中产生的大量社交情感数据不仅有助于情感分析的研究，也为疫情的宏观调控提供了新的可能性。基于自然语言处理技术识别并分析文本中蕴含的情感信息一方面有利于政府及时了解网民的情绪波动、把控整体舆情的变化态势，另一方面也有助于第一时间发现疑似病例，从而更有针对性地进行舆情检测、预警和引导，从物理和网络两个维度更有力地进行疫情防控。本次评测任务给定了通用微博和疫情微博两个数据集，旨在对通用微博进行情感分析的基础上，借由迁移学习方法，更好地将通用微博的情感知识迁移到疫情微博中，从而能够快速构建针对疫情相关微博的情绪分类模型。

本文提出的模型主要基于对 RoBERTa 以及 BERT 模型的改进，使用迁移学习将通用微博中获取的知识迁移到疫情微博中，最后将多个模型利用 Stacking 技术进行集成，在疫情微博数据上取得了较好的结果。在 SMP2020 微博情绪技术分类评测任务中线上总排名最终达到第三，其中疫情微博模型的 F1 值在验证集上达到 0.6982，排行第二，在测试集达到 0.6932，排行第四，证明了迁移学习

和集成学习在情绪分类任务上的有效性和较好的泛化能力。

关键词：情感分析，社会媒体情绪分类，迁移学习，stacking 集成

1. 引言

1.1 任务描述

本次评测任务描述如下：

微博情绪分类任务旨在识别微博中蕴含的情绪。输入是一条微博文本，输出是该微博所蕴含的情绪类别。在本次评测中，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。本次评测任务需要对通用微博数据集和疫情微博数据集分别做情绪分类。

1.2 数据介绍

数据集共分为两部分：

通用微博训练数据集是随机获取到的微博文本，不针对特定的话题，覆盖的范围较广。通用训练数据集包括 27,768 条微博。

疫情微博训练数据集是在疫情期间使用相关关键字筛选获得的微博文本，其内容与新冠疫情相关。疫情训练数据集包括 8,606 条微博。

与传统的情感极性二分类任务不同的是，本次评测是一个细粒度情绪多分类任务。一方面，同一文本可能会蕴含不同的情绪，导致任务中微博的具体情感难以辨别；另一方面，具体情感标签中如 happy 和 surprise、sad 和 angry，这些标签本身由于标注人员的主观性存在歧义。因此本次评测任务相比普通的情感极性分类任务存在更大的挑战。

1.4 报告结构

本次报告共分为四个部分：

第一部分是引言，包括任务描述、数据介绍以及论文的整体架构。

第二部分是模型及方法介绍，首先介绍本次评测所使用的通用微博模型及疫情微博模型，然后介绍本次实验过程中所使用到的其他技巧性的尝试。

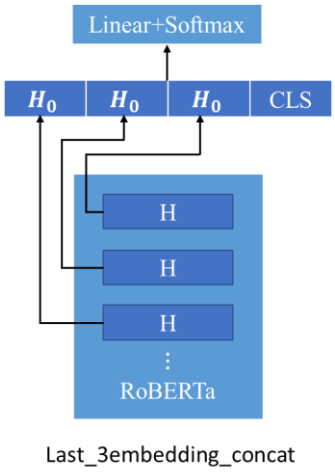
第三部分是实验结果分析，包括实验所用不同模型、结果后处理、融合策略等的对比实验分析。

第四部分是最后对本次比赛的总结。

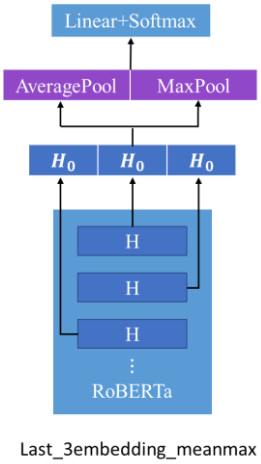
2. 模型及方法介绍

本文模型结构主要对 RoBERTa 以及 BERT 模型进行改进，使用迁移学习将通用微博中获取的信息迁移到疫情微博中，将多个模型利用 Stacking 进行集成，在疫情微博上取得了较好的结果。

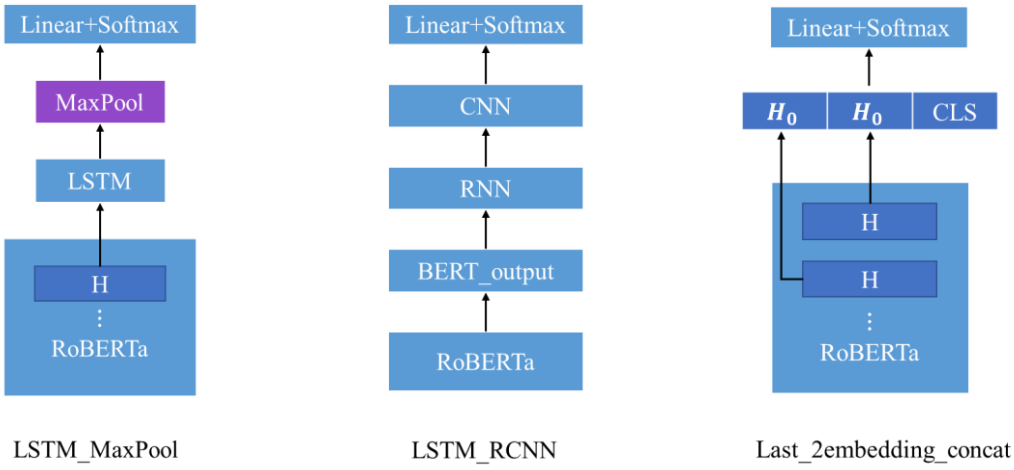
2.1 通用微博模型



模型结构 1



模型结构 2



LSTM_MaxPool　　LSTM_RCNN　　Last_2embedding_concat

模型结构 3　　模型结构 4　　模型结构 5

图 2.4 模型结构图

本文在 RoBERTa 以及 BERT 模型基础上进行改造，构建大量模型架构，最终根据实验效果，选取了其中部分模型架构，预训练权重选择 robert-wwm-ext-large 以及 bert-wwm-ext，下面对这几种模型架构简要介绍。

1. Last_3embedding_concat：取预训练模型的最后三层 embedding 向量与 cls 向量进行拼接，传入 linear 层得到预测结果；
2. Last_3embedding_meanmax：取预训练模型的最后三层 embedding 向量分别进行 mean-pooling 和 max-pooling，将 pooling 的结果拼接传入 linear 进行分类；
3. LSTM_MaxPool：取预训练模型的最后一层 embedding 向量传入双向 LSTM 中，叠加 max-pooling 层和 linear 层进行分类；
4. LSTM_RCNN：取预训练模型的最后一层 embedding 向量输入双向 LSTM，叠加 CNN 和 linear 层进行分类预测；
5. Last_2embedding_concat：取预训练模型的最后两层 embedding 向量与 cls 向量进行拼接，传如 linear 层进行分类

根据实验结果，最终通用微博模型选取模型结构 1，并且由于模型融合效果不佳，

通用微博模型最终使用该单模型。

2.3 疫情微博模型

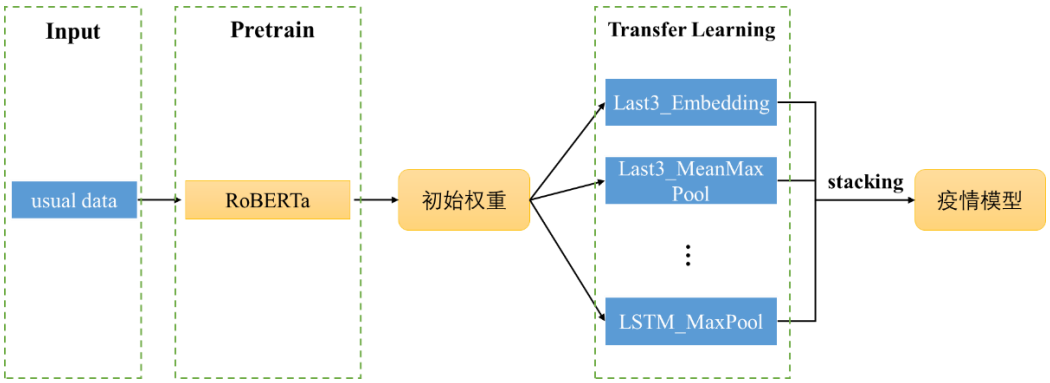


图 2.5 疫情微博模型训练框架

最终疫情微博模型训练流程如图所示。首先使用通用微博数据集在预训练模型 RoBERTa 上进行后训练，将获得的预训练模型参数作为疫情模型的初始权重，然后通过相同方式在不同的模型结构上进行训练，最终选取线下结果较优的多个单模型进行 Stacking 集成，获得了较好的效果。

2.4 其他技巧

在构建单模型时使用多折交叉融合，其中 usual 数据使用 6 折交叉，virus 数据使用 5 折交叉。具体过程为将数据分为多折，选取每一折效果最优的模型，预测时将每折结果进行概率平均，可以充分利用全部数据集进行模型构建。

对于学习率，使用多层学习率以及学习率衰减策略。任务预测层（Task Specific Layer）学习率使用 1e-4，预训练模型内部学习率使用 1e-5，且内部学习率随层数降低逐步衰减。

考虑类别不平衡的情况，尝试了同义替换、过采样、欠采样等方式改善数据

集分布，通过标签平滑、focal loss 等方式改善模型训练过程，通过 F1 指标优化进行结果后处理。实验表明 F1 指标优化方法对最终预测结果有较大提升。

训练过程中使用 FGM 在 embedding 层进行对抗扰动，使用 multi-sample-dropout 策略加快收敛，缓解过拟合风险。这部分策略对于最终结果没有非常明显的提升。

尝试不同的预训练模型，包括 NEZHA、Electra、ALBERT、BERT、RoBERTa、ERNIE 等，考虑到最终实验结果和模型训练时长，最终选取线下结果较好的 RoBERTa-large 预训练模型，以及选取较优的 BERT-base 模型增加模型的多样性。

尝试使用通用微博数据进行领域预训练，通过情感极性微博数据集进行相关任务预训练，最终结果并未有明显提升。

3. 实验结果及分析

3.1 对比试验

1. 多折交叉验证

多折交叉融合(usual)	线上F1(usual)
6折	0.7870
7折	0.7783
10折	0.7833

图 3.1 交叉融合实验

对于 usual 数据集，通过 6 折、7 折和 10 折进行对比实验，最终选取 6 折。

对于 virus 数据集，由于数据量较小，保持 5 折。

2. 不同的模型结构

数据集	模型	F1（线上）
Usual	Last3_Embedding	0.7870
	LSTM_MaxPool	0.7818
	RNN_CNN	0.7817
Virus	Last3_Embedding	0.6936
	Last3_MeanMaxPool	0.6905
	LSTM_MaxPool	0.6875

图 3.2 模型结构实验

对于 usual 数据集和 virus 数据集，last_3embedding 模型结构都表现的最好。

3. 结果后处理

数据集	F1处理前	F1处理后
Usual	0.7834	0.7870
Virus	0.6918	0.6936

图 3.3 F1 指标优化处理对比

F1 指标优化方法对 usual 与 virus 数据集均有较大幅度的效果提升

4. 结果融合

数据集	融合方式		F1(线上)
usual	概率平均融合		0.7837
	投票融合		0.7848
	加权融合		0.7816
	Stacking		0.7784
virus	概率平均融合		0.6872
	投票融合		0.6868
	加权融合		0.6906
	Stacking+GBDT	Step=30	0.6941
		Step=50	0.6982
		Step=70	0.6976

图 3.4 集成融合实验分析

对于通用微博 usual 数据集，将多个模型进行融合后效果均没有单模型结果好，所以最终提交时仅选取最优单模型提交。单模型本身复杂度较低，过拟合风险较小。最终结果也表明在 usual 数据集上验证集 F1 值和最终测试集 F1 值波动

很小。

对于疫情微博 virus 数据集，使用投票融合、概率平均融合和加权融合的集成策略均没有得到有效的提升，当使用stacking集成后，线上结果有较大的提升，故本文最终提交方案选取了 Stacking+GBDT 方法对多模型进行集成融合。

4. 总结

参加此次比赛获得了很大的收获，其中之一就是要相信自己的线下结果。因为此次赛制要求，在最终测评时需要替换测试数据集，验证集结果仅为一个参考，不参与最终排名，因此更考验模型的泛化能力。最后结果也表明，本队伍的最终测试集的 F1 值和线下测试结果基本一致。

另外也总结了本次比赛中有待提升的地方：

1. 由于赛事时间原因，虽然尝试了很多思路，但其中许多思路仅仅是浅尝辄止，在发现效果没有明显提升后都选择了放弃，并没有更深入的进行修改尝试。
2. 比赛中花费较多时间进行调参，对问题本身的考虑有所欠缺。虽然分析了一些 Bad Case 与问题难点，做了一些尝试，但是由于时间限制并没有很好地解决这些问题。
3. 由于本次比赛是给定全新数据集进行预测，并且测试集数据量较大，时间也较短，所以需要提前将模型进行复现，方便直接进行推理预测。这就需要及时规划好实验结果整理、代码整理，充分利用团队协作。