ERNIE-CNN for SMP2020-EWECT

作者：黄江平，刘凡，杨苗苗

重庆邮电大学

摘要

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，数据集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。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。在本次评测中，我们使用了 ERNIE 预训练模型作为嵌入层，再传入卷积神经网络进行特征提取，最后进行分类。在此次评测中，此模型在通用微博测试集和疫情微博测试集的 F1 值分别为 73.63%和 56.05%，虽然评测成绩不是很理想，但是希望通过提交评测技术报告能够对中文信息处理做出一点点贡献。

关键词：自然语言处理，情感分析，社会媒体情绪分类

1.引言

情感分析技术一直是自然语言处理领域研究的重点内容之一。2020 年，新冠肺炎疫情成为了全国人民关注的焦点，众多用户针对此次疫情在新浪微博等社交媒体平台上发表自己的看法，蕴含了非常丰富的情感信息。基于自然语言处理技术自动识别社交媒体文本中的情绪信息，可以帮助政府了解网民对各个事件的态度，及时发现人民的情绪波动，从而更有针对性地制定政策方针，具有重要的社会价值。尽管之前的社交媒体情感分析技术已经取得了不错的进展，但是如何将之前的研究成果快速高效地应用到疫情相关的数据当中，仍然是一个值得研究的问题。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。

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，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。

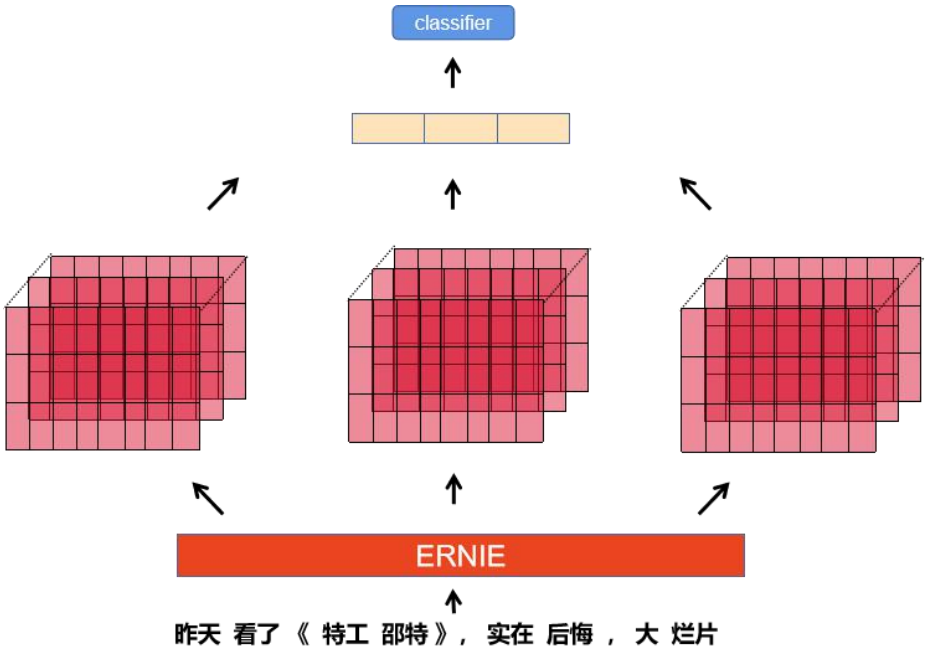
其中通用微博训练数据集包括 27,768 条微博，验证集包含 2,000 条微博，测试数据集包含 5,000 条微博。疫情微博训练数据集包括 8,606 条微博，验证集包含 2,000 条微博，测试数据集包含 3,000 条微博。

本次评测报告结构如下，第 2 节介绍了我们团队提交的模型以及对于数据的处理方法。第 3 节介绍了使用 TextRCNN，DPCNN，FastCNN 等模型在数据集上的进行的相关实验。第 4 节进行了本次评测的总结。

2.模型及方法介绍

随着深度学习的发展，预训练模型在 NLP 的诸多任务中取得了很不错的成绩，在文本分类上也是如此，因此，在一些 NLP 任务中使用预训练模型可以得到更好的效果，而且在模型的训练中还可以加速训练，使得模型收敛更快通常会使得模型达到更好的效果。因为预训练模型中的参数都是从大量数据中训练得来，比起在自己的数据集上从头开始训练参数，在预训练模型参数基础上继续训练的方式肯定要快一些。还有就是预训练模型是通过海量数据训练得来，可以更好地学到了数据中的普遍特征，比起在自己的数据集上从头开始训练参数，使用预训练模型参数通常会有更好的泛化效果。

在本次的评测任务中我们使用了 ERNIE 预训练模型作为嵌入层，再用卷积网络进行卷积，最后将结果进行全连接进行分类。整体结构如下所示，



2.1 预训练模型 ERNIE

百度利用大规模文本语料库和知识图训练了一个增强语言表征模型（ERNIE），其可以同时利用词汇、句法和知识信息。ERNIE 模型通过建模海量数据中的实体概念等先验语义知识，学习完整概念的语义表示。即在 Masked LM 中通过对词和实体概念等语义单元进行 mask 来预训练模型，使得模型对语义知识单元的表示更贴近真实世界。引入多源数据语料训练 ERNIE。包括百科类，新闻资讯类、论坛对话类数据来训练模型。尤其是论坛对话语料的引入，文章认为，“对话数据的学习是语义表示的重要途径，往往相同回复对应的 Query 语义

相似”。基于该假设，ERINE 采用 DLM（Dialogue Language Model）建模 Query-Response 对话结构，将对话 Pair 对作为输入，引入 Dialogue Embedding 标识对话的角色，利用 Dialogue Response Loss 学习对话的隐式关系，通过该方法建模进一步提升模型语义表示能力。使用 ERNIE 进行特征提取(1) 对实体概念知识的学习来学习真实世界的完整概念的语义表示；ERNIE 对训练语料的扩展尤其是论坛对话语料的引入，更加的增强模型的语义表示能力，对微博的语义特征提取有很好的效果。

2.2 卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）是一种前馈神经网络，它的人工神经元可以响应一部分覆盖范围内的周围单元。它包括卷积层(convolutional layer)和池化层(pooling layer)。第一次卷积可以提取出低层次的特征。第二次卷积可以提取出中层次的特征。第三次卷积可以提取出高层次的特征。特征是不断进行提取和压缩的，最终能得到比较高层次特征，简言之就是对原式特征一步又一步的浓缩，最终得到的特征更可靠。利用最后一层特征可以做各种任务：比如分类等。我们使用 ERNIE 进行特征提取后，再使用卷积神经网络进行卷积，再将得到的结果进行全连接，最后进行情绪标签预测。

2.4 其他技巧

在分类任务中，数据集可能存在标签类别不平衡的情况，某种类别特别多，在分类模型当中，经常对标签使用 one-hot 的形式，然后去预测样本属于每一个标签的概率，如果不考虑多标签的情况下，通常选择概率最大的作为我们的预测标签。在实际过程中，这样对标签编码可能存在两个问题：（1）可能导致过拟合；（2）模型对于预测过于自信，以至于忽略到可能的小样本标签。通用微博数据集如下表所示，其中“fear”标签数量较少，会影响到模型的训练效果

情绪标签	Angry	Fear	Happy	Neural	Sad	Surprise	合计
数量	8344	1220	5379	5749	4990	2086	27768
长度	0–30	30–50	50–70	70–90	90–100	>100	合计
数量	14887	7059	2624	1530	452	1473	27768

表 1：数据集标签和长度分布

3.实验结果及分析

数据集设置，按照 8:2 的比例分割评测所给训练集数据，训练数据 26374 条，开发集 4000 条，其中通用微博划分 3000 条，疫情微博划分 1000 条，通用微博

测试集 3000 条，疫情微博测试集 2000 条。

模型	平均指标	通用微博Macro_F	通用微博Accuracy	疫情微博Macro_F	疫情微博Accuracy
TextCNN	0.5969	0.6424	0.6887	0.5515	0.7235
TextRCNN	0.5919	0.6615	0.7000	0.5223	0.7035
TextRNN_Att	0.5912	0.6636	0.6920	0.5188	0.6755
FastText	0.5862	0.6479	0.6960	0.5246	0.7171
TextRNN	0.5482	0.6316	0.6643	0.4648	0.6500
DPCNN	0.5361	0.6084	0.6547	0.4639	0.6230
Transformer	0.4814	0.511	0.6003	0.4519	0.6540
ERNIE	0.6868	0.7504	0.7783	0.6232	0.76
ERNIE_RNN	0.6851	0.7342	0.7737	0.636	0.7715
ERNIE_CNN	0.6873	0.729	0.7677	0.6456	0.7765
ERNIE_RCNN	0.6786	0.7233	0.7633	0.6339	0.7735
Bert_CNN	0.6679	0.7239	0.7613	0.612	0.7635
Bert	0.6654	0.7333	0.7657	0.5976	0.7465
Bert_RNN	0.6634	0.7082	0.7487	0.6186	0.7605
Bert_RCNN	0.6548	0.7034	0.74	0.6063	0.7495
Bert_DPCNN	0.6262	0.701	0.74	0.5515	0.7135

数据集	数量
train	26374
dev	4000
usual_test	3000
virus_test	2000

表 2：数据集数量分布

实验参数设置，其中未使用预训练模型的模型将初始学习率设为 0.001，填充长度为 75，训练迭代次数设为 20。使用预训练模型将初始学习率设为 1e-5，训练迭代次数 5 次，在训练过程中，都应用了 Adam 优化。

表 3：实验结果

我们分别尝试了目前比价经典的分类模型，其中卷积神经网络比循环神经网络表现更加好一些，因为卷积神经网络相比于循环神经网络能够更加好的抓取关键词信息对于短文本，CNN 配合 Max-pooling 池化(如 TextCNN 模型)速度快，而且效果也很好。因为短文本上的关键词比较容易找到，而且 Max-pooling 会直接过滤掉模型认为不重要特征。虽然 Attention 也突出了重点特征，但是难以过滤掉所有低分特征。。TextCNN、TextRCNN、FastText 等未使用预训练模型的模型在分类效果上要比使用了预训练模型的模型要差一点。这表明使用预训练模型可以提高模型文本分类效果，但是在训练中，使用预训练模型需要的计算资源和耗费的时间也要多得多。表 3 显示，在使用了预训练模型的模型中，使用了 ERNIE 作为嵌入层表现要比使用 Bert 预训练模型更加好一些。在其中我们选择了表现最好的 ERNIE_CNN 模型作为本次评测提交的模型，其中 ERNIE 作为嵌入层，再使用卷积神经网络进行特征卷积，再将卷积得到的特征进行全连接分类。

4.总结

参加本次评测收获良多。对于一个评测来说，也是一个竞赛，参加一个竞赛的过程中也是对个人的全方面的能力的考验，遇到不懂的问题，通过查阅资料，在老师指导下找到解决的办法，这也是一个个人全面提升的过程，在这个过程中所得到的经验对于以后的学习、工作、生活也很重要，我们所收获的不仅仅是理论知识和经验，更是有对于问题的探索精神。对于本次评测任务内容，微博是现如今常用的交流方式，微博文本一般都长度差距较大，内容上含有较多的表情符号和一些自制符号、新生成词等，内容非常复杂多变，处理起来难度较大。但是随着深度学习和大数据的发展，各类模型在 NLP 的各类任务中也表现不俗，我们在此次评测中尝试了现如今已出现的多种经典的分类模型，通过此次评测也对它们有了更加深入的了解，虽然本次评测任务成绩不太理想，但是也提交了本次评测的技术报告，为促进中文信息处理做出一点点我们的努力，也希望中文文本信息处理发展能够越来越好。