互联网新闻情感分析

我们都上哈工深

团队简介

本团队名为“我们都上哈工深”，团队中的四名成员目前均就读于哈尔滨工业大学（深圳），是计算机科学与技术学院智能计算研究中心的硕士研究生。近年来，智能计算研究中心主要以自然语言处理和机器学习技术为核心，开展基于海量信息的智能金融、健康医疗大数据挖掘、生物信息学等领域的研究。而作为自然语言处理的学习者，团队的四名成员希望通过这次比赛，巩固知识，提升技能，同时认识更多优秀的同学，一起交流进步。

摘要

随着网络新闻服务的飞速发展，网络上产生了大量的新闻文本信息，探索新闻文本背后的情绪表达，可以为政府和企业提供潜在价值。本赛题给定互联网新闻文本标题及新闻文本内容，要求判断新闻的情感极性（包括正面极性，中性极性和负面极性），是自然语言处理领域的典型分类任务。针对该任务，本文采用了RoBERTa模型，并改造了多个上层模型并通过投票融合的方式取得了较好的结果。在CCF BDCI的新闻情感分析的评测任务上，该方法在最终的B榜评测数据上，F1分值达到了0.81697。

关键词

新闻文本，情感分析，文本分类

1 引言

随着互联网的快速发展，网络新闻服务已成为主要的网络信息载体。越来越多的民众通过网络新闻了解时事动态，并通过评论的形式表达观点、抒发情感。了解新闻内容中的情感，可以掌握舆论的态度和情绪变化，有助于政府和企业实现舆情分析和监控，并及时应对处理突发的舆情事件。在如今信息爆炸的时代，这些新闻文本每天以指数级的数量增长，仅靠人工进行分析需要消耗大量的人力和时间。因此，自动识别网络新闻中所表达的正负面情感，具有重要的理论意义和实用价值的。而互联网新闻的情感分析可以抽象为自然语言处理领域的文本三分类问题，根据新闻的标题和内容，判定该新闻背后的情感极性。

2 任务分析

2.1 赛题描述

本赛题是对互联网新闻长文本进行情感极性分类，其中正面情绪对应0，中性情绪对应1，负面情绪对应2。而评分指标采用Macro-F1值，具体如下：

(1)

(2)

(3)

其中，TP是真样例，FP是假阳例，FN是假阴例，通过以上公式得到该类F1值，将每一类F1值求平均，即得到Macro-F1值。

2.2 数据分析

本赛题的原始特征是新闻的标题和新闻的内容。因此我们对新闻标题和新闻内容进行简单统计分析。图1中前三个个小图分别是关于新闻标题，新闻内容以及新闻总长度的字数分布情况。可以观察到，大部分的新闻长度都集中在0到5000之间。

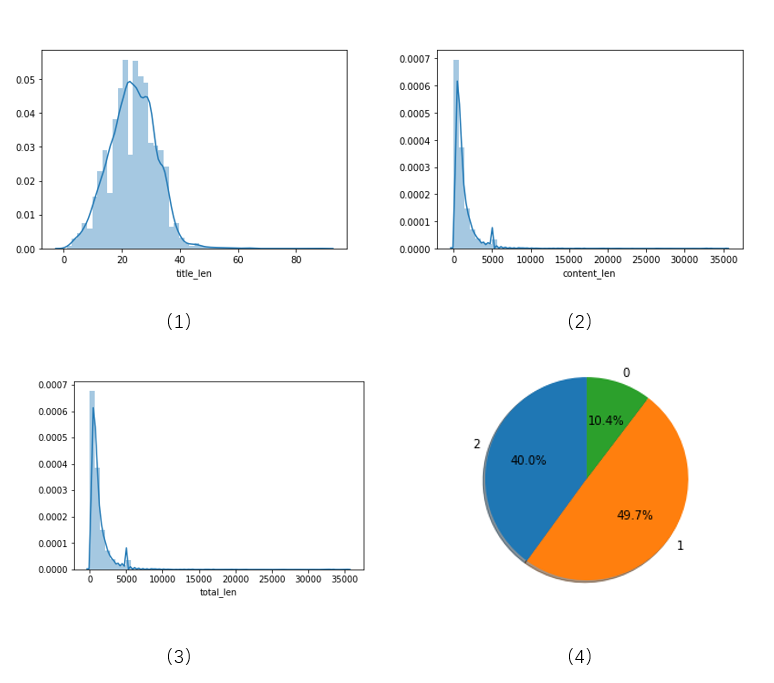


图1：对数据集的简单统计分析

此外，我们还对训练数据集的新闻情感极性类别进行统计，如图1中第四小图所示，发现三个类别的分布并不是特别均匀，特别是0类即正面情绪样本仅占10.4%，需要在后续的方案中考虑这一因素。

3 实验

3.1 模型结构

基于RoBERTa模型[1]，进行上层结构改造，具体包括如下方法：

1．模型分为两部分，可以看做下层是n个RoBERTa模型（参数共享），把新闻文本分成n段后分别输入，上层使用双向GRU来连接n个RoBERTa的pooler\_output输出，最后进行分类。这样可以解决用RoBERTa处理长文本的问题。见图2模型结构1。

图2：基于RoBERTa模型的五种结构

2．提取bert模型内部隐藏层第一状态位输出。通过提取最后2层或最后3层的第一个状态输出向量，与原先的bert模型pooler\_output拼接，最后进行分类。见图2模型结构2,3,4。

3．利用bert模型最后一个隐含层所有状态位信息。这些状态信息可以连接更深层的网络模型，如双向LSTM[3]、双向GRU[4]。通过更深层的网络模型提取文本的更高维度特征，然后通过提取隐含层状态、平均池化、最大池化等操作聚合双向GRU输出和隐含层状态特征，最后拼接bert模型的pooler\_output进行分类。见图2模型结构5。

3.2 模型参数与融合细节

比赛中，我们尝试了多种bert预训练模型[2]，其中准确率最高的是RoBERTa-wwm-ext-large[5]，但其显存占用也较多。以bert\_gru为例，为节省显存，finetune时单GPU的batch 大小参数即per\_gpu\_train\_batch\_size设为4，gradient\_accumulation\_steps设为4，这样每次输入1个样本，梯度累计4次后，再进行反向传播更新参数，以牺牲一定训练速度来节省显存；参数learning\_rate设为1e-5，采用三角学习率，首先warm\_up，学习率逐渐变大，再linear lr decay，学习率逐渐变小，有效改善训练效果。

对于每个模型，我们采用了分层抽样5折交叉验证，分层抽样保证了每折数据集中的各类别样本比例保持不变，每折选择验证集上F1值最高的模型对测试集进行预测，最后5折融合采用概率平均。不同模型融合采用简单投票，得票数最多的类别为最终预测结果。

3.3 实验结果与分析

表1：实验结果

|  |  |  |
| --- | --- | --- |
| 模型 | 数据 | 复赛A榜F1分值 |
| bert\_gru | 头512 | 0.8110 |
| 头256+尾256 | 0.8126 |
| 头300+尾300 | 0.8117 |
| 头512（伪标签） | 0.8123 |
| 头256+尾256（随机种子24） | -- |
| bert\_last2embedding\_cls | 头512 | 0.8157 |
| 头512（随机种子24） | 0.8133 |
| 头512（数据清洗） | 0.8153 |
| 头512（数据清洗，伪标签） | 0.8221 |
| bert\_lstm\_gru | 头512 | 0.8091 |
| bert\_last3embedding | 头512 | 0.8150 |
| bert\_last3embedding\_cls | 头512 | 0.8157 |
| 融合（以上12个模型简单投票） | -- | 0.8233 |

以上模型均为分层5折交叉验证后的结果，数据列中“头512”指的是截取新闻头部512个字符训练和预测；“头256+尾256”是截取头部256个字符和尾部256个字符分别输入BERT, 然后在上层使用双向GRU连接；伪标签是将预测结果加入训练集，在训练过程中引入测试集分布，从而提高模型性能，我们选择每折加入10%的测试集；数据清洗中，一个是将url、html等无关信息去除，另一个是规范标点符号，新闻中很多标点符号本应该为中文标点符号，却写成了英文标点符号，我们进行了改正。

从结果可以看出，“头256+尾256”比“头512”的效果更好，可能头尾包含的情感信息比头部要更多。我们采用的数据清洗方法反而会使效果变差，伪标签的加入能有效提高单模分数。

4 结论

本文介绍了“我们都上哈工深”团队参加互联网新闻情感识别比赛的整体思路和具体方案。比赛中，团队采用了基于RoBERTa的多种改造模型，在每一模型基础上进行5折交叉验证，然后投票融合进一步提高分数，最终取得了不错的性能。

致谢

本次参赛所使用的代码是基于HunggingFace团队的开源项目transformers以及参考郭达雅的开源Baseline进行修改和实现的，主要使用的预训练模型为崔一鸣提供的RoBERTa中文预训练模型，在此向他们表示衷心的感谢；同时，比赛过程中也获得很多参赛选手的支持和帮助，在此对他们的无私和热心表示感谢。

参考

[1] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[2] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[3] Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//Fifteenth annual conference of the international speech communication association. 2014.

[4] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[5] ymCUI, Chinese-BERT-wwm, https://github.com/ymcui/Chinese-BERT-wwm.