



引用格式:姚妮,高政源,娄坤,等. 基于 BERT 和 BiGRU 的在线评论文本情感分类研究[J]. 轻工学报,2020,35(5):80-86.

中图分类号:TP183 文献标识码:A

DOI:10.12187/2020.05.011

文章编号:2096-1553(2020)05-0080-07

基于 BERT 和 BiGRU 的 在线评论文本情感分类研究

Research on sentiment classification for online reviews based on BERT and BiGRU

姚妮,高政源,娄坤,朱付保

YAO Ni,GAO Zhengyuan,LOU Kun,ZHU Fubao

郑州轻工业大学 计算机与通信工程学院,河南 郑州 450001

College of Computer and Communication Engineering, Zhengzhou University of Light Industry,
Zhengzhou 450001, China

关键词:

深度学习;情感分类;
BERT;Word2Vec;
BiGRU

Key words:

deep learning;
sentiment classification;
BERT;Word2Vec;
BiGRU

摘要:针对互联网用户在线评论文本情感分类不准确的问题,提出一种基于 BERT 和 BiGRU 的在线评论文本情感分类模型. 该模型首先使用 Word2Vec 框架对文本内容进行词向量表示,然后利用 BERT 预训练语言模型提取词向量的深层动态表示,最后将其输入 BiGRU 网络进行情感分类. 实验结果表明,与双向 LSTM 结合 Attention 机制模型(W2V-BiLSTM-Attention)、传统卷积神经网络模型(W2V-CNN)和传统循环神经网络模型(W2V-RNN)相比,本文模型的 *MicroF1* 值最高(0.91),分类效果最好.

收稿日期:2020-07-01

基金项目:河南省 2020 年科技发展计划项目(202102210384);郑州轻工业大学 2019 年众创空间孵化项目(2019ZCKJ228)

作者简介:姚妮(1978—),女,湖南省桑植县人,郑州轻工业大学实验师,主要研究方向为智能医疗、信息处理技术.

Abstract: Aiming at the problem of inaccurate sentiment classification for online comment texts of Internet users, an online reviews sentiment classification model was proposed based on BERT and BiGRU. The model used the Word2Vec framework to represent the word vector of the text content, then extracted the deep dynamic representation of the word vector by the BERT pre-training model, and finally input it into the BiGRU network for sentiment classification. The experimental results demonstrated that compared with the dual-path LSTM combined with Attention mechanism model (W2V-BiLSTM-Attention), traditional convolutional neural network model (W2V-CNN) and traditional recurrent neural network model (W2V-RNN), the *MicroF1* value of this model was the highest (0.91) with the best classification results.

0 引言

随着信息技术水平的不断提升和互联网产业的飞速发展,越来越多的互联网应用已经渗入到人们生活的方方面面.普通用户与网络应用之间的交互越来越频繁,互联网用户群体的角色逐渐从互联网内容信息的浏览者演变为创造者.在这个过程中,用户可以在媒体平台上提出情感态度型的观点和评论,对其进行检测和分类不仅可以产生巨大的商业价值,还可以维护互联网环境的安全与整洁.因此,对互联网在线评论数据中的主观性文本进行情感分析有着重要意义^[1].

传统的情感分析方法主要是使用情感词典进行情感分析,B. Liu 等^[2]提出了几种构建情感词典的方法,通过对情感极性的标注和匹配,实现了对情感类型的计分和分类;周咏梅等^[3]使用信息熵来识别微博中评论的情感类型,结合 SO-PMI 算法选择情感词,构建了中文微博的情感词典.但由于信息量的不断增长,以及新型词汇的不断出现,基于情感词典的情感分析方法效果欠佳.近年来,机器学习方法的研究和应用不断深入,已逐渐成为情感分析的主流. B. Pang 等^[4]最早使用机器学习方法进行电影评论数据的情感分析,实验结果表明,基于机器学习的智能算法明显优于基于情感词典的规则算法.姜杰^[5]结合领域规则和机器学习方法提取更为丰富的情感特征,将其融入机器学习分类模型,在微博情感分类实验中分类性能明显优于规则算法.由于文本数据

的复杂多变,所以传统的机器学习方法无法学习到文本中的深层次语义信息,进而导致在一些情感分析任务中无法准确分类.

深度学习方法具有更好的特征表示能力和更高的分类能力,因此基于深度学习的情感分析方法成为研究热点.王利利^[6]通过规范化字符层面上的卷积神经网络(CNN)进行情感分类研究. R. Socher 等^[7]提出了循环神经网络(RNN)、递归神经张量网络(RNTN)等多个递归神经网络,其中 RNTN 模型通过使用句法分析树来获取词语的情感信息,然后通过求和来确定语句的情感类别.随着迁移学习的不断发展,预训练语言模型为基于大数据的情感分类提供了新的思路.鉴于此,本文拟基于 BERT (bidirectional encoder representations from transformers) 预训练语言模型,结合双向长短期记忆神经网络(BiGRU)设计 BERT + BiGRU 分类模型,对互联网在线文本评论数据进行情感分类,以提高情感分类的准确性.

1 数据与文本格式化

1.1 数据来源

本文采用国内图灵联邦平台上的情感分类练习赛所提供的数据集^[8]来进行模型的训练和实验评估.该数据集包括互联网线上各大电商平台商品评论数据、新闻媒体平台的新闻评论数据等共 16 000 余条.数据样例如图 1 所示,其中在 label 列中,0 代表消极评论(负向),1 代表积极评论(正向),2 代表中性评论(中性).

	id	text	label
1	10315	性价比不错,但降价太快。。,买时4199元十天前,现在3999	2
2	9906	房间温度就最低,而且外面也不是很冷,真是想不明白,换了一个房间,还是一样,只好退房了!	1
3	15220	热水器很漂亮,而且比实体店实惠多了。配的配件齐全,安装后使用了一下,加热很快。超赞!	1
4	2928	总是Thinkpad系列的,系统的稳定性比较好,散热也做得不错,VISTA界面很漂亮。	1
5	16726	华语乐坛未来的巨星接班人,蒙牛酸酸乳巨星梦想学院也为他提供了赴韩国进行音乐游学机会。	1
6	4452	该要怎样做,是一本讲解设计之道而不是设计之艺术的书。翻译得也不错,启发很多,受益匪浅。对	1
7	4116	有些区别的可能是自己没有一个人住过,所以没共鸣,书没看完就放一边了,还好就买了一本她的书	0
8	5206	最喜欢邓佑真,一如她名字里的真字,她真是十分讨喜的主角。画面感很好,大概是因为要电视	2
9	15291	剧作。3屏幕有些偏色,图片显示也不够靓丽,由于是STN屏幕的原因,强光环境下屏幕显示不清。	2
10	7651	共了双床房但是印象大打折扣补充点评2008年3月24日温泉住宿客人是98池是46另外收费的	2
11	1193	则超过三天就要求用户取消订单重新订认为不人性化,应当在10天内用户在用户的要求下都可以才好	0
12	12045	外出携带,京东客服电话太难打,基本是摆设。触摸板那里温度高,包装箱被拆开过,不知何故。	0
13	16710	是背景很好罢了,选择在这样一个竞争激烈的时代,写了那么一个催人奋进的故事,文笔不咋地。	0
14	16011	价格不高,觉得是占了便宜,没想到去了之后,房间有很怪的味道,想换房间说没有了,四周都在	1
15	3154	蒙牛手工饼干搭配不错慢慢复习非家西点铺	1

图 1 数据样例

Fig. 1 Data samples

1.2 Word2Vec 词向量表示

Word2Vec 框架由 T. Mikolov 等^[9]于 2013 年提出,它通过神经网络模型将词映射到低维的实数向量空间上,从而得到词向量表示,并通过该向量空间上的词相似度来表示文本语义上的相似度. 该框架由两个神经网络语言模型组成:连续词袋(CBOW)模型和 Skip-gram 模型. 考虑到训练过程中的效率问题,两种模型均采用只包含输入层、隐藏层和输出层的浅层神经网络进行训练.

CBOW 模型通过已知上下文单词对当前单词出现的概率进行预测,其结构如图 2 所示. 该模型的输入为需要预测的词其周围词的 One-Hot 编码向量,继而通过前向传播加反向传播进行训练,最后得到词向量矩阵 $W_{V \times N}$. CBOW 模型在输入层到隐藏层的映射中进行权值共享以提高训练效率.

Skip-gram 模型通过中间单词预测上下文单词,其结构如图 3 所示. 该模型输入的是某个词的 One-Hot 编码后的向量,经由隐藏层线性组合到输出层,输出为该单词上下文的 One-Hot 编码,通过反向传播优化损失函数得到最后的词向量表示 $W_{V \times N}$.

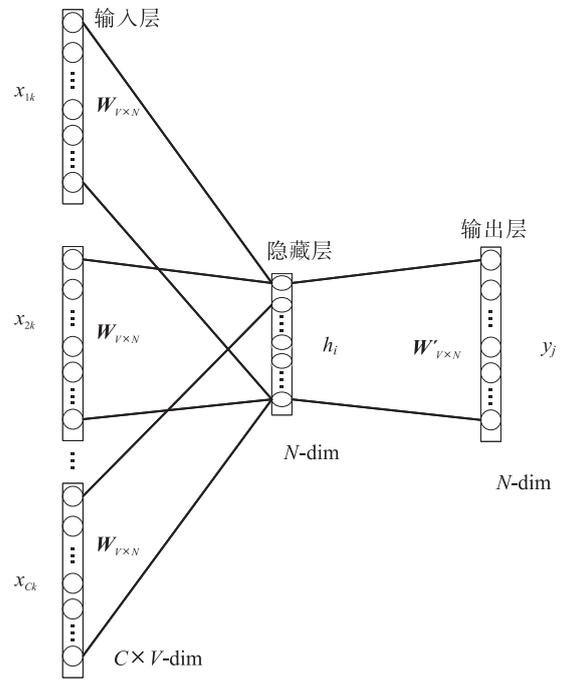


图 2 CBOW 模型结构

Fig. 2 CBOW model structure

2 网络结构设计

2.1 BERT 预训练语言模型

BERT 是由谷歌公司的 J. Devlin 等^[10]在 2018 年提出的一种 NLP 预训练语言模型. 该模型主要由双向 Transformer 编码器实现,其结构如图 4 所示,其中 E_1, \dots, E_N 表示字的文本输

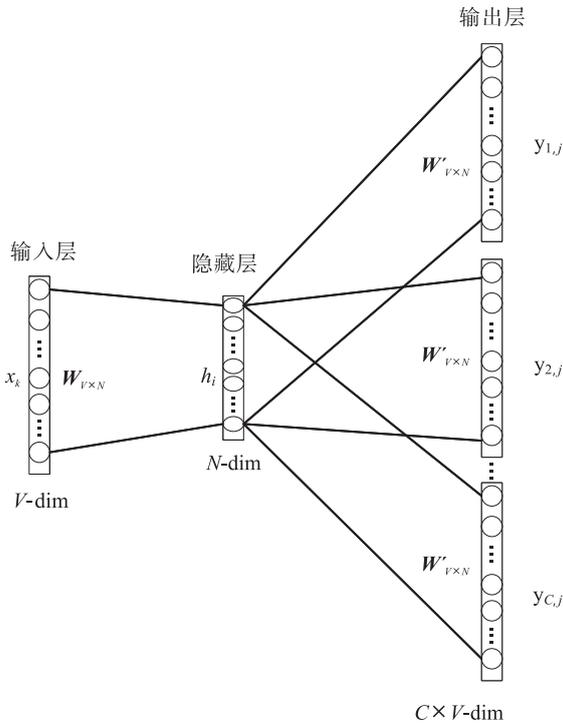


图 3 Skip-gram 模型结构

Fig. 3 Skip-gram model structure

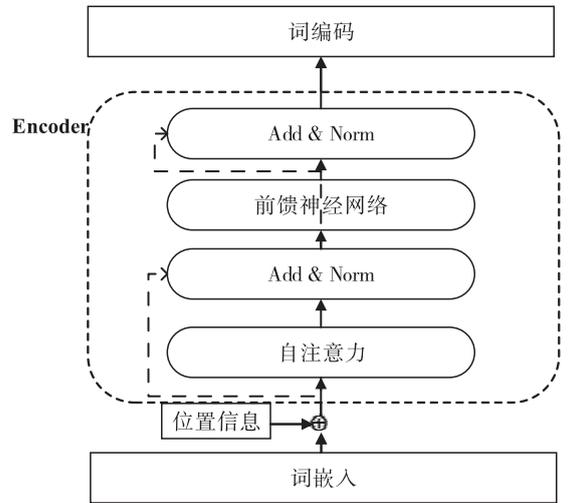


图 5 Transformer 中的 Encoder 结构

Fig. 5 Encoder structure in Transformer

个字的位置信息进行求和,再经自注意力层来帮助 Encoder 在对每个字进行编码时查看该字的前后信息,然后再经过 Add&Norm 层,其中,Add 表示将自注意力层的输入和输出结果进行相加, Norm 表示对相加后的输出结果进行归一化.得到的向量列表会传到一层全连接的前馈神经网络,在该网络内部,也会有相应的 Add&Norm 层处理,最后输出全新的归一化后的词向量列表.该向量列表能有效学习每个单词的前后信息,从而获得更好的词向量表示.

2.2 BiGRU 网络结构

J. L. Elman^[11]在 1990 年提出的 GRU (gate recurrent unit) 是一种循环神经网络 RNN (recurrent neural network),已被应用于自然语言处理领域.在循环神经网络结构中,对于每一个神经元来说,当前时刻的输出主要由当前时刻的输入和之前时刻的输出决定,因此,该类型的网络结构可以对有时序依赖的数据进行更好的表示学习.由于 RNN 在网络模型训练中进行反向传播时,梯度不能在较长序列中长期传递下去,因此便会出现梯度消失的情况.而 GRU 网络可用来解决长期记忆和反向传播中梯度消失等问题,其结构如图 6 所示.

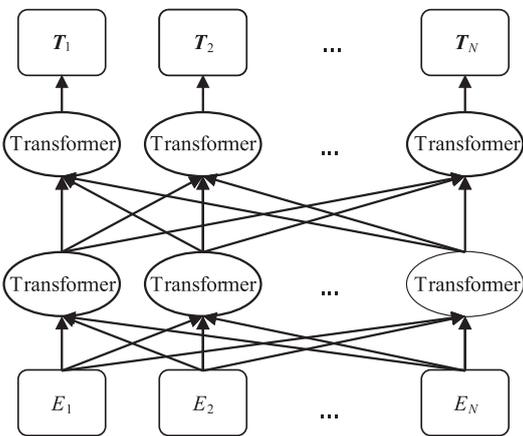


图 4 BERT 模型结构

Fig. 4 BERT model structure

入,经过双向 Transformer 编码器得到文本的向量化表示 T_1, \dots, T_N .

Transformer 模型是 BERT 的核心构成,也是一个基于自注意力机制的序列到序列模型,主要结构是 Transformer 中的编码 (Encoder) 部分,其结构如图 5 所示. Encoder 结构先将一段文本的字嵌入作为输入,之后与该段文本中每

由图6可以看出,GRU网络通过训练可以学习如何遗忘信息和记住信息,因此具有更好

的捕捉长距离依赖关系的能力.但是GRU网络只能处理单向的时间序列,而文本中的前后信息关联较大,单向处理会遗漏许多信息,为此提出了BiGRU网络,其结构如图7所示.

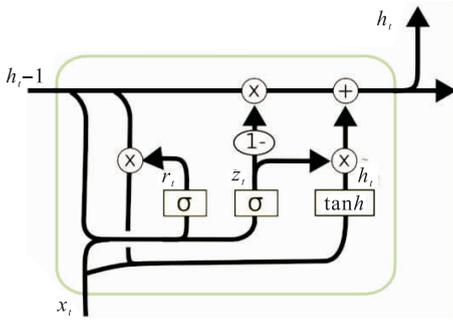


图6 GRU网络结构

Fig.6 GRU network structure

2.3 BERT + BiGRU 网络结构

综合考虑本机实验环境平台,选用由哈工大讯飞联合实验室发布的BERT-wwm-ext模型^[12],基于BERT,结合BiGRU网络设计的BERT + BiGRU网络结构如图8所示.由于BERT支持的最大输入token长度为512,因此该网络模型首先将长文本截成k段,分别输入

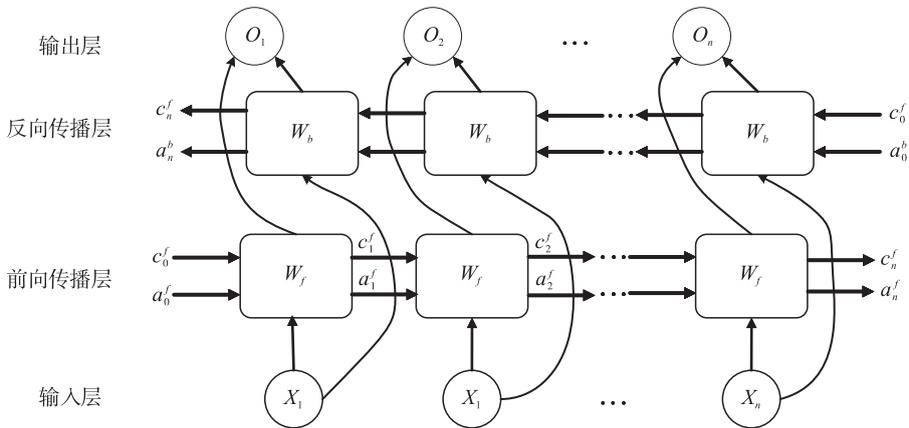


图7 BiGRU网络结构

Fig.7 BiGRU network structure

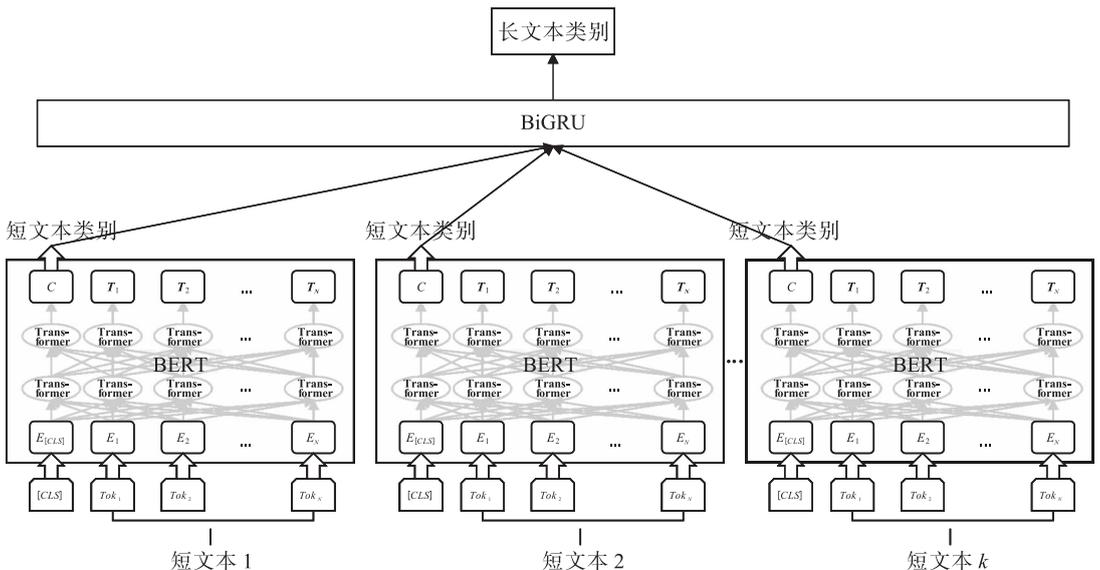


图8 BERT + BiGRU 网络结构

Fig.8 BERT + BiGRU network structure

BERT 预训练语言模型进行词向量的抽取,得到每段文本的包含上下文信息的词向量表示,这样既可以保证不损失文本信息,又可以通过调节输入文本长度来降低显存占用,从而更有效地利用资源;然后采用 BiGRU 模型对 k 个由 BERT 预训练语言模型得到的向量进行拼接,得到最终的特征向量;最后通过全连接层结合 softmax 进行情感分类。

3 实验结果与分析

3.1 评价指标

为了有效且全面地评估模型性能,实验采用的评价指标主要包括精确率 (*Precision*)、召回率 (*Recall*)、*F1* 和 *MicroF1*。精确率指分类器预测为正样本且预测正确的样本占全部预测为正样本的比例;召回率指分类器预测为正样本且预测正确的样本占所有真正样本的比例;*F1* 值指综合精确率和召回率的一个评估指标(精确率和召回率的调和平均值),无论针对类别平衡的数据还是类别不平衡的数据,该指标都能全面地评估模型;*MicroF1* 指各类别 *F1* 值的微平均,是综合评价多分类模型性能的指标。主要计算公式分别为

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = (2 \times Precision \times Recall) / (Precision + Recall)$$

其中,真正例 *TP* 表示被模型预测为正的样本个数;真负例 *TN* 表示被模型预测为负的样本个数;假正例 *FP* 表示被模型预测为正,但其实为负的样本个数;假负例 *FN* 表示被模型预测为负,但其实为正的样本个数。

3.2 结果与分析

实验平台使用 Pytorch1.2, CUDA 9.0 和具有 16 G 显存的英伟达 Tesla P100 GPU。采用数据集对 BERT 预训练语言模型进行参数微调。

微调的过程主要涉及输入文本的 max-seq-length、batch-size、迭代步数等。考虑到显存资源,在实验中单 GPU 的 batch 大小参数,即 per_gpu_train_batch_size 设置为 4, gradient_accumulation_steps 设置为 2, 学习率参数 learning_rate 设置为 $1e-5$;在训练过程中采用三角学习率,首先进行 warm_up,学习率逐渐变大,之后再继续进行 linear lr decay,学习率逐渐变小,这样可以有效地改善训练效果。

实验使用五折交叉验证方法对双向 LSTM 结合 Attention 机制模型 (W2V-BiLSTM-Attention)、传统卷积神经网络模型 (W2V-CNN)、传统循环神经网络模型 (W2V-RNN) 和本文提出的 BERT + BiGRU 网络模型进行训练和评估,其对比结果如表 1 所示。

表 1 BERT + BiGRU 模型与其他模型的对比结果

Table 1 Comparison of BERT + BiGRU model and other models

模型	类别	<i>Precision</i>	<i>Recall</i>	<i>F1</i>	<i>MicroF1</i>
W2V-CNN	0	0.814 7	0.898 9	0.854 7	
	1	0.804 1	0.880 5	0.840 6	0.82
	2	0.789 6	0.414 3	0.543 4	
W2V-RNN	0	0.858 9	0.891 0	0.874 6	
	1	0.857 1	0.855 3	0.856 2	0.83
	2	0.774 1	0.709 7	0.740 5	
W2V-BiLSTM-Attention	0	0.873 9	0.894 5	0.884 1	
	1	0.856 6	0.880 5	0.868 4	0.86
	2	0.790 7	0.692 7	0.738 5	
BERT + BiGRU	0	0.934 7	0.924 0	0.929 3	
	1	0.933 7	0.905 0	0.919 1	0.91
	2	0.787 4	0.867 6	0.825 5	

W2V-CNN 和 W2V-RNN 是神经网络产生以来较为传统的情感分类方法,由表 1 可知,W2V-RNN 的 *MicroF1* 优于 W2V-CNN,这是由于 W2V-RNN 的 RNN 网络具有结合上下文信息的特点,因此较 W2V-CNN 网络展现了更好的分类效果。

W2V-BiLSTM-Attention 网络是在传统 RNN 的基础上,基于双向 LSTM 和 Attention 模型的

网络结构设计的. 由表 1 可知, W2V-BiLSTM-Attention 的 *MicroF1* 显著优于 W2V-CNN 和 W2V-RNN, 这是因为 W2V-BiLSTM-Attention 不仅考虑句子中不同的维度信息, 而且通过最后的 Attention 层注意到文本句子对结果明显倾向的特征表示, 因此结合双向 LSTM 网络于 Attention 模型后, 可以更准确地针对评论文本进行情感分类; BERT-BiGRU 的 *MicroF1* 值最高, 这是因为采用 BERT 所获得的字或词向量是动态的, 同样的字或词在不同的语境中会有不同的字或词向量表达, 具有更精确的情感分类效果, 且采用 BiGRU 模型对多个 BERT 模型得到的向量进行拼接, 可以结合文本的上下文信息, 避免截断后的文本出现信息损失.

4 结语

本文基于 BERT 预训练语言模型, 结合 BiGRU 设计了 BERT + BiGRU 网络模型. 该模型首先通过 Word2Vec 进行词向量表示, 再通过 BERT-wwm-ext 得到词向量的动态表达, 最后将其输入 BiGRU 网络进行情感分类. 实验结果表明, 较 W2V-BiLSTM-Attention、W2V-CNN 和 W2V-RNN 3 种模型, 本文模型的分类效果最好. 但本文采用的 BERT 预训练语言模型只是 BERT 模型中的一个 base 版本, 后续将对其他预训练模型的使用场景和分类效果进行对比研究, 以进一步提高情感分类的准确性.

参考文献:

- [1] 赵妍研, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834.
- [2] LIU B, ZHANG L. A survey of opinion mining and sentiment analysis [M] // AGGARWAL C C, ZHAI C X. Mining text data. New York: Springer, 2012: 415 - 463.
- [3] 周咏梅, 杨佳能, 阳爱民. 面向文本情感分析的中文情感词典构建方法[J]. 山东大学学报(工学版), 2013, 43(6): 27.
- [4] PANG B, LEE L, VAITHYANATHAN S. Thumbs up sentiment classification using machine learning techniques [C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2002: 79.
- [5] 姜杰. 社交媒体文本情感分析[D]. 南京: 南京理工大学, 2017.
- [6] 王利利. 基于深度学习的中文文本情感分类研究及应用[D]. 徐州: 中国矿业大学, 2019.
- [7] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions [C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 151.
- [8] 图灵联邦. 情感分类大赛 dataset [EB/OL]. (2020 - 01 - 23) [2020 - 05 - 31] <https://www.turingtopia.com/competitionnew/detail/319f33ab29c04d9583e7f5c208dea119/dataset>.
- [9] MIKOLOV T, CHEN K, CORRADO G S, et al. Efficient estimation of word representations in vector space [C] // Proceedings of the 2013 International Conference on Learning Representations. [S. l. : s. n.], 2013.
- [10] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] // Proceedings of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 4171.
- [11] ELMAN J L. Finding structure in time [J]. Cognitive Science, 1990, 14(2): 179.
- [12] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT [J]. (2019 - 10 - 29) [2020 - 05 - 31] <https://arxiv.org/pdf/1906.08101.pdf>.