文章编号：2095-2163(2021)12-0193-04 中图分类号：TP391 文献标志码：A

基于 python 的社交情感分析注意力模型

薛 涛
（运城师范高等专科学校 数计系，山西 运城 044000）

摘 要：针对现有方法没有充分探索表情符号对文本情感极性的影响问题，设计了一种社交情感分析注意力模型，并基于 python 实现了该模型。首先构建一个富含表情符号的语料库，利用注意力方法描述表情符号对文本的影响，设计并实现了基于双向长短期记忆的情感分析模型。实验采用社交平台数据集，将本模型与现有模型进行全面对比。结果显示，本模型具有较高的准确度，能实现较好的社交情感分析性能。

关键词：情感分析；注意力模型；Python

Social sentiment analysis model based on python

XUE Tao
（Department of Mathematical Calculation, Yuncheng Advanced Normal College, Yuncheng Shanxi 044000, China）

【Abstract】Aiming at the problem that existing methods have not fully explored the influence of emoji on the emotional polarity of text, this paper designs an attention model for social sentiment analysis implemented based on python. This paper first constructs a dataset composed of rich emoticons. Then the attention method is used to describe the influence of emoji on the text with a sentiment analysis model based on two-way long and short-term memory. In the experimental part, we use the dataset of the social platform to comprehensively compare the model in this article with other existing models. The results show that the model in this paper has high accuracy and can achieve better social sentiment analysis performance.

【Key words】sentiment analysis; attention model; python

0 引 言

人们通过社交平台来表达感受、情绪和态度，社交平台的帖子中通常包含丰富的信息，因此社交媒体成为热门研究对象。其中，情感分析是最基本且关键的研究主题之一[1-3]。情感分析的目的是分析社交媒体的极性，以判断人们对某些事件所持有的正面、负面或中性态度[4-5]。有研究者提出将社交媒体中的表情符号应用于情感极性预测，目前大多数现有的方法不仅依赖于手工特征，还分别考虑了表情符号和纯文本的情感，但并没有充分探索表情符号对文本情感极性的影响。表情符号在纯文本的情感极性中起着重要作用，对于情感原本是中性的纯文本，在纯文本后添加开心或沮丧的表情会使帖子表达不同的情绪极性。

本研究提出了一种深度学习模型，结合表情符号对文本情感极性的影响以进行情感分析。该模型使用双向长短期记忆模型来构建社交平台帖子的表示，使用注意力模型计算每个单词的权重。研究的主要贡献有两点：首先建立了带有表情符号、包含超过 1 万条帖子的语料库；其次，联合训练微博帖子中的表情符号和单词，获得包含其上下文信息的表情符号表示。

1 符号语料库

大多数现有的情感分析语料库仅包含一小部分带有表情符号的内容，这些语料库并不适用于基于表情符号的情感分析。因此，需要收集和注释带有表情符号的文本。

由新浪微博收集了 250 000 条微博帖子，从中提取了 85 000 条包含表情符号的帖子。根据每个表情符号的出现次数，对微博帖子进行排名，并选择至少出现 10 次的表情符号集。用表情符号分割每条微博帖子，选择只包含一个表情符号的微博帖子，并过滤掉帖子中的 URL、用户名和主题标签以清理数据，并选择至少出现 10 次的表情符号集。用表情符号分割每条微博帖子，并保留长度大于 5 的微博帖子。在筛选出的 35 000 条微博帖子中，随机抽取了 18 000 条微博帖子进行下一步标记，并使用 Jieba 中文文本分词工具进行分词。

基金项目：全国高等院校计算机基础教育研究会计算机基础教育教学研究项目(2018-AFCEC-378)。

作者简介：薛 涛(1979-)，男，硕士，讲师，主要研究方向：计算机科学与技术基础教育、程序设计、数据库开发等。

收稿日期：2021-11-02

采用手工标注的方式来构建语料库。情感极性分为正面、中性和负面,分别用 0、1、2 表示。首先,仅根据文本来判断每个帖子的极性,即从文本中删除表情符号,仅使用每条微博帖子的纯文本来确定帖子的极性;然后,结合文本和表情符号来确定每个帖子的极性。语料库的极性结果见表 1。

表 1 语料库的极性

Tab. 1 Corpus polarity %

	正面	中性	负面
文本极性	38	36	26
总体极性	58	9	33

由此可见,表情符号的出现会改变帖子的情感极性。表 2 展示了情感极性变化的社交帖子情况。

表 2 情感极性变化的情况

Tab. 2 Changes of emotional polarity

	正→中	正→负	中→正	中→负	负→正	负→中
情感极性变化百分比	4%	2%	54%	28%	2%	10%

2 情感分析模型

本文提出的社交情感分析注意力模型结构如图 1 所示。

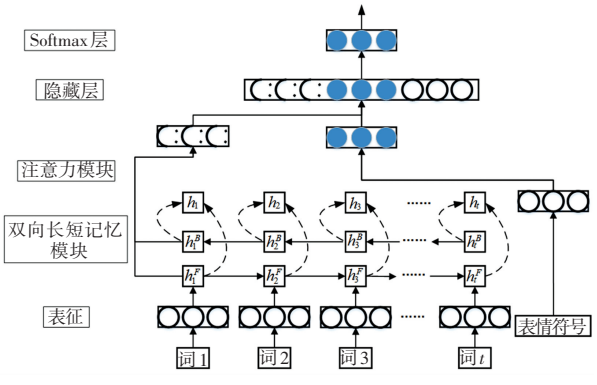

图 1 情感分析模型

Fig. 1 Sentiment analysis model

应用双向长短记忆(Bi-directional Long Short Term Memory,Bi-LSTM)模型学习句子的表征,将表征作为特征对情感的极性进行分类。本文使用 PyTorach 来实现该模型,PyTorach 是一个基于 Python 的深度学习框架。模型初始化过程如下:

```
def _init_(self, ntoken, ems, n, nclass, semb, hid_size=50,no_att=False):
    super(EmSentClass, self)._init_() //调用父类 EmSentClass
    self. ems = ems
    self.n = n
    self.embed = nn.Emb(ntoken, semb, padding_idx=0) //创建词嵌入模型
    self.no_att = no_att
    self.ems = nn.Emb(ems, semb)
    self.items = nn.Emb(n, semb)
    self.word = AttentionalBiGRU(semb, semb//2, no_att = self.no_att) //注意力模型
    self.semb = semb
    self.lin_out = nn.Linear(semb * 3,nclass) //设置网络全连接层
```

LSTM 能捕获序列中的长距离依赖关系。一个 LSTM 模型由多个 LSTM 单元组成,其中每个 LSTM 单元对神经网络中的记忆进行建模。LSTM 单元包含的门结构允许 LSTM 存储和访问随时间变化的信息。给定一个包含词 w_t 的短文本,使用嵌入矩阵 $\boldsymbol{W}_e$ 将这些词嵌入到向量 $\boldsymbol{x}_t = \boldsymbol{W}_e w_t$ 中,该向量的维度是 d。Bi-LSTM 包含一个前向 LSTM 以读取从 $\boldsymbol{x}_1$ 到 $\boldsymbol{x}_T$ 的文本和一个后向 LSTM 以读取从 $\boldsymbol{x}_T$ 到 $\boldsymbol{x}_1$ 的文本,即:

$$\begin{aligned} \boldsymbol{h}_t^F &= L^F(x_t) \\ \boldsymbol{h}_t^B &= L^B(x_t) \end{aligned} \tag{1}$$

Bi-LSTM 将每个词 w_t 映射到一对隐藏向量 $\boldsymbol{h}_t^F$ 和 $\boldsymbol{h}_t^B$ 中,那么一个词可以表示为一对向量的串联,即 $\boldsymbol{h}_t = [\boldsymbol{h}_t^F, \boldsymbol{h}_t^B]$。因此,得到 $[h_0, \cdots, h_T]$,然后将其输入到平均池化层以获得句子的表示 s。

为了表明表情符号对文本情感极性的影响,提出了一种基于表情符号的注意力机制。给定一个微博帖子,每个词对情感极性的贡献是不一样的,表情符号的交互权重也不均等。EA 机制结合单词和表情符号来衡量微博帖子中单词的权重。

在微博帖子 $\{w_1, \ldots, w_T; \boldsymbol{E}\}$ 中,w_i 表示单词,$\boldsymbol{E}$ 表示表情符号。首先,w_i 和 $\boldsymbol{E}$ 都被转换为向量表示,即 $\boldsymbol{x}_i$ 和 $\boldsymbol{e}$。

聚合这些词的表示以形成句子表示,句子表示 s 是隐藏状态 h_i 的加权和,即:

$$s = \sum_{i=1}^{T} a_i h_i \tag{2}$$

其中,权重 a_i 用于衡量第 i 个词的重要性,其计算方式为:

$$a_i = \frac{e^{f(h_i, e)}}{\sum_{j=1}^{T} e^{f(h_j, e)}} \tag{3}$$

其中,函数 $f(\cdot)$ 表示单词的重要性,函数 $f(\cdot)$

的定义为：

$$f(h_i, e) = v^{\mathrm{T}} \tanh(W_h h_i + W_E e + b) \quad (4)$$

其中，W_h、W_E 是可学习的参数；v^{T} 表示 v 的转置；b 是偏置。

串联了 3 种类型的特征，如下所示：

$$l_c = [h_0^F, h_T^B] \oplus s \oplus e \quad (5)$$

其中，h_0^F 和 h_T^B 表示最后一步中前向和后向 LSTM 的隐藏状态。

训练的目标是最小化交叉熵损失，在引入基于表情符号的注意力机制后，获得了用于文本情感分析的特征 l_c。模型使用线性变换，将 l_c 投影到 C 种类别的目标空间中：

$$d_c = W_c l_c + b_c \quad (6)$$

之后，使用一个 softmax 层来获得微博帖子情感的概率分布：

$$p_c = \frac{e^{d_c}}{\sum_{k=1}^{C} e^{d_k}} \quad (7)$$

其中，C 是情感标签的数量，p_c 是情感标签 c 的预测概率。

softmax 层的 python 实现如下所示：

```
def softmax(self,mat,mak):
exp = torch.exp(mat) * Variable(mak,requires_grad=False)
sum_exp = exp.sum(1,True)+0.000 001
sm = exp/sum_exp.expand_as(exp)//定义 softmax 函数
torch.set_printoptions(threshold=19 000) //设置输出选项
d = { }
bs = len(self.rev)
global w_dict
for i in range(bs):
r =self.rev[i].split()
att = softmax[i]
att_ = att.data.cpu().numpy() //把 tensor 转换成 numpy 的格式
comb = tuple(zip(r,att_))
if len(att) ! = 1:
w_dict[self.rev[i]] = comb
if bs ! = 1:
w_dict = {}
if bs == 1:
try:
if list(w_dict.values())[0].__len__() < 2:
w_dict = {}
except:
w_dict = {}
return sm
```

设 $p_c^g(d)$ 是帖子的目标分布，$p_c(d)$ 是预测的情绪分布，D 是微博帖子的集合。训练目标是最小化集合 D 中的 $p_c^g(d)$ 和 $p_c(d)$ 之间的交叉熵损失，则损失函数定义为：

$$L = -\sum_{d \in D} \sum_{c=1}^{C} p_c^g(d) \log(p_c(d)) \quad (8)$$

3　实验评估

为了获得单词和表情符号的嵌入表示，使用 word2vec.3 的 SkipGram 模式，对由 350 万条微博组成的大规模语料库上训练单词和表情符号嵌入。

实验中使用 5 重交叉验证。原始数据被随机分成 5 个相等的部分，其中 4 个部分用于训练，第 5 部分用于测试。从 4 个训练部分中随机选择一个部分作为开发集来调整超参数。分类结果通过准确度来衡量。准确度定义为 T/N，其中 T 表示预测的与真实情绪评级相同的情绪评级数量，N 表示微博的总数量。由于多分类中类不平衡问题，还使用了宏观精度来进行更公平的比较。

将词嵌入和表情符号嵌入的维度设置为 200。LSTM 单元中隐藏状态和单元状态的维度设置为 100。在训练期间，使用 Adadelta 作为优化方法。训练的批次大小为 16，动量为 0.9，初始学习率 α 为 0.01。

为了评估本模型的性能，将其与 E-only[6]、SVM、LSTM 和 Bi-LSTM 等算法进行了比较。其中，E-only 是仅使用表情符号来判断情感的极性，Bi-LSTM 将微博帖子的文本和表情符号作为 Bi-LSTM 模型的输入进行情感分析，实验对比了各个模型的精度、召回率、F-度量和准确度，表 3 给出了所有模型进行情感分析的实验结果。由于类不平衡问题，算法在中性极性的性能要远低于其它极性。

实验结果从表 3 的结果可见，由于模型利用了包括文本、表情符号特征，以及表情符号对文本的影响，本文模型表现最佳。这表明基于表情符号的注意力，可以有效地捕捉表情符号对文本情感极性的影响。此外，LSTM 优于 SVM，表明与具有稀疏指标特征的离散模型相比，神经网络模型能更好地提取文本和表情符号特征。

表 3 实验结果对比

Tab. 3 Comparison of experimental results

模型	极性	精度	召回率	F - 度量	准确度
E-only	正	0.88	0.93	0.91	
	中	0.38	0.18	0.21	
	负	0.88	0.90	0.89	0.86
SVM	正	0.82	0.84	0.83	
	中	0.37	0.23	0.38	
	负	0.80	0.83	0.81	0.62
LSTM	正	0.89	0.94	0.90	
	中	0.39	0.19	0.22	
	负	0.88	0.91	0.90	0.86
Bi-LSTM	正	0.87	0.95	0.91	
	中	0.43	0.16	0.22	
	负	0.91	0.91	0.90	0.87
本文模型	正	0.89	0.95	0.92	
	中	0.47	0.27	0.34	
	负	0.92	0.92	0.92	0.88

Bi-LSTM 模型与本文模型对不同情感极性的准确率比较,结果见表 4。从中可以看出,在情感变化方面,本文模型在大多数情况下优于 Bi-LSTM 模型。

表 4 极性变化的准确度对比

Tab. 4 Accuracy comparison of polarity changes

	正—>中	正—>负	中—>正	中—>负	负—>正	负—>中
Bi-LSTM	0.11	0.85	0.96	0.95	0.86	0.11
本文模型	0.17	0.88	0.98	0.97	0.89	0.18

4 结束语

本研究设计并实现了基于注意力模型的情感分析模型。该模型考虑了表情符号对文本情感极性的影响。与现有的模型相比,本模型实现了较好的性能。未来的工作将在以下两个方向上进一步研究表情符号对短文本情感极性的影响。首先,将研究扩展到其它类型的短文本。其次,将采用其它神经网络模型以探索表情符号对文本的影响。

参考文献

[1] LIU S, SHEN H, ZHENG H, et al. Ct lis: Learning influences and susceptibilities through temporal behaviors [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(6): 1-21.

[2] ZHAO S, XIAO Y, GUO J, et al. Curriculum cyclegan for textual sentiment domain adaptation with multiple sources [C]//Proceedings of the Web Conference 2021. 2021: 541-552.

[3] CHEN Z, CAO Y, YAO H, et al. Emoji-powered sentiment and emotion detection from software developers' communication data [J]. ACM Transactions on Software Engineering and Methodology (TOSEM), 2021, 30(2): 1-48.

[4] HAYATI S A, MUIS A O. Analyzing incorporation of emotion in emoji prediction [C]// Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2019: 91-99.

[5] DUARTE L, MACEDO L, OLIVEIRA H G. Exploring emojis for emotion recognition in portuguese text[C]//EPIA Conference on Artificial Intelligence. Springer, Cham, 2019: 719-730.

[6] LE T A, MOELJADI D, MIURA Y, et al. Sentiment analysis for low resource languages: A study on informal Indonesian tweets [C]//Proceedings of the 12th Workshop on Asian Language Resources (ALR12). 2016: 123-131.

(上接第 192 页)

[4] 袁华波. 基于 Retinex 算子的盲解卷积方法研究[D]. 西安:西安电子科技大学,2017.

[5] 程序员大本营. 自然图像先验与图像复原[EB/OL]. [2018-07-28]. https://www.pianshen.com/article/3941672358/

[6] 张姣. 混合正则化约束的湍流退化图像复原算法[J]. 激光与红外,2017 (7):884-888.

[7] 林子强. 明场显微光切片的三维重建技术[D]. 广州:暨南大学,2018.

[8] 孙必慎. 计算视觉核心问题:自然图像先验建模研究综述[J]. 智能系统学报,2019 (1):71-81.