张华辉,吴敏敏,徐航. 基于交互对抗网络的方面级情感分类方法[J]. 智能计算机与应用,2026,16(1):150-156. DOI:10.20169/j.issn.2095-2163.24032802

基于交互对抗网络的方面级情感分类方法

张华辉, 吴敏敏, 徐 航
(莆田学院 计算机与大数据学院, 福建 莆田 351100)

摘 要: 情感分类是自然语言处理领域的热点研究问题之一,方面级的文本情感分类旨在识别文本不同方面间的情感极性。针对方面级情感分类模型存在特征提取能力弱、方面词与上下文间交互不充分的问题,提出基于交互对抗网络的方面级情感分类模型(Aspect-level Sentiment classification model based on Interactive Adversarial Networks, ASIAN)。首先,通过 Transformer 的双向表征编码器模型作为编码器,将方面词和上下文进行单独建模提取隐含层特征。其次,构建交互注意力网络,将隐含层特征进行交互学习。最后,对交互信息进行联合学习,做交叉熵损失、回传参数。此外,ASIAN 添加了对抗训练旨在进一步优化分类效果。在 SemEval-2014 任务 4 中的 Laptop、Restaurant 数据集和 ACL-2014 的 Twitter 数据集上,ASIAN 与大多数基线模型相比有较高的分类准确率。

关键词: 情感分类; 自然语言处理; 交互对抗网络; 对抗训练
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)01-0150-07

Aspect-level sentiment classification method based on Interactive Adversarial Networks

ZHANG Huahui, WU Minmin, XU Hang
(College of Computer and Data Science, Putian University, Putian 351100, Fujian, China)

Abstract: Sentiment classification is one of the hot research problems in the field of natural language processing, aspect-level text sentiment classification aims to identify the sentiment polarity between different aspects of a text. To address the problems of weak feature extraction and insufficient interaction between aspect words and context in aspect-level sentiment classification models, Aspect-level Sentiment classification model based on Interactive Adversarial Networks (ASIAN) is proposed. Firstly, the bidirectional encoder representation from Transformers model is employed as an encoder, the aspect words and the context are modeled separately to extract the hidden layer features. Secondly, an interactive attention network is constructed, which interacts the hidden layer features. Finally, the interaction information is learned jointly to do cross-entropy loss and back-propagation parameters. In addition, ASIAN adds an adversarial training aimed at further optimizing classification results. On the Laptop and Restaurant datasets from SemEval-2014 Task 4 and the Twitter dataset from ACL-2014, ASIAN has high classification accuracy compared to most baseline models.

Key words: sentiment classification; natural language processing; Interactive Adversarial Network; adversarial training

0 引 言

在文本情感分类任务中,依据不同的文本粒度可以将情感分类问题分为篇章级、句子级、方面级[1]。其中,方面级情感分类问题是自然语言处理领域的一个热点研究方向,方面级情感分类是旨在识别文本中不同方面的情感,如评论性文本:“这家饭店的牛肉很好吃,环境很舒适,但是价格很贵”中,有“牛肉”、“环境”、“价格”三个方面词,对应的情感分别为积极、积极、消极。在海量的社交网络数据中,高效地分类出文本不同方面的情感具有一定的挑战性。

早期,基于传统机器学习方法的文本情感分类模型在文本研究领域占据一定地位,如黄佳锋等学者[2]采用集成学习进行中文情感分类,通过定义 3 种复杂的特征提取模式,构造基于信息增益的随机

基金项目: 国家自然科学基金(62103209); 福建省自然科学基金项目(2020J05213); 福建省中青年教师教育科研项目(JAT220298)。
作者简介: 张华辉(1995—),男,助教,主要研究方向:自然语言处理,情感分析。Email:zhanghuahui@ptu.edu.cn。
收稿日期: 2024-03-28

子空间的集成算法做分类器。王伟等学者[3]提出一种基于 LDA(Latent Dirichlet Allocation)主题模型的文本情感分类方法，该方法通过情感词典提取情感信息，使用主题模型发掘情感信息中的关键特征，并融入到情感向量空间中，最后利用支持向量机实现中文评论文本的情感分类。类似地，还有很多其他基于传统机器学习方法的文本情感分类模型[4-6]，但是，上述基于传统机器学习方法的文本情感分类模型都存在特征工程过程复杂、特征的选取非常依赖于经验、训练出来的分类器往往分类精度低以及泛化能力弱等问题。

近年来，深度网络模型的出现，很好地规避了繁杂的特征提取过程，而且深度网络模型的文本情感分类效果有大幅提升[7-9]。如 Zeng 等学者[10]提出基于 BERT 的局部注意力机制的情感分类模型，以方面词为中心进行文本权重的动态衰减或遮盖，该模型的思想简单、效果不错。张立等学者[11]提出一种基于多方面依存建模的图卷积网络模型，有效建模了文本中多方面词的依存关系，在多个公开数据上取得了不错的效果。

基于深度学习的情感分类模型，不仅在情感分类精度上对比传统机器学习方法分类模型有大幅提升，而且减少了不必要的人工提取特征过程。针对现有的文本情感分类模型存在特征提取能力弱、方面词与上下文交互不充分的问题，本文提出基于交互对抗网络的方面级情感分类模型，主要工作和贡献如下：

(1)提出基于 BERT 的交互注意力网络模型，通过 BERT 编码提取语义特征，建立交互注意力网络将方面词与上下文充分交互。通过实验验证了该网络优于大多数基线模型。

(2)模型引入了对抗样本，添加了对抗训练过程，旨在进一步增强优化模型的分类效果，实验表明对抗训练对不同数据集上的分类精度有不同幅度的提升。

(3)在公开的 Laptop、Restaurant 和 Twitter 数据集取得了不错的效果，最终的分类准确率分别达 77.12%、82.59%、73.27%。

1　对抗训练

对抗样本是指在原数据集中通过人工添加肉眼不可见或在经处理不影响整体的肉眼可见的细微扰动所形成的样本[12]，这类样本会导致训练好的模型以高置信度输出与原样本不同的分类结果。

对抗样本的概念首次由 Szegedy 等学者[13]提出，经 Goodfellow 等学者[14]证明了高维网络实际呈现为线性状态，为对抗训练奠定了理论基础。对抗训练是一种常见的对抗样本攻击的防御方法[15]，根据不同的对抗训练损失可以分为监督学习和无监督学习。

监督学习中，对抗样本的标签是原训练样本的标签，对抗训练的模型损失为：

$$-\log p(y \mid x + r_a;\ \boldsymbol{\Theta}) \quad \text{where } r_a = \arg\min_{r,\ \|r\| \leqslant \varepsilon} \log p(y \mid x + r;\hat{\boldsymbol{\Theta}}) \tag{1}$$

其中，$\boldsymbol{\Theta}$ 表示模型的参数；$\hat{\boldsymbol{\Theta}}$ 表示 $\boldsymbol{\Theta}$ 的常量参数；r_a 表示扰动；ε 表示扰动的阈值参数；x 表示正常样本；y 表示 x 对应标签；$x + r_a$ 表示对抗样本。

无监督学习中，对抗训练的模型损失为：

$$\mathrm{KL}[p(\cdot \mid x;\hat{\boldsymbol{\Theta}}) \,\|\, p(\cdot \mid x + r_a;\boldsymbol{\Theta})] \quad \text{where } r_a = \arg\max_{r,\ \|r\| \leqslant \varepsilon} \mathrm{KL}[p(\cdot \mid x;\hat{\boldsymbol{\Theta}}) \,\|\, p(\cdot \mid x + r;\hat{\boldsymbol{\Theta}})] \tag{2}$$

其中，KL[· ‖ ·]表示 KL 散度。

目前，比较有效的对抗训练算法有 FGM 算法[16]和 PGD 算法[17]。其中，FGM 算法是一种基于当前输入梯度 *L2* 范数缩放的一次性计算扰动算法，计算扰动公式为：

$$r_a = \varepsilon \cdot g / \|g\|_2, \quad \text{where } g = \nabla_x \log p(y \mid x;\boldsymbol{\Theta}) \tag{3}$$

其中，g 表示梯度，$\|\cdot\|_2$ 表示 *L2* 范数。

PGD 算法基于某个步长的多次迭代优化扰动算法，其生成的对抗样本公式为：

$$x^{t+1} = \prod_{x+R} \{x^t + \eta \cdot \mathrm{sign}[\nabla_x L(\boldsymbol{\Theta},x,y)]\} \tag{4}$$

其中，R 表示扰动集合；η 表示步长；L 表示损失函数；Π_{x+R} 表示以某个扰动阈值为半径的球上投影，若迭代扰动幅度过大会投影回球面。PGD 最优化内部损失和外部损失，公式如下：

$$\min_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}) \tag{5}$$

where $p(\boldsymbol{\Theta}) = E_{(x,y)\sim D}[\max_{r_a \in R} L(\boldsymbol{\Theta},x + r_a,y)]$，其中 D 表示一种数据分布。

2　ASIAN 模型

本文提出 ASIAN 模型进行方面级情感分类，ASIAN 的结构见图 1 所示，依次分别为输入层、编码层、注意力层和分类层。图 1 中，输入层负责接收将原始数据集预处理后的固定格式输入序列；编码层

负责将输入序列进行字符转换、编码和生成对抗样本;注意力层将编码后的样本及对抗样本依次通过交互注意力网络提取特征;分类层将提取的特性进行三分类,同时,回传损失、梯度下降,优化分类权重。

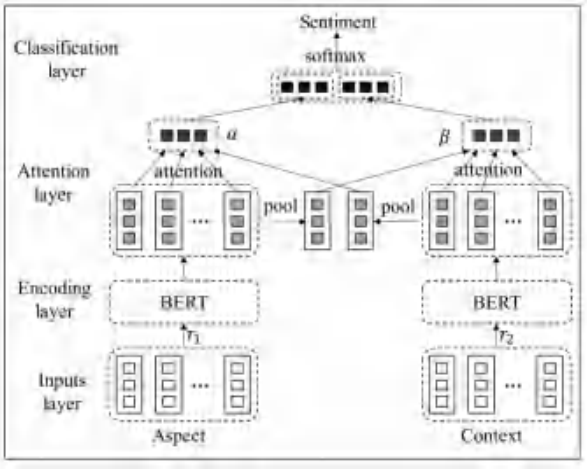

图 1 ASIAN 模型

Fig. 1 Model of ASIAN

2.1 输入层

将方面词和上下文分别预处理成“[CLS]+Aspect+[SEP]”和“[CLS]+Context+[SEP]”格式的固定输入。如存在方面词“atmosphere”和对应的评论性文本“The design and atmosphere is just as good”时,则预处理成“[CLS] atmosphere [SEP]”和“[CLS] The design and atmosphere is just as good [SEP]”。其中,[CLS]是开始标识符,[SEP]是分隔标识符。

2.2 编码层

将预处理好的固定格式输入进行编码,编码过程如图 2 所示。BERT 编码过程主要分为 2 个方面。一方面提取关于 Aspect 和 Context 的相关标识生成词嵌入、段落、单词位置信息,另一方面利用多头注意力机制和前馈网络提取特征。

(1)将方面词和上下文通过 BERT 词典映射替换每一个单词,生成新的 Embedding Information,根据[CLS]和[SEP]标识信息生成判断每个单词是否在同一段的 Segment Information 和生成每个单词在段中位置信息 Position Information。最后,将 Embedding Information、Segment Information 和 Position Information 叠加,再经过归一化和随机失活操作后,作为下一层次的输入。为方便叙述,假设方面词和上下文经过以上操作的结果用字母 X_1 和 X_2 表示。

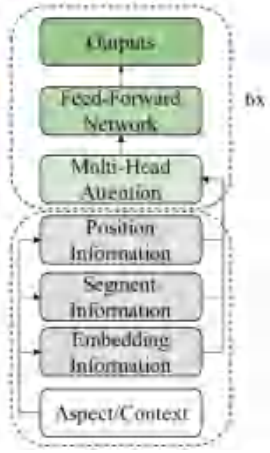

图 2 BERT 编码过程

Fig. 2 BERT encoding process

(2)当将 X_1 经过多头注意力机制(Multi-head Attention, MA)提取特征时,设多头注意力机制头数为 8,随机初始化可学习矩阵 $\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$,$i \in [1, 8]$。则有:

$$\mathrm{MA}_i = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}_i \tag{6}$$

其中,$\boldsymbol{Q}_i = \boldsymbol{X}_1 \cdot \boldsymbol{W}_i^Q$、$\boldsymbol{K}_i = \boldsymbol{X}_1 \cdot \boldsymbol{W}_i^K$、$\boldsymbol{V}_i = \boldsymbol{X}_1 \cdot \boldsymbol{W}_i^V$;$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示文本向量化后的查询、键和值矩阵;$\sqrt{d_K}$ 表示 K 的维度。将 8 种不同方向的注意力 MA_i 进行拼接并附加一个激活函数到输出 O_1:

$$\boldsymbol{O}_1 = \mathrm{Tanh}[(\mathrm{MH}_1 \oplus \mathrm{MH}_2 \cdots \mathrm{MH}_8) \cdot \boldsymbol{W}^*] \tag{7}$$

其中,Tanh 表示激活函数;“$\oplus$”表示拼接;$\boldsymbol{W}^*$ 表示可学习矩阵。再将 $\boldsymbol{O}_1$ 通过 Feed-Forward Network 得到 FFN($\boldsymbol{O}_1$),公式如下:

$$\mathrm{FFN}(\boldsymbol{O}_1) = \max[0, \mathrm{ffn}(\boldsymbol{O}_1)]\boldsymbol{W}_2 + \mathrm{bia}_2 \tag{8}$$

$$\mathrm{ffn}(\boldsymbol{O}_1) = (\boldsymbol{O}_1 \cdot \boldsymbol{W}_1 + \mathrm{bia}_1) \tag{9}$$

其中,bia_1 和 bia_2 表示偏置,$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 表示随机初始化的可学习矩阵。FFN($\boldsymbol{O}_1$) 通过归一化和随机失活操作后,此刻的结果作为下一同结构的输入放入,如此重复 6 次后的结果为编码层最终的输出。为方便表示,用 BERT( ) 函数表示整体编码层编码过程,可以表示为:

$$\boldsymbol{E}_{\mathrm{a}} = \mathrm{BERT}(X_1) \tag{10}$$

$$\boldsymbol{E}_{\mathrm{c}} = \mathrm{BERT}(X_2) \tag{11}$$

其中,$\boldsymbol{E}_{\mathrm{a}}$ 和 $\boldsymbol{E}_{\mathrm{c}}$ 分别表示 Aspect 和 Context 编码后的结果。

2.3 注意力层

首先,提取 Aspect 和 Context 编码后的平均池化向量 $\bar{\boldsymbol{E}}_{\mathrm{a}}$ 和 $\bar{\boldsymbol{E}}_{\mathrm{c}}$,公式如下:

$$\bar{\boldsymbol{E}}_{\mathrm{a}}=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{E}_{\mathrm{a}}^{i} \tag{12}$$

$$\bar{\boldsymbol{E}}_{\mathrm{c}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}_{\mathrm{c}}^{i} \tag{13}$$

其中，m 和 n 分别表示 Aspect 和 Context 的长度。

然后，将 $\bar{\boldsymbol{E}}_{\mathrm{a}}$ 和 $\boldsymbol{E}_{\mathrm{c}}$、$\boldsymbol{E}_{\mathrm{a}}$ 和 $\bar{\boldsymbol{E}}_{\mathrm{c}}$ 交互学习特征，得到特征矩阵 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$，计算过程如下：

$$\boldsymbol{\alpha}_i=\frac{\exp[\gamma(\boldsymbol{E}_{\mathrm{a}}^{i},\bar{\boldsymbol{E}}_{\mathrm{c}})]}{\sum_{j=1}^{m}\exp[\gamma(\boldsymbol{E}_{\mathrm{a}}^{j},\bar{\boldsymbol{E}}_{\mathrm{c}})]} \tag{14}$$

$$\boldsymbol{\beta}_i=\frac{\exp[\gamma(\boldsymbol{E}_{\mathrm{c}}^{i},\bar{\boldsymbol{E}}_{\mathrm{a}})]}{\sum_{j=1}^{n}\exp[\gamma(\boldsymbol{E}_{\mathrm{c}}^{j},\bar{\boldsymbol{E}}_{\mathrm{a}})]} \tag{15}$$

其中，$\gamma(\,)$ 函数的展开如下：

$$\gamma(\boldsymbol{E}_{\mathrm{a}}^{i},\bar{\boldsymbol{E}}_{\mathrm{c}})=\tanh(\boldsymbol{E}_{\mathrm{a}}^{i}\cdot\boldsymbol{W}_{\mathrm{a}}\cdot\bar{\boldsymbol{E}}^{\mathrm{T}}_{\mathrm{c}}+\mathrm{bia}_{\mathrm{a}}) \tag{16}$$

其中，$\boldsymbol{W}_{\mathrm{a}}$ 和 $\mathrm{bia}_{\mathrm{a}}$ 分别表示权重矩阵和偏置。$\gamma(\boldsymbol{E}_{\mathrm{a}}^{j},\bar{\boldsymbol{E}}_{\mathrm{c}})$、$\gamma(\boldsymbol{E}_{\mathrm{c}}^{i},\bar{\boldsymbol{E}}_{\mathrm{a}})$、$\gamma(\boldsymbol{E}_{\mathrm{c}}^{j},\bar{\boldsymbol{E}}_{\mathrm{a}})$ 的计算过程与上式一致，不详尽罗列。

最后，将注意力矩阵 $\boldsymbol{\alpha}$、$\boldsymbol{\beta}$ 和 $\boldsymbol{E}_{\mathrm{a}}$、$\boldsymbol{E}_{\mathrm{c}}$ 重构得到最终的交互注意力矩阵 $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$，公式如下：

$$\boldsymbol{E}_1=\sum_{i=1}^{m}\alpha_i\cdot\boldsymbol{E}_{\mathrm{a}}^{i} \tag{17}$$

$$\boldsymbol{E}_2=\sum_{i=1}^{n}\beta_i\cdot\boldsymbol{E}_{\mathrm{c}}^{i} \tag{18}$$

2.4　分类层

首先，将交互注意力矩阵 $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$ 进行拼接操作得到 $\boldsymbol{E}$，然后，将 $\boldsymbol{E}$ 进行非线性变换操作得到 $\boldsymbol{X}$，过程如下：

$$\boldsymbol{E}=\boldsymbol{E}_1\oplus\boldsymbol{E}_2 \tag{19}$$

$$\boldsymbol{X}=\tanh(\boldsymbol{W}\cdot\boldsymbol{E}+\mathrm{bia}) \tag{20}$$

其中，$\boldsymbol{W}$ 和 bia 分别表示权重矩阵和偏置。

那么，关于 X 的情感预测过程为：

$$y_i=\frac{\exp(X_i)}{\sum_{i=1}^{Z}\exp(X_i)} \tag{21}$$

其中，Z 表示情感极性类别数，y_i 表示训练样本的预测结果。设模型参数用 $\boldsymbol{\Theta}$ 表示，那么模型的训练损失 Loss1 为：

$$\mathrm{Loss1}=-\sum_{i=1}^{Z}y^{*}\cdot\ln(y_i)+\delta\sum_{\theta\in\boldsymbol{\Theta}}\theta^2 \tag{22}$$

其中，y^{*} 表示真实标签的分布，δ 表示 $L2$ 规范化系数。

特殊地，ASIAN 融合了对抗学习算法学习特征，ASIAN 通过本文第 2 节介绍的 PGD 算法，生成基于当前模型梯度的微小扰动 r_1 和 r_2。将扰动叠加在训练样本上生成对抗样本。对抗样本作为一种特殊的文本数据增强，也和训练样本一致，依次通过编码层、注意力层、分类层，设对抗样本的情感预测结果用 Y_i 表示，那么有：

$$\mathrm{Loss2}=-\sum_{i=1}^{Z}y^{*}\cdot\ln(Y_i)+\delta\sum_{\theta\in\boldsymbol{\Theta}}\theta^2 \tag{23}$$

即对抗样本的标签，采用对应训练样本的标签训练优化。

综上，ASIAN 模型的最终损失为：

$$\mathrm{Loss}=\mathrm{Loss1}+\mathrm{Loss2} \tag{24}$$

ASIAN 依据 2 个不同的损失函数学习和优化参数。

3　实验及结果分析

3.1　数据集

本文采用的 3 个公开数据集是：SemEval-2014 任务 4 中的 Laptop 和 Restaurant 数据集[18]、ACL-2014 的 Twitter 数据集[19]。每个数据集由积极、中性、消极三类样本构成，3 类样本的训练和测试数据量见表 1。此外，每次训练模型时，从训练数据中随机抽取 20%作为验证集。

表 1　实验数据集

Table 1　Experimental dataset

Datasets	Positive		Negative		Neural	
	Train	Test	Train	Test	Train	Test
Laptop[18]	994	341	870	128	464	169
Restaurant[18]	2 164	728	807	196	637	196
Twitter[19]	1 561	173	1 560	173	3 127	346

3.2　评估标准

（1）准确率（ACC）：在总样本中，分类正确的样本比例，公式如下：

$$\mathrm{ACC}=\frac{1}{Z}\left(\sum_{i=1}^{Z}\frac{T_i}{T_i+F_i}\right) \tag{25}$$

其中，Z 表示情感类别总数；T 表示分类准确的样本量；F 表示分类错误的样本量。

（2）$F1$ 值：由精度和召回率表示，对模型进行综合衡量。精度表示查准的概率，召回率表示查全的概率，公式如下：

$$F1=\frac{1}{Z}\left(\sum_{i=1}^{Z}\frac{2\cdot P_i\cdot R_i}{P_i+R_i}\right) \tag{26}$$

3.3 实验环境搭建及超参数配置

在 Linux 操作系统和 Nvidia GeForce RTX 2080 环境下,基于 PyTorch 深度学习框架进行实验。优化器采用 Adam、初始化采用 Xavier 正态初始化、学习率为 1e-5、L2 正则化衰减率为 0.01、batch 大小为 32、BERT 版本为:bert-base-uncased、dropout 依据不同数据集灵活调整、对抗样本生成采用 PGD 算法、PGD 扰动步长为 0.3、阈值为 1。

3.4 对比实验

(1)TD-LSTM 模型。由 Tang 等学者[20]提出的一种增强方面词与上文、方面词与下文联系的情感分类模型。

(2)ATAE-LSTM 模型。由 Wang 等学者[21]提出的一种方面词融合上下文单词共同嵌入编码的情感分类模型。

(3)RAM 模型。由 Chen 等学者[22]提出的一种基于记忆切片分配权重,结合多层注意力机制提取特征的情感分类方法。

(4)AOA 模型。由 Huang 等学者[23]提出一种提取方面词和上下文交互的注意力矩阵行、列特征向量,行、列特征向量再交互的情感分类方法。

(5)MGAN 模型。由 Fan 等学者[24]提出的一种多粒度注意网络,增强方面词与上下文的交互,减轻粗粒度注意力机制中的信息损失。

(6)BERT 模型。由 Devlin 等学者[25]提出的 BERT 模型,当时刷新了 11 项自然语言处理任务,BERT 在情感分类模型也有不俗的表现。

(7)ASGCN 模型。由 Zhang 等学者[26]提出基于图神经网络的情感分类方法,充分利用了句法信息和单词依存关系进行情感分类。

(8)GCAN-C 模型。由赵志影等学者[27]提出一种结合图卷积神经网络的注意力模型,该模型引入了句法分析树挖掘句法信息和结构。

ASIAN 与基线模型的对比实验结果见表 2。表 2 中,TD-LSTM 和 ATAE-LSTM 模型主要考虑的是方面词与上下文间的词级交互,分别是将方面词与上文内容拼接嵌入,方面词与下文内容拼接嵌入或方面词直接与上下文中的单词依次拼接嵌入,在 3 个公开数据集上的效果一般,提取特征的能力有待提升。

RAM、AOA 和 MGAN 通过多层注意力机制或多粒度注意力机制构造模型,模型提取得到的特征来源比较丰富。RAM、AOA 和 MGAN 模型在 Laptop、Restaurant 和 Twitter 数据集上的性能平均高出 ATAE-LSTM 模型约 3%,侧面体现 ATAE-LSTM 模型的构造较为简单,提取特征的层次浅。同时,也展现了基于多层注意力机制或多粒度注意力机制构造的模型,提取特征的能力往往比较优秀。

表 2 对比实验结果

Table 2 Comparison of experimental results %

Models	Laptop		Restaurant		Twitter	
	ACC	*F*1	ACC	*F*1	ACC	*F*1
TD-LSTM[20]	68.13	-	75.63	-	70.80	69.00
ATAE-LSTM[21]	71.83	68.43	78.00	66.73	66.62	64.01
RAM[22]	74.49	71.35	80.23	70.80	69.36	67.30
AOA[23]	74.50	-	81.10	-	-	-
MGAN[24]	75.39	72.47	81.25	71.94	72.54	70.81
BERT[25]	75.29	71.91	81.94	71.54	-	-
ASGCN[26]	75.55	71.05	80.86	72.19	72.15	70.40
GCAN-C[27]	76.49	**72.60**	82.20	**73.62**	72.49	70.93
ASIAN	**77.12**	71.71	**82.59**	73.45	**73.27**	**71.72**

上述基线模型均采用循环神经网络,而 BERT 和 ASGCN 则是 2 种完全不同于上述模型的深度神经网络模型。BERT 和 ASGCN 在 Laptop、Restaurant 和 Twitter 数据集上的效果与 MGAN 模型接近,优于大多数的循环神经网络基线模型,说明了 BERT 和图神经网络在情感分类问题有一定研究前景。

ASIAN 模型采用基于 BERT 的交互注意力网络提取特征,通过生成和学习对抗样本的方式,优化模型的分类效果。在 Laptop、Restaurant 和 Twitter 数据集上的准确率分别为 77.12%、82.59%和 73.27%,比基于循环神经网络的 MGAN 模型分别高 1.73%、1.34%、0.73%,对比 BERT 模型分别高 1.83%、0.65%,对比基于图神经网络的 ASGCN 模型分别高 1.57%、1.73%、1.12%,比 GCAN 分别高 0.63%、0.39%、0.78%。数据显示,与基于循环神经网络、BERT、图神经网络的基线模型对比,ASIAN 具有一定的优势,能提升情感分类的准确率。

值得注意的是,ASIAN 在 Laptop、Restaurant 数据集的 *F*1 值分别低于 GCAN-C 模型 0.89%、0.17%,说明了每个模型对数据集的适应程度不一样,GCAN-C 模型在 2 个数据集上分类正样本的能力要强于 ASIAN 模型。

3.5 消融实验

本小节主要验证 ASIAN 模型结构设计的合理性,将 ASIAN 不生成和学习对抗样本的情况,用 ASASIAN_NO 表示。同时,这一部分还引入了基于循环神经网络的 IAN 模型[28]和基于双向循环神经

网络的 BIAN 模型[29]进行对比验证分析，ASIAN 与各模型的消融对比实验结果见表 3。

表 3　消融实验结果

Table 3　Ablation experiment results　%

Models	Laptop		Restaurant		Twitter	
	ACC	*F*1	ACC	*F*1	ACC	*F*1
IAN[28]	72.10	63.72	78.71	67.71	69.22	66.90
BIAN[29]	72.30	66.70	79.40	69.90	–	–
ASIAN_NO	74.92	68.79	80.36	69.09	70.23	68.14
ASIAN	**77.12**	**71.71**	**82.59**	**73.45**	**73.27**	**71.72**

数据显示，ASIAN_NO 在 Laptop、Restaurant 和 Twitter 数据集上的分类准确率达 74.92%、80.36%、70.23%，对比 IAN 准确率分别提升 2.82%、1.65%、1.01%，对比 BIAN 高 2.62%、0.96%，实验结果表明即便是基础版本的 ASIAN 模型，对比 2 种不同版本的交互注意力网络模型，也有不俗的文本分类效果，提取特征的能力更出色。ASIAN 对比 ASIAN_NO 在 Laptop、Restaurant 和 Twitter 数据集上的准确率分别提升 2.2%、2.23%、3.04%，说明 ASIAN 采用生成和学习对抗样本的策略有效，能优化模型的分类效果。

4　结束语

早期的情感分类大多聚焦于传统机器学习算法，主要存在特征提取过程复杂、分类精度低、泛化能力弱等问题，深度网络模型的出现，一定程度缓解了上述问题。针对方面级情感分类模型存在特征提取能力弱、方面词与上下文交互不充分的问题，提出基于交互对抗网络的方面级情感分类方法，采用基于 BERT 的交互注意力网络提取方面词和上下文的交互信息，生成和学习对抗样本优化模型的分类。通过在 3 个公开数据集的对比实验结果数据表明，ASIAN 在 Laptop、Restaurant、Twitter 数据集上的准确率分别达到了 77.12%、82.59%、73.27%，优于 8 种深度网络基线模型。消融实验结果显示，ASIAN 生成和学习的对抗样本数据增强策略有效，对不同数据集有不同幅度的提升，最高提升幅度可达 3.04%。

虽然 ASIAN 取得了不错进展，但是 ASIAN 也有一定的局限性，如 ASIAN 在 Laptop 和 Restaurant 数据上正样本分类正确的能力较 GCAN-C 模型偏弱。

参考文献

[1] 张华辉，邱晓莹，徐航. 文本情感分类方法研究综述[J]. 延边大学学报(自然科学版)，2023，49(3)：275-282.

[2] 黄佳锋，薛云，卢昕，等. 面向中文网络评论情感分类的集成学习框架[J]. 中文信息学报，2018，32(9)：113-122.

[3] 王伟，周咏梅，阳爱民，等. 一种基于 LDA 主题模型的评论文本情感分类方法[J]. 数据采集与处理，2017，32(3)：629-635.

[4] 刘丽，岳亚伟. 面向高校学生微博的跨粒度情感分析[J]. 计算机应用研究，2019，36(6)：1618-1622.

[5] 尹春勇，章荪. 面向短文本情感分类的端到端对抗变分贝叶斯方法[J]. 计算机应用，2020，40(9)：2536-2542.

[6] 王立志，慕晓冬，刘宏岚. 采用改进粒子群优化的 SVM 方法实现中文文本情感分类[J]. 计算机科学，2020，47(1)：231-236.

[7] 杨玉亭，冯林，代磊超，等. 面向上下文注意力联合学习网络的方面级情感分类模型[J]. 模式识别与人工智能，2020，33(8)：753-765.

[8] 刘书齐，王以松，陈攀峰. 基于 CNN-ATTBiLSTM 的文本情感分析[J]. 贵州大学学报(自然科学版)，2019，36(2)：85-89.

[9] 张华辉，冯林，廖凌湘等. 基于对抗学习的自适应加权方面级情感分类算法[J]. 小型微型计算机系统，2023，44(4)：766-772.

[10] ZENG B，YANG H，XU R，et al. LCF：A local context focus mechanism for aspect-based sentiment classification[J]. Applied Sciences，2019，9(16)：3389.

[11] 张立，肖志勇. 多目标依存建模在特定目标情感分类中的应用[J]. 中文信息学报，2022，36(5)：133-144.

[12] 潘文雯，王新宇，宋明黎，等. 对抗样本生成技术综述[J]. 软件学报，2020，31(1)：67-81.

[13] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv，1312.6199，2013.

[14] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv，1412.6572，2014.

[15] 杜小虎，吴宏明，易子博，等. 文本对抗样本攻击与防御技术综述[J]. 中文信息学报，2021，35(8)：1-15.

[16] MIYATO T，DAI A M，GOODFELLOW I J. Adversarial training methods for semi-supervised text classification[EB/OL]. (2016-11-07). https://arxiv.org/abs/1605.07725.

[17] MADRY A，MAKELOV A，SCHMIDT L，et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv，1706.06083，2017.

[18] PONTIKI M，GALANIS D，PAVLOPOULOS J，et al. SemEval-2014 Task 4：Aspect based sentiment analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). ACL，2014：27-35.

[19] DONG Li，WEI Furu，TAN Chuanqi，et al. Adaptive recursive neural network for target-dependent Twitter sentiment classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2：Short Papers). ACL，2014：49-54.

[20] TANG D，QIN B，FENG X，et al. Effective LSTMs for Target-Dependent Sentiment Classification[C]//Proceedings of the 26th International Conference on Computational Linguistics：Technical Papers. Osaka，Japan：The COLING 2016 Organizing Committee，2016：3298-3307.

[21] WANG Y，HUANG M，ZHU X，et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of 2016

Conference on Empirical Methods in Natural Language Processing. ACL, 2016: 606-615.

[22] CHEN Peng, SUN Zhongqian, BING Lidong, et al. Recurrent attention network on memory for aspect sentiment analysis[C]// Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. ACL, 2017: 452-461.

[23] HUANG Binxuan, OUYANG Lan, CARLEY K M. Aspect level sentiment classification with attention - over - attention neural networks[M]//THOMSON R, DANCY C, HYDER A, et al. Social, Cultural, and Behavioral Modeling. Cham: Springer, 2018: 197-206.

[24] FAN Feifan, FENG Yansong, ZHAO Dongyan. Multi-grained attention network for aspect-level sentiment classification[C]// Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. ACL, 2018: 3433-3442.

[25] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding [C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2019, 1: 4171-4186.

[26] ZHANG Chen, LI Qiuchi, SONG Dawei. Aspect - based sentiment classification with aspect - specific graph convolutional networks[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP - IJCNLP). ACL, 2019: 4568-4578.

[27] 赵志影, 邵新慧, 林幸. 用于方面情感分析的结合图卷积神经网络的注意力模型[J]. 中文信息学报, 2022, 36(7): 154-163.

[28] MA Dehong, LI Sujian, ZHANG Xiaodong, et al. Interactive attention networks for aspect-level sentiment classification[J]. arXiv preprint arXiv, 1709. 00893, 2017.

[29] 陈虹, 杨燕, 杜圣东. 用户评论方面级情感分析研究[J]. 计算机科学与探索, 2021, 15(3): 478-485.