小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0039
2026 年 2 月 第 2 期
Vol.47 No.2 2026

面向教学评估的图注意力网络情感分析模型

柯昌博,任知临,张伯雷

(南京邮电大学 计算机学院,南京 210023)
E-mail:1222046026@njupt.edu.cn

摘 要:随着深度学习的发展,预训练模型、图神经网络等技术的广泛应用,课堂教学评价已成为人工智能和智慧教育领域的研究热点.本文提出了图注意力网络和预训练模型 BERT 相结合的细粒度情感分析模型,其中方面类别情感分析(ACSA)被分为方面检测(ACD)和方面情感分析(ASC)且共用同 BERT 参数,同时利用句子依存关系的图注意力网络强化上下文语义来提升模型性能,实验证明了在公开数据集中,本文的模型具有更高的准确率.在案例分析中,本文使用了自主设计并标注的包含 13 个方面的课堂教学评价数据集,并通过实验证明了本文的模型在多方面、小样本情感分析等任务上的优越性.最后,总结现有神经网络模型在情感分析任务中遇到的挑战,并展望未来研究的潜在方向.

关 键 词:图注意力网络;BERT 模型;课堂教学评价;注意力机制;依存关系
中图分类号:TP311 **文献标识码**:A **文 章 编 号**:1000-1220(2026)02-0274-08

Teaching Evaluation-oriented Graph Attention Network Sentiment Analysis Model

KE Changbo,REN Zhilin,ZHANG Bolei
(School of Computer Science,Nanjing University of Posts and Telecommumications,Nanjing 210023,China)

Abstract:With the development of deep learning,the application of pre-trained models and graph neural networks has made classroom teaching evaluation a research hotspot in artificial intelligence and smart education. This paper proposes a fine-grained sentiment analysis model that combines a graph attention network with the pre-trained model BERT. In this model,Aspect Category Sentiment Analysis (ACSA) is divided into Aspect Category Detection (ACD) and Aspect Sentiment Classification (ASC),with both tasks sharing the same BERT parameters. The model enhances contextual semantics by using a graph attention network based on sentence dependency relationships,improving its performance. Experiments show that the proposed model achieves higher accuracy on public datasets. In the case study,a classroom teaching evaluation dataset with 13 aspects,designed and annotated in this paper,is used to demonstrate the model's superiority in multi-aspect and few-shot sentiment analysis tasks. Finally,the paper summarizes the challenges faced by neural network models in sentiment analysis and suggests potential directions for future research.

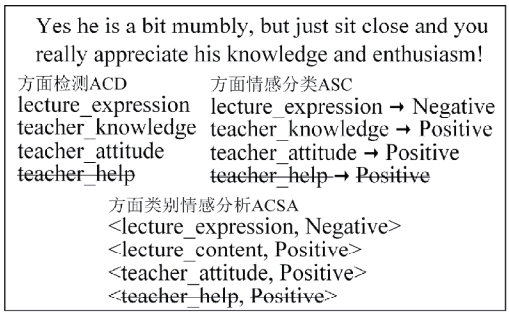
Keywords:graph attention network;Bert;teaching evaluation;attention network;dependency relations

0 引 言

随着人工智能的快速发展,以及 OBE 教育理念的进一步推广,以智慧教育为导向的评价体系需要多维度、多角度,对开设课程的教学内容、教学设计、教学质量和教学成果等全面准确地评估.在这些工作中普遍存在着大量的评价文本,根据评价标准以及上下文信息进行高效且准确的分析成为当前教学评价体系的研究热点.

传统课堂教学评估中采用的调查问卷方式,忽视了意见反馈文本中包含的丰富且主观的意见,且处理大量文本效率低下,难以充分利用评价文本.为了解决上述问题,近年来研究者使用各种细粒度情感分析方法,如预训练模型,图神经网络等理解文本语义,从而自动化获取文本中的意见.

实际课堂教学评价过程中对教师评价角度较多,且教学评价文本中可能同时包含多个情感信息,尽量全面准确地提取出全部情感信息成为自动化分析教学评价文本的研究焦点.方面类别情感分析(Aspect Category Sentiment Analysis,ACSA)旨在提取句子中包含的方面及其对应的情感.然而,随着教学评价角度的增加,即:方面数量的增加,模型性能会逐渐降低,且实际应用中不同方面的样例数量难以保持均衡,因此方面数量较多时保持模型的性能成为挑战.



此样例来自本文提出的包含 13 个方面项的教学评价数

收稿日期:2025-01-20 收修改稿日期:2025-03-13 基金项目:国家杰出青年项目(62125203)资助;国家自然科学基金重点项目(61932013)资助;国家自然科学基金面上项目(62072253)资助;江苏省自然科学基金项目(BK20221327)资助. 作者简介:**柯昌博**,男,1984 年生,博士,副教授,研究方向为车联网安全、人工智能;**任知临**(通信作者),男,1999 年生,硕士研究生,研究方向为人工智能;**张伯雷**,男,1988 年生,博士,副教授,研究方向为社会计算和强化学习.

据集,在数据集中包含方面 teacher_knowledge 的样例较少,包含 lecture_expression、teacher_attitude 和 teacher_help 的样例较多为主流方面,现有模型容易额外生成不必要的主流方面如 teacher_help 并忽视样例较少的方面,进而影响情感分类结果降低了模型的全面性和可靠性.

对于现有情感分析模型,Li 等人[1]将二元组提取任务分为方面检测和方面情感分类两个子任务,并假设对应方面的情感极性为句子中每个单词的情感极性之和,在已有公开数据集中取得了良好的结果. Hu 等人[2]使用 LSTM 层获取句子的向量化表示,利用正则化层加强两个子任务之间的联系,但方面项并没有紧密联系句子上下文语义,多方面项时容易忽视非主流方面. Bai 等人[3]提出将双向依存关系应用于图注意力网络以强化句子的上下文语义,利用外部知识提高模型理解能力,使情感分类性能得到提高. Schmitt 等人[4]使用端到端的 LSTM 结合卷积神经网络共同建模方面检测和方面情感分类任务,但池化得到的单一向量包含的信息量有限,随着方面数量增加句子中包含多个二元组时难以同时检测多个方面并分析情感极性. Zhang 等人[5]将句法信息和长距离的依存关系融入分类模型中,但单词节点距离作为权重难以提高句子语法关系较为复杂时的准确率. Huang 等人[6]结合句法依赖图和图注意力网络建模加强了句法结构和方面词相关性,使得情感分析的结果更为准确. Liang 等人[7]使用两个图卷积网络分别用于句法分析树和依存关系图,但影响了计算效率. Zhang 等人[8]通过层次句法图和层次词汇图联合建模分层统合上下文语义,使用图神经网络结合外部知识加强单词间语义表达,在公开数据集中提高了情感分类准确性. Tian 等人[9]利用 BERT 在迁移学习中获取上下文语义,在方面数量较多时,能更好地捕捉方面之间地差异性. Wan 等人[10]通过 BERT 获取句子的上下文信息并与方面项紧密结合,利用 BIO 标签检测方面项和 CLS 分类情感,但在方面项较多时降低了模型性能. Sun 等人[11]对 BERT 模型进行改进,使用问答和推理两种方式,当样例数量不平衡时容易分类到主流方面导致召回率较低. Liu 等人[12]利用 BART 模型通过文本生成的方式提取句子中对应方面和情感极性取得了较好的成果,但在不同数据集下主要依赖模型本身的性能.

使用 LSTM 的模型轻量化、训练快适用于算力要求较低的场景,但难以从大量训练数据中理解文本信息并总结经验. GCN 在每个节点上应用相同的权重,限制了对不同节点的差异性捕捉能力,而图注意力网络可以通过多头注意力机制计算不同的权重从而更灵活地处理节点间的关系,并且在方面数量增加时更容易区分不同的方面,增强了模型的泛化能力. 因此,本文提出了一种结合 BERT[13]和图注意力网络的方面类别情感分析模型. 本文将方面类别情感分析任务分为方面检测(Aspect Category Detection,ACD)和方面情感分类(Aspect Sentiment Classification,ASC)两个子任务,在两个子任务中共用训练同一个 BERT 加强方面项和情感极性之间的联系,并利用结合依存关系的图注意力网络强化上下文语意,在方面检测中使用注意力机制作为分类层为每个方面单独计算注意力权重,优化非主流方面项感知,同时将注意力权重应用于情感分类任务进一步区分不同方面的情感极性,最后得到句子中包含的多个方面情感二元组. 由于公开数据集中方面数量较少,而实际教学评价中考察方面较多,因此在案例分析部分,本文收集真实的教学评价文本数据并标注情感信息,最终构建了一个包含 2820 条数据、涵盖 13 个方面的教学评价数据集,通过实验,本文在方面数量性能影响实验、小样本情感分析实验、多方面多情感数据集实验上均优于其他模型. 本文的贡献如下:

1)提出了结合预训练模型 BERT 和图注意力网络的细粒度情感分析模型.

2)利用图注意力网络与依存关系相结合的方法,加强神经网络对上下文语义的理解.

3)收集并标注了教学评价数据集,有助于验证模型在多方面多情感文本上的表现.

1　方面类别情感分析模型

如图 1 所示,情感分析模型包含方面检测 ACD 和方面情感分类 ASC 两个子模型. 在方面检测 ACD 中,输入句子可以输出其包含的全部方面项;在方面情感分类 ASC 中,每次输入句子和一个方面项,可以输出对应方面的情感极性. 在编码层,句子中单词 w_i 通过嵌入转化为向量,方面检测将句子的

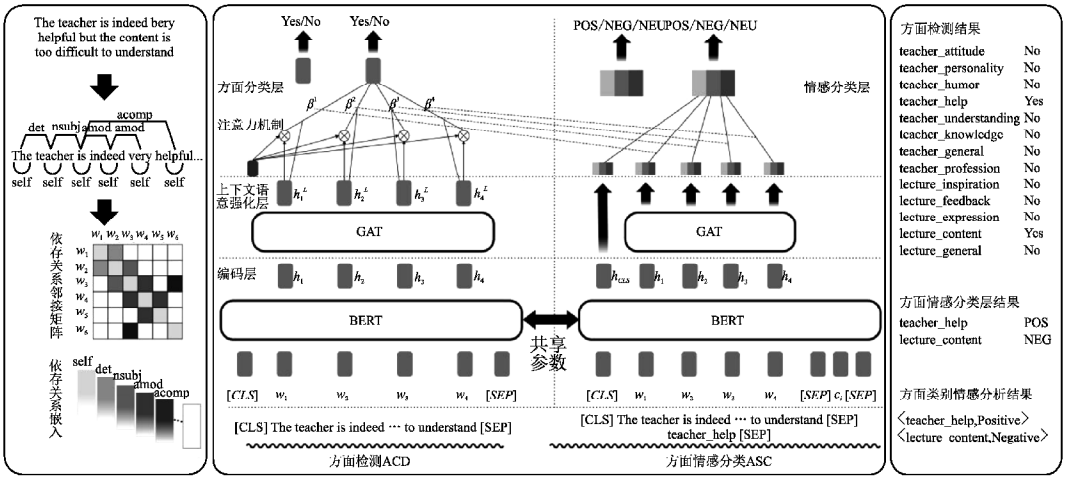


图 1　模型结构

Fig. 1　Model architecture

向量化表示,输入到 BERT 模型中;方面情感分类将句子的向量化表示与方面项 c_i 的向量化表示,也输入到 BERT 模型中,输出为句子带着上下文语意的向量化表示 h_i,且子模型共用同一个 BERT 参数. 在上下文语意强化层,将 BERT 的输出作为图注意力网络的初始输入,使用结合依存关系的图注意力网络强化句子的上下文语意表示,输出强化后对应句子中单词的向量化表示 h_i^L. 在方面分类层,通过注意力机制计算每个方面对应句子中词向量的注意力权重 β,并结合图注意力输出 h_i^L 分类方面项(Yes 包含 No 不包含此方面),在情感分类层中计算 CLS 标签 h_{CLS}和每个词向量 h_i^L 的情感极性,并利用方面分类层中的注意力权重加权求和,分类得到对应方面 c_i 的情感极性(Pos,Neu,Neg). 最后,方面检测和方面情感分分类结果组成方面情感二元组.

方面集合 $C=\{c_1,c_2,\cdots,c_N\}$,其中,c_i 为方面项. 情感极性集合 $P=\{Pos,Neu,Neg\}$,其中 Pos 为积极情感,Neu 为中立情感,Neg 为消极情感. 给定句子 $S=\{w_1,w_2,\cdots,w_n\}$,w_i 为句子中的单词. 句子中包含的 $K(K\leqslant N)$ 个方面 $C^S=\{c_1^S,c_2^S,\cdots,c_K^S\}$,且 $C^S\subset C$,对应 K 个情感极性 $P^S=\{P_1^S,P_2^S,\cdots,P_K^S\}$,且 $P_i^S\subset P$.

在方面检测子模型中,句子 S 以“[CLS] + S + [SEP]”组成序列,转化为词向量输入到 BERT 模型中,得到句子中单词 w_i 的上下文语意的输出表示 $H^D=\{h_1^D,h_2^D,\cdots,h_n^D\}$.

在方面情感分类子模型中,方面集合 C 中每个方面项 c_i 分别与句子 S 以“[CLS] + S + [SEP] + c_i + [SEP]”组成序列,转化为词向量输入到 BERT 模型中,得到句子中单词 w_i 关于特定方面 c_i 的上下文语意的输出表示 $H^A=\{h_1^A,h_2^A,\cdots,h_n^A\}$,$h^x=h_{CLS}^A$表示 CLS 标签对应的输出.

1.1 上下文语义强化层

依存关系表示句子中单词之间的语法关系,图 1 左侧部分描述了句子中依存关系的处理过程. 评论句子的依存关系由语法分析工具获取,图中弧上标记了两个单词之间的依存关系 r,如句子“The teacher is indeed very helpful …”中 is 和 indeed 的之间的状语关系 amod,is 和 teacher 之间的名词性主语关系 nsubj,且每个单词包含自环关系 self. 依存关系图对应生成依存关系邻接矩阵,并通过嵌入将依存关系转换至维度 d_r 的特征向量.

在图注意力网络部分,每个图注意力层由上一层的输出作为输入,编码层的输出 H^D 作为图注意力网络初始第 0 层词向量的输入 H^0,在图注意力计算后使用前馈神经网络计算得到当前层输出. 图注意力网络计算见公式(1):

$$a_i^l=\Big\|_{z=1}^{Z}\sigma\Big(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{lz}\left(W_V^{lz}h_j^{l-1}+\omega_1 W_{V_r}^{lz}r_{ij}\right)\Big)\tag{1}$$

其中 l 表示当前层数,Z 表示注意力头数,σ 表示非线性激活函数,权重 $W_V^{lz}\in R^{\frac{d}{Z}\times d}$,$W_{V_r}^{lz}\in R^{\frac{d}{Z}\times d_r}$,单词 i 与 j 之间的依存关系 $r_{ij}\in R^{d_r}$,$\mathcal{N}(i)$表示与单词 i 存在弧的单词集合,$\omega_1\in R$ 为超参数.

权重 α_{ij}^{lz}计算见公式(2) ~ 公式(4):

$$\alpha_{ij}^{lz}=\frac{exp\left(f(h_i^{l-1},h_j^{l-1})+\omega_2 f(h_i^{l-1},r_{ij})\right)}{\sum_{j'\in\mathcal{N}(i)}exp\left(f(h_i^{l-1},h_{j'}^{l-1})+\omega_2 f(h_i^{l-1},r_{ij'})\right)}\tag{2}$$

$$f(h_i^{l-1},h_j^{l-1})=\frac{(W_Q^{lz}h_i^{l-1})^T(W_K^{lz}h_j^{l-1})}{\sqrt{d/Z}}\tag{3}$$

$$f(h_i^{l-1},r_{ij})=\frac{(W_Q^{lz}h_i^{l-1})^T(W_{K_r}^{lz}r_{ij})}{\sqrt{d/Z}}\tag{4}$$

其中 $\omega_2\in R$ 为超参数,$f(\cdot)$为缩放点积注意力函数,权重 $W_Q^{lz},W_K^{lz}\in R^{\frac{d}{z}\times d}$,$W_{K_r}^{lz}\in R^{\frac{d}{z}\times d_r}$.

图注意力网络的输出为 $H^l=\{h_1^l,h_2^l,\cdots,h_n^l\}$,在图注意力网络后使用前馈神经网络,见公式(5):

$$h_i^l=\delta(W_1a_i^{l-1}+b_1)W_2+b_2\tag{5}$$

其中 $W_1,W_2\in R^{d\times d}$,δ 为非线性激活函数.

GAT 层的输出是 $H^{DL}=\{h_1^{DL},h_2^{DL},\cdots,h_n^{DL}\}=\{h_1^L,h_2^L,\cdots,h_n^L\}$,$L$ 表示总层数.

在方面情感分类部分,使用与方面检测中相同结构的 GAT 层,将 H^A 输入到得到 GAT 层中输出 H^{AL},见公式(6):

$$H^{AL}=\{h_1^{AL},h_2^{AL},\cdots,h_n^{AL}\}=GAT(H^A)\tag{6}$$

1.2 方面分类层

注意力机制可以根据不同的方面计算句子中单词对应的注意力权重,其中 $q_j\in R^d$,$W_j\in R^{d\times d}$,$j\leqslant N$,δ 为非线性激活函数,注意力机制见公式(7):

$$\beta_j=softmax(q_j^T\delta(W_jH^{DL}+b_j))\tag{7}$$

由注意力权重和单词的向量表示计算方面预测结果,其中 $W_j\in R^{d\times 1}$,见公式(8):

$$\hat{y}_j=sigmoid(W_jH^{DL}\beta_j^T+b_j)\tag{8}$$

1.3 情感分类层

对于每个单词的向量表示,计算每个单词对应的情感极性,其中 $i\in n$,$W_i\in R^{d\times d}$,$W_A\in R^{d\times 3}$,见公式(9):

$$p_i=W_A\delta(W_ih_i^{AL}+b_i)+b_A\tag{9}$$

对于每个方面,结合方面检测中得到的注意力权重,计算句子中单词的情感极性之和,见公式(10):

$$p_j^m=softmax(\sum_{i=1}^{n}p_i\beta_{ij})\tag{10}$$

分类标签 CLS 的输出 h_j^x,计算其情感极性,见公式(11):

$$p_j^x=softmax(W_Ah_j^x+b_A)\tag{11}$$

分类标签 CLS 的情感极性结果与句子中单词的情感极性之和作为句子的情感极性,见公式(12):

$$p_j=softmax(p_j^m+p_j^x)\tag{12}$$

1.4 损失计算

方面检测任务中,损失表示为对于每个方面的二分类预测之和,损失函数见公式(13):

$$\ell_D=-\sum_{j=1}^{N}y_j\log\hat{y}_j+(1-y_j)\log(1-\hat{y}_j)\tag{13}$$

方面情感分类任务中,对于句子中包含的 K 个方面,损失表示为对于每个方面的情感多分类预测的损失之和. 使用交叉熵损失,损失函数见公式(14):

$$\ell_A=-\sum_{j=1}^{K}\sum_{s\in P}y_{j_s}\log p_{j_s}\tag{14}$$

模型对方面检测任务和方面情感分类任务同时训练,总损失值为两任务损失之和,见公式(15):

$$\ell=\ell_D+\ell_A+\lambda\|\theta\|_2^2\tag{15}$$

其中 λ 为 L2 正则化超参数,θ 为模型中的可训练参数.

2 实验与评价

2.1 模型参数设置

在本文的实验中,BERT 使用 BERT-base 模型,其中包含

12 个 Transformer 块,隐藏层大小为 768. 图注意力网络使用 SpaCy 分词并获取依存关系,使用 NVIDIA RTX4070 GPU 训练,模型可调参数如表 1 所示.

表 1　可调参数设置

Table 1　Adjustable parameter setting

可调参数	值
优化器	ADAM
batch size	MAMS:16 / Rest14:32
learning rate	0.00002
方面检测图注意力网络层数 L_1	1
方面情感分类图注意力网络层数 L_2	1
图注意力网络头数 z	12
正则化超参数 λ	0.00001
超参数 ω_1,ω_2	0.8,0.6
词嵌入维度 d	768
依存关系嵌入维度 d_r	64

1)图注意力网络层数与头数和依存关系嵌入维度确定

为了验证图注意力网络层数的影响,本实验设置方面检测和方面情感分类的图注意力网络层数取值范围为{0,1,2},则共有 9 种层数组合. 实验结果参见图 2,本实验在 Rest14(Tay 等人[14])数据集上测试,取每个 epoch 的方面检测 F1 值和方面情感分类准确率之和的平均值. 其中横坐标括号内数值分别表示方面检测和方面情感分类图注意力网络层数. 实验结果中(1,1)为最优结果,说明图注意力网络能够进一步强化语义信息提高模型的性能且同时适用于两个子任务,而使用两层图注意力网络的性能普遍低于使用一层图注意力网络的模型,原因可能是出现了 Over-Smoothing 问题,图中相邻的节点随着网络深度增加而相似,使得难以区分特征.

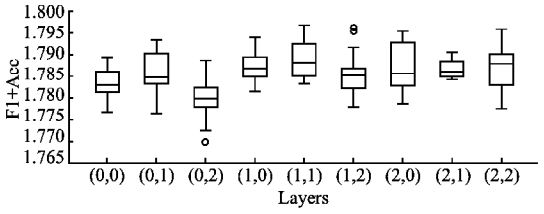


图 2　图注意力网络层数确定

Fig. 2　Determination of GAT layers experiments

为了验证图注意力网络中头数和依存关系嵌入维度的影响,本实验设置图注意力网络头数 z 分别取值为 12、6、4、3,且设置依存关系嵌入维度 d_r 等于每个头中词向量维度,则对应分别取维度大小为 64、128、192、256. 实验结果参见图 3,本实验在 Rest14(Tay 等人)数据集上测试,取每个 epoch 的方面检测 F1 值和方面情感分类准确率之和的平均值. 实验结果表明,设置头数 z 为 12 嵌入维度 d_r 为 64 时模型可以最快收敛达到最优结果,说明图注意力网络头数应保持与相连接的 BERT 模型中自注意力层的头数相同.

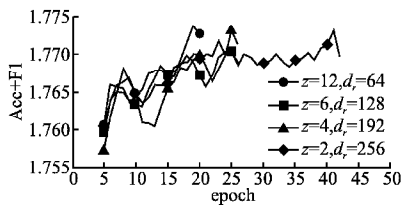


图 3　图注意力网络头数与依存关系嵌入维度确定

Fig. 3　Determination of GAT z and d_r experiments

2)超参数确定

在模型中包含 ω_1,ω_2 两个超参数,本实验设置参数取值范围为{0.2,0.4,0.6,0.8,1.0},得到 25 种超参数组合并分别在 Rest14(Tay 等人)数据集上测试,取方面检测 F1 值和方面情感分类准确率之和最高的结果. 本实验结果参见表 2.

表 2　超参数确定实验结果

Table 2　Result of hyperparameters determine experiments

ω_1,ω_2	0.2	0.4	0.6	0.8	1.0
0.2	1.7788	1.7812	1.7814	1.7796	1.7833
0.4	1.7822	1.7842	1.7829	1.7821	1.7845
0.6	1.7785	1.7847	1.7828	1.7811	1.7841
0.8	1.7840	1.7795	**1.7857**	1.7843	1.7846
1.0	1.7831	1.7841	1.7804	1.7796	1.7805

观察实验结果,本实验可以得出 ω_1,ω_2 通分别取值 0.8 和 0.6 时平均值达到最高,模型性能达到最优.

2.2　方面类别情感分析实验

为了验证方面类别情感分析任务的有效性即方面检测与方面情感分类在公开数据集上共同的有效性,本实验在 Rest14(Pontiki 等人[15])数据集上进行测试. 在方面检测中使用的精确率、召回率和 F1 值作为评价指标,在方面情感分类中使用四极性,三极性,二极性的准确率作为评价指标,其中四极性表示情感极性集合为{Positive,Neutral,Negative,Conflict},在三极性实验中去除 Conflict,在二极性实验中去除 Neutral. 本文的模型与以下基线模型进行比较:

XRCE[16]:基于机器学习的混合分类方法.

NRC-Canada[17]:提出序列标记,SVM 解决细粒度情感分析问题

AT-LSTM&ATAE-LSTM[18]:词嵌入融合方面嵌入,基于 LSTM 和注意力机制分析情感极性

Multi-task framework(MTL)[4]:基于 LSTM 的多任务学习模型,在 CNE-net 实验中的将其编码器改为 BERT-base,本文引用此实验数据

BERT-pair-NLI-B[11]:基于 BERT 的自然语言推理模型

BERT-pair-QA-B[11]:基于 BERT 的问答模型

表 3　方面类别情感分析对比

Table 3　Comparison of ACSA experiments

模型	方面检测			方面情感分类		
	精确率	召回率	F1	四极性	三极性	二极性
XRCE	83.23	81.37	82.29	78.1	–	–
NRC-Canada	91.04	86.24	88.58	82.9	–	–
AT-LSTM	–	–	–	–	83.1	89.6
ATAE-LSTM	–	–	–	–	84.0	89.9
MTL	91.87	90.44	91.15	–	–	–
BERT-pair-NLI-B	93.57	90.83	92.18	84.6	88.7	95.1
BERT-pair-QA-B	93.04	89.95	91.47	85.9	89.9	95.6
CNE-net	**93.76**	90.83	**92.27**	87.1	91.3	**96.4**
Our model	90.74	**92.68**	91.70	**87.4**	**91.6**	95.8

CNE-net[19]:基于注意力机制的分类器对 BERT 模型输

出进行分类

实验结果如表 3 所示,相比目前已知的最好模型 CNE-net 在方面检测中 F1 值降低了 0.57%,召回率提升了 1.85%,精确率降低了 3.02%,方面情感分类中四极性准确率提升了 0.3%,三极性准确率提升了 0.3%,二极性准确率降低了 0.6%. BERT-pair-NLI-B、BERT-pair-QA-B 使用推理和问答模型容易生成更多的方面项,使精确率较高,而额外的方面会导致召回率降低. CNE-net 中将所有方面项连接组成一个长输入序列提高了精确率,但随着方面项的增加,过长的输入序列会导致训练时间陡增且生成额外的方面项,导致召回率降低. 在教师评价情境下,与其它模型相比,本文的模型具有的较高方面检测召回率和情感分类准确率.

2.3 情感分类实验

方面情感分类实验中包含 3 组公开数据集,分别为 Rest14(Tay等人)、Rest14-hard(Xue等人[20])和MAMS-ACSA(Jiang 等人[21])数据集. Rest14-hard 数据集来自于 Rest14 中包含至少两个情感极性的样本集. MAMS-ACSA 为三极性的餐馆评论句子数据集,相比 Rest14 数据量更大且方面数量更多,可以验证在更多方面时情感分类的稳定性和准确性. 本实验数据集中情感极性数量参见表 4.

表 4 实验数据集统计信息

Table 4 Dataset statistics

数据集	极性	Pos.	Neg.	Neu.
Rest14	Train	1873	712	433
	Dev	306	127	67
	Test	657	222	94
Rest14-Hard	Test	21	20	12
MAMS	Train	1929	2084	3077
	Dev	241	259	388
	Test	245	263	393

实验以方面情感分类准确率的平均值和方差作为评价指标,与以下基线模型进行比较:

GCAE[20]:采用带有门控机制的 CNN 模型,输出给定方面的情感极性

AS-Capsules[22]:通过多个基于 RNN 的 capsule 模型输出方面情感二元组

CapsNet[21]:通过基于 capsule 网络的模型理解各个方面与上下文之间的关系

AC-MIMLLN[1]:通过基于注意力机制的多实例多标签学习检测方面并分析情感极性

CapsNet-BERT:基于 BERT 的 CapsNet

BERT-pair-QA-B:基于 BERT 的问答模型

AC-MIMLLN-BERT:基于 BERT 的 AC-MIMLLN

BART generation[12]:基于 BART 的自然语言生成模型

实验结果参见表 5. 本实验可以得到以下结论,与基于预训练的 CapsNet-BERT、BERT-pair-QA-B、AC-MIMLLN-BERT、BART generation 和非预训练的 GCAE、AS-Capsules、CapsNet、AC-MIMLLN 相比,本文的模型在 MAMS 和 Rest14-hard 数据集上超过了所有其他模型,相比于 BART generation 分别提升了 0.888% 和 1.887%,在 Rest14 数据集上略逊于 BART generation. 其中 MAMS-ACSA 数据集中包含 8 个方面,Rest14 数据集包括 5 个方面,这说明本文的模型在方面数量多的数据集中表现效果更好. Rest14-hard 数据集中的样本包含多种情感极性,这说明本文的模型可以很好的预测句子中给定不同方面的多个情感极性.

表 5 方面情感分类的实验结果,准确率(方差)%

Table 5 Results of ASC experiments Acc(Var)%

模型	Rest14	Rest14-hard	MAMS-ACSA
GCAE	81.336(±0.883)	54.717(±4.920)	72.098
CapsNet	81.172(±0.631)	53.962(±0.924)	73.986
AS-Capsules	82.179(±0.414)	60.755(±2.773)	75.116(±0.473)
AC-MIMLLN	81.603(±0.715)	65.283(±2.264)	76.427(±0.704)
CapsNet-BERT	86.557(±0.943)	51.321(±1.412)	79.461
BERT-pair-QA-B	87.523(±1.175)	69.433(±4.368)	79.134(±0.973)
AC-MIMLLN-BERT	89.250(±0.720)	74.717(±3.290)	81.198(±0.606)
BART generation	**90.545(±0.315)**	77.358(±2.160)	83.130(±0.478)
Our model	90.174(±0.324)	**79.245(±0.944)**	**84.018(±0.292)**

3 教学评价数据情感分析案例研究

本文收集真实的教学评价数据进行案例研究. 数据来自网站 www.ratemyprofessors.com 中学生对教师的教学评价. 并选取网站上评分接近3.0且评价数量多于150条的教师,收集教学评价文本 1000000 条,再通过无监督学习 ABAE[23]算法进行方面提取,在 ABAE 设置 100 个簇,每个簇在向量空间

表 6 各方面样例数量

Table 6 Number of cases in each category

方面	teacher								lecture				
	attitude	personality	humor	help	understanding	knowledege	general	profession	inspiration	feedback	expression	content	general
Pos.	149	64	192	273	26	65	545	60	27	16	46	178	97
Neu.	2	6	6	6	0	0	93	2	0	2	17	52	13
Neg.	50	174	134	177	19	3	378	63	1	30	93	212	120
总和	201	244	332	456	45	68	1016	125	28	42	156	442	230

中找到 20 个与簇中心最近的单词,同时参考词频统计结果确定了 13 个方面. 本文对教学评价语句进行方面情感标注,最后得到训练集和测试集分别为 2500 条和 320 条的包含 3 种情感极性的数据集. 数据集中各方面样例数量如表 6 所示,部分教学评价数据集样例如表 7 所示.

3.1　方面数量性能影响实验

为了验证方面数量对方面情感分类的影响,设置方面数分别为 5、7、9、11、13,将本文的模型与 BERT-pair-QA-B 和 BERT-pair-NLI-B 比较. 本实验按照数据集中各方面样例数量的大小,取样例数量最多的 5 个方面组成方面数 5 的数据集,以此类推. 实验结果如图 4 所示.

表 7　教学评价数据样例

Table 7　Samples of teaching evaluation dataset

教学评价文本	方面情感极性二元组
Kaplan is a nice guy and very friendly, but he makes the class unnecessarily difficult.	< teacher_personality, Positive > < teacher_general, Positive > < lecture_content, Negative > < teacher_profession, Negative >
This old man is very funny, he is good at math, but is also good at bewildering you.	< teacher_humor, Positive > < teacher_profession, Neutral >
OOP conccpts arc boring cnough, but coupled with Mercer's teaching style it makes the class awful.	< lecture_content, Negative > < lecture_general, Negative >
she does not listen to the complaints made by students, it was very frustrating when she kept ignore.	< teacher_personality, Negative > < lecture_feedback, Negative >

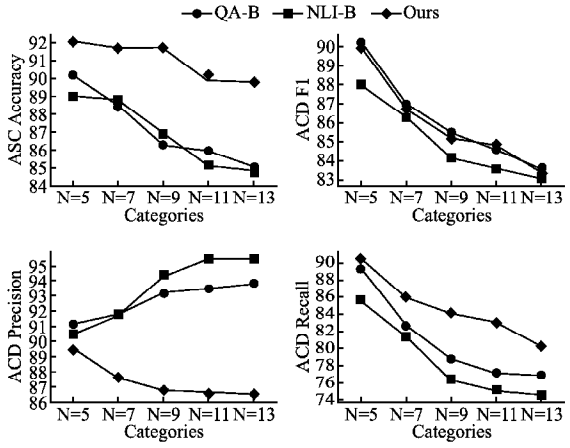


图 4　方面数量对性能的影响

Fig. 4　Category amount affection

从实验结果中可以看出本文的模型与 BERT-pair-QA-B 和 BERT-pair-NLI-B 的方面检测 F1 值比较接近,方面检测的精确率在 87% 左右相对较低,方面检测的召回率和方面情感分类的准确率保持较高. 本文推测由于方面项的增多类别间的语意相似度增加,注意力机制中向量 q 在向量空间中与其他向量距离更加接近,降低了精确率,又因为注意力机制中每个方面都对应一个独立的向量 q 对于数据集中包含样本更少的方面增加了召回率. 而文本问答和文本推理类模型在方面数量分布不均时倾向输出主流方面从而使精确率较高而召回率较低. 从整体角度 3 个模型在方面检测 F1,方面检测召回率和方面情感分类准确率均随方面数增加而降低. 本实验说明本文的模型在多方面情感分类时方面检测和方面情感分类都能保持较高的性能.

3.2　小样本情感分析实验

为了验证样本数量较少时模型的性能,本实验以 3.2 节实验中方面数为 9 的数据为本实验数据集,从中取出 400 个样例作为测试集,并取每种方面每种情感极性的样例数分别为 10、20、30、40、50 个作为训练集,将本文的模型与 BERT-pair-QA-B 和 BERT-pair-NLI-B 比较.

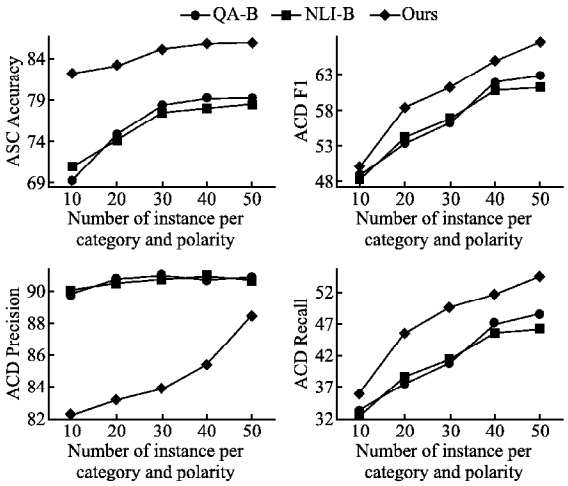


图 5　小样本情感分析结果

Fig. 5　Result of few-shot dataset ACSA

实验结果参见图 5,可以看出本文的模型相比 BERT-pair-QA-B 和 BERT-pair-NLI-B 方面情感分类准确率、方面检测 F1 值、方面检测召回率较高,方面检测精确率较低,从整体角度 3 个模型在方面情感分类准确率,方面检测 F1 值、方面检测召回率方面都随着样例数增加而增长. 本实验说明本文的模型在数据量较小时方面检测和方面情感分类都能保持较高的性能,模型迁移能力更强.

3.3　多方面多情感数据集实验

为了验证模型在样例包含多方面多情感时模型能性能,本实验抽取数据集中包含 2 到 3 个方面情感二元组的样例共 320 个组成测试集,其他 2500 个样例组成训练集,将本文的模型与 BERT-pair-QA-B 和 BERT-pair-NLI-B 比较. 在方面检测中记录测试集每个方面的准确率和总体的准确率,在方面情感分类中记录每个方面的情感准确率和总体的情感准确率,实验结果如图 6 所示.

从实验结果中可以看出,本文的模型在方面检测准确率和方面情感分类总体准确率均高于其他模型,对于训练样例较少的方面 teacher_knowledge、lecture_feedback、teacher-_understanding、lecture_inspiration,方面检测的准确率接近或超过其他模型,这验证了 3.1 节所述方面数量性能影响实验中的猜想. 在方面 lecture_general 包含 230 个样例但是方面检测准确率却只有 25.9%,本文推测是由于 lecture_general 与 teacher_general 等方面语义上比较接近,当评价者意图描述教师讲授水平优秀时,通常表达对教师本人的称赞,使得模型较难与 teacher_general 区分. 方面 lecture_content 包含 442 个样例准

确率只有 41.5%,本文推测是由于 lecture_content 对应的语义范围较大如课堂准备、教学设计、教学方式,且容易与其他方面语义相交,实际数据集中“将一门的很难的课讲的简单”、“讲解过程中偏离主题”、“PPT 设计差”等均被归为方面 lecture_content,导致模型难以准确分类,teacher_attitude 同理. 表 8 中展示了 4 个真实样例的预测结果.

预测结果中,3 个模型的方面检测任务都有可能缺少方面项,QA-B 模型会增加标签中没有的主流方面项如 Case 4,这符合 3.1、3.2 两组实验中本文的模型召回率比较高的特点.3 个模型的方面情感分类任务中 NLI-B 和 QA-B 两个模型容易得到错误的情感极性分类结果如 Case 1、Case 3,且 QA-B 模型容易使不同方面获得相同情感极性如 Case 3、Case4.

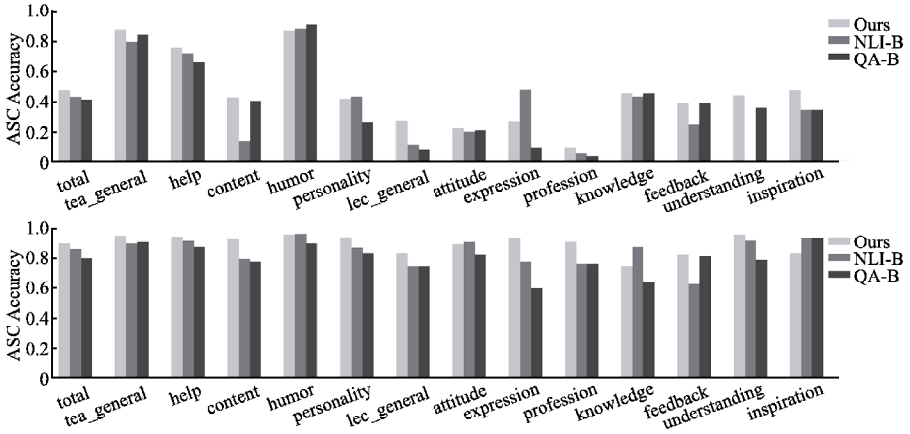


图 6 多方面多情感数据集对比

Fig. 6 Comparison of multi-category multi-polarity dataset ACSA

表 8 教学评价数据样例预测结果对比

Table 8 Case prediction comparison of teaching evaluation dataset

Case 1	He's very smart and lecture is not boring.
Gold	< teacher_knowledge, Positive > < teacher_humor, Positive >
Ours	< teacher_knowledge, Positive > < teacher_humor, Positive >
NLI-B	< teacher_knowledge, Negative > < teacher_humor, Positive >
QA-B	< teacher_humor, Negative >
Case 2	Einhaus knows his stuff but he persents them in a boring way.
Gold	< teacher_humor, Negative > < teacher_profession, Positive >
Ours	< teacher_humor, Negative > < teacher_profession, Positive >
NLI-B	< teacher_humor, Negative >
QA-B	< teacher_humor, Negative >
Case 3	Boring lectures, useless class, but fair grader and knowledgable person.
Gold	< teacher_humor, Negative > < teacher_help, Negative > < teacher_knowledge, Positive >
Ours	< teacher_humor, Negative > < teacher_knowledge, Positive >
NLI-B	< teacher_humor, Negative > < teacher_help, Negative >
QA-B	< teacher_humor, Negative > < teacher_knowledge, Negative >
Case 4	Burg did his best to make it fun.
Gold	< teacher_attitude, Positive > < teacher_humor, Positive >
Ours	< teacher_humor, Positive >
NLI-B	< teacher_humor, Positive >
QA-B	< teacher_humor, Positive > < teacher_personality, Positive >

4 总 结

实现智慧教育背景下课堂教学评价过程中评价文本可利用性和处理文本高效性是本文的主要目的,实际教学评价中对教师素质以及教学水平考察角度较多,公开数据集方面数量少难以比较多方面情况下模型的性能,所以在案例分析部分提出了由本文收集并标记的包含 13 个方面的教学评价数据集,实验结果表明本文的模型在方面数量较多时仍能保证较高的性能,且训练样例减少时对模型性能的影响较小,更适合应用于新领域或半监督学习.

在实验中也注意到一些问题,短文本情感分析不适用于一些长度较长的评价文本,且评价文本中包含一些俗语、颜文字、省略句等,神经网络在没有其他知识信息作为参考时难以分析情感,未来的工作中将尝试获取短文本间的语义信息和融入一些更高级的先验知识如知识图谱以弥补不足.

References:

[1] Li Y, Yin C, Zhong S, et al. Multi-instance multi-label learning networks for aspect-category sentiment analysis[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 3550-3560.

[2] Hu M, Zhao S, Zhang L, et al. Can: constrained attention networks for multi-aspect sentiment analysis[C]//Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019.

[3] Bai X, Liu P, Zhang Y. Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020: 503-514.

[4] Schmitt M, Steinheber S, Schreiber K, et al. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 1109-1114.

[5] Zhang C, Li Q, Song D. Syntax-aware aspect-level sentiment classification with proximity-weighted convolution network[C]//42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019: 1145-1148.

[6] Huang B, Carley K M. Syntax-aware aspect level sentiment classifi-

cation with graph attention networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,2019:5469-5477.

[7] Liang S, Wei W, Mao X L, et al. BiSyn-GAT +: bi-syntax aware graph attention network for aspect-based sentiment analysis[C]// Findings of the Association for Computational Linguistics, 2022: 1835-1848.

[8] Zhang M, Qian T. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2020:3540-3549.

[9] Tian Y, Chen G, Song Y. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2021:2910-2922.

[10] Wan H, Yang Y, Du J, et al. Target-aspect-sentiment joint detection for aspect-based sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2020:9122-9129.

[11] Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence[C]//Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 380-385.

[12] Liu J, Teng Z, Cui L, et al. Solving aspect category sentiment analysis as a text generation task[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 4406-4416.

[13] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019:4171-4186.

[14] Tay Y, Tuan L A, Hui S C. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence,2018:5956-5963.

[15] Pontiki M, Galanis D, Pavlopoulos J, et al. SemEval-2014 task 4: aspect based sentiment analysis[C]//8th International Workshop on Semantic Evaluation,2014:27-35, doi:10.3115/v1/s14-2004.

[16] Brun C, Popa D N, Roux C. XRCE: hybrid classification for aspect-based sentiment analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation,2014:838-842.

[17] Kiritchenko S, Zhu X, Cherry C, et al. NRC-Canada-2014: detecting aspects and sentiment in customer reviews[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: 437-442.

[18] Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2016:606-615.

[19] Dai Z, Peng C, Chen H, et al. A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2020:6955-6965.

[20] Xue W, Li T. Aspect based sentiment analysis with gated convolutional networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,2018:2514-2523.

[21] Jiang Q, Chen L, Xu R, et al. A challenge dataset and effective models for aspect-based sentiment analysis[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,2019:6280-6285.

[22] Wang Y, Sun A, Huang M, et al. Aspect-level sentiment analysis using as-capsules[C]//World Wide Web Conference,2019:2033-2044.

[23] He R, Lee W S, Ng H T, et al. An unsupervised neural attention model for aspect extraction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 388-397.