基于情感增强 BERT 与多任务生成对抗网络的虚假评论检测模型

李丹[1]，谢语涵[2]，韩潇帅[2]，吕晨[3]

（1. 上海财经大学浙江学院经济与信息管理系，浙江 金华 321000；2. 上海财经大学信息管理与工程学院，上海 200433；
3. 上海财经大学计算机与人工智能学院计算经济交叉科学教育部重点实验室，上海 200433）

摘　要：针对当前虚假评论检测模型存在的深层情感特征挖掘不足、语义依赖关系缺失以及泛化性能不佳的问题，提出一种基于情感加权 BERT 与多任务对抗学习的虚假评论识别模型 DEBR-GAN。首先，借助情感词典辅助预训练 BERT，通过情感加权机制对评论文本中的潜在情感信息进行提取，从而增强对评论中细微情绪变化的捕捉能力；随后，采用循环神经网络（RNN）对 BERT 输出的语义特征进行处理，充分挖掘评论中词语之间的时序依赖及上下文关系，以提高对文本细节的敏感性；接着，为提升模型在多领域场景下的鲁棒性与泛化能力，DEBR-GAN 借鉴了生成对抗网络（GAN）的对抗学习思想，将虚假评论检测器视为特征生成器，用于提取跨领域共享的有效特征，同时，通过设置类别鉴别器和评分鉴别器，在反向传播过程中采用梯度反转技术，与生成器进行对抗博弈，有效消除类别信息和用户评分偏好对特征提取过程的干扰，从而保证检测器在识别虚假评论时具有高准确性。实验结果表明，在大众点评数据集上，DEBR-GAN 模型的 F1 值高达 0.926，与未引入多任务对抗学习模块的模型相比，其分类准确率提高了 5.1 百分点，而相较于当前最佳基线模型则提升了 3.51 百分点。此外，该模型在处理不同情感倾向和语义结构的评论时均表现出较高的识别准确率，充分验证了情感增强与对抗学习相结合在虚假评论检测中的有效性与优越性。

关键词：情感增强；生成对抗网络；虚假评论检测；社交网络评论；BERT

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0252154

False Comment Detection Model Based on Sentiment-Enhanced BERT and Multi-Task Generative Adversarial Networks

LI Dan[1], XIE Yuhan[2], HAN Xiaoshuai[2], LÜ Chen[3]

(1. Department of Economics and Information Management, Shanghai University of Finance and Economics Zhejiang College, Jinhua 321000, Zhejiang, China;
2. School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China;
3. Key Laboratory of Computational Economics and Interdisciplinary Sciences, Ministry of Education, School of Computer Science and Artificial Intelligence, Shanghai University of Finance and Economics, Shanghai 200433, China)

【Abstract】Current false comment detection models face several problems such as insufficient mining of deep emotional features, lack of semantic dependency relationships, and poor generalization performance. In response to these, a false comment recognition model, DEBR-GAN, based on emotion-weighted BERT and multi-task adversarial learning, is proposed. First, using an emotion dictionary to assist in pretraining BERT, the potential emotional information in the comment text is extracted through an emotion weighting mechanism, thereby enhancing the ability to capture subtle emotional changes in the comments. Subsequently, a Recurrent Neural Network (RNN) is used to process the semantic features output by BERT, fully exploring the temporal dependencies and contextual relationships between words in comments for improving sensitivity to text details. Furthermore, to enhance the robustness and generalization ability of the model in multi-domain scenarios, DEBR-GAN draws on the adversarial learning concept of the Generative Adversarial Networks (GAN), treating the fake comment detector as a feature generator for extracting effective features shared across domains. Simultaneously, by setting category discriminators and rating discriminators, gradient reversal techniques are used in the backpropagation process to engage in adversarial games with the generator. This effectively eliminates the interference of category information and user rating preferences in the feature extraction process, thereby ensuring that the detector is highly accurate in identifying fake comments. The experimental results show that, on the Dianping dataset, the F1 value of the DEBR-GAN model is as high as 0.926. Compared with those of the model without the multi-task adversarial learning module and the current best baseline model, the classification accuracy of DEBR-GAN is increased by 5.1 and 3.51 percentage points, respectively. In addition, DEBR-GAN exhibits high

基金项目：国家自然科学基金（62476164）；教育部人文社会科学研究课题（24YJCZH197）。

作者简介：李丹，女，讲师、硕士，主研方向为自然语言处理、深度学习；谢语涵，博士；韩潇帅，硕士；吕晨（通信作者），副教授、博士。

收稿日期：2025-02-24　　**修回日期：**2025-04-17　　**E-mail：**lyu.chen@sufe.edu.cn

计算机工程 www.ecice06.com

recognition accuracy in handling comments with different emotional tendencies and semantic structures, thereby verifying the effectiveness and superiority of combining emotional enhancement and adversarial learning in false comment detection.
【Key words】 sentiment enhancement; Generative Adversarial Networks (GAN); fake comment detection; social network comment; BERT

0 引言

随着社交网络与电子商务的深度融合,用户生成内容(UGC)已成为消费者决策的重要依据[1]。大众点评、Yelp 等用户评论社交网络(URSN)通过整合多维评价构建产品评估体系[2]。然而,虚假评论的泛滥严重破坏了信息生态的可靠性。例如,TripAdvisor 曾曝出 5 500 万条可疑评论[3],亚马逊网站上 30%的评论涉嫌付费刷单(星级虚高 0.38)。《2023 年网络内容安全报告》显示,在全球电商平台中,约 23.7%的评论涉嫌虚假性诱导,这导致消费者信任度下降,市场公平性受损,国内也破获了涉案金额达 3.7 亿元的刷评产业链。尽管当前平台已部署过滤机制,但是虚假评论依然通过语义伪装持续渗透,亟需发展更精准的检测技术以维护生态的可信度。

现有研究通过语义分析、用户行为建模及深度学习(如 BERT、图神经网络)尝试识别虚假评论,但这些模型往往仅依赖单一维度的文本特征进行建模,未能充分挖掘文本信息的丰富性与复杂性,忽视了从多维度、多角度深入剖析文本内容的潜力,从而限制了检测效果的提升。例如,“帮妈妈买了个包,性价比超级高,又推荐同事买了钱包!”这条评论虽然在情感上表现出强烈的正面特征,但从语义关联性分析可知,它应当是虚假评论。此外,现有方法往往过度依赖文本的表层语义,忽视了虚假评论中情感表达与语义逻辑之间的矛盾性(如“五星好评”与负面情感词汇并存)。最后,现有模型易受特定领域类别(如餐饮、旅游)的用户行为模式干扰,难以适应跨类别场景检测需求。如文献[4]构建的餐饮类多模态检测模型,在迁移至本文数据集时 F1 值从 0.814 降至 0.731。

针对上述问题,本文提出了一种基于情感增强 BERT 和生成对抗网络(GAN)的虚假评论检测模型 DEBR-GAN。本文的主要贡献如下:

1)情感加权 BERT 模型的创新应用。本文提出了一种基于情感加权机制的 BERT 模型,通过情感词典辅助预训练 BERT,并对评论文本中的潜在情感信息进行加权提取,显著提高了模型对细微情绪变化的捕捉能力,为虚假评论检测中的情感特征挖掘提供了新思路。

2)时序依赖与上下文关系挖掘。结合循环神经网络(RNN)对 BERT 输出的语义特征进行进一步处理,充分挖掘评论中词语的时序依赖及上下文关系,提升模型对文本细节的敏感性与虚假评论的识别能力。

3)多任务对抗学习提升泛化能力。借鉴 GAN 的思想,通过设置类别鉴别器和评分鉴别器,采用梯度反转技术有效剥离领域特异性信息,增强模型在多领域场景下的鲁棒性与泛化能力,显著提高虚假评论检测的准确性。

1 相关工作

从研究内容的角度,可将国内外对虚假评论的研究大体分为 3 类,即对虚假评论文本的检测[5]、对虚假评论发布者的检测[6]以及对虚假评论用户组的检测[7]。从研究过程的角度,大致可以分为 3 个步骤,即虚假评论数据集的构建及标注、虚假评论特征提取以及虚假评论识别。

1.1 社交网络评论分析

社交网络评论形式涵盖文本、图片及视频,其中,图片与视频评论直观但耗时,因此,文本评论数量更多。文本评论包含语言风格、语义内容及情感倾向三要素。语言风格通常表示评论文本中含有某些特定类型词语的表达方式,是决定用户评论质量的关键指标[8]。有学者认为虚假评论与真实评论的语言特征存在明显差异,如问号和感叹号[9]、评论文本的长度或每句话的平均字数[10]、分类语言与动态语言[11]、语言风格[12]等;语义信息是评论文本中所蕴含的深层次含义、意图或概念。研究者们发现,大部分的虚假信息会包含若干个错误内容[13],以有限的误导性陈述,巧妙地嵌入大量看似真实的文本,内容中存在潜在的逻辑冲突或叙事不连贯[14],用多义词误导消费者[15],并且内容包含消费者兴趣映射[16]等特征,因而通过将主题信息和语义信息相结合[17]、从外部知识获取文本内容的背景知识[18]以及用户特征和内容特征相融合[19]等方式来增强潜在语义的识别能力。

情感倾向是消费者对自身消费经历的情感评价,可以在一定程度上影响潜在消费者的购买决策。情感分析又分为粗粒度[20]和细粒度[21]分析,细粒度的情感分析能更好地理解用户行为和意图,提升

产品性能及服务质量。研究者们采用基于情感词典[22]、机器学习[23]、深度学习[24]等方法对情感倾向进行分类。基于词典的方法通过句法依赖和词性特征提取文本情感特征,其分类性能较好,但是其对文本的语义结构特征获取不足,更好的方法是将词典和语义结构特征有效融合,提高情感分类的准确性。

1.2 虚假评论数据集特征提取

虚假评论数据集特征提取一般分为基于统计特征和基于文本内容两类。在统计特征方面,主要根据文本的语言特征,从字、词、词性频率[25]、文本情感极性比例[26]以及不同级别的 n-gram 模型等维度提取特征,然后将这些统计特征进行处理,再使用特征工程以及一些机器学习方法完成分类模型的构建;在文本内容特征方面,学者主要从文本长度[27]、文本相似性[28]、实体关系[29]等方面提取文本特征。

情感是消费者对产品使用感受的直观表现,有很多学者将情感特征与其他特征相融合,如将情感极性与逻辑回归相结合[27]、情感极性和语言结构相结合[30]、情感知识和语言特征单词[31]及句子相结合[32-33]等,从而提取深层次的文本特征。这类模型较多使用自然语言处理的方法实现,包括使用神经网络、预训练模型等对文本进行编码,然后构建检测模型,学习文本内容的相关特征,最终完成虚假评论识别。这类模型的性能相比统计模型有了较大的提升。

1.3 虚假评论识别方法

在虚假评论识别方法方面,早期学者采用支持向量机[25]、马尔可夫随机场[34]、随机森林[35]、朴素贝叶斯[36]等机器学习模型,对评论文本、评论者、评论群组进行虚假检测,整体性能表现较好,但是会存在识别精度低等问题。随着深度学习的发展,研究者们开发不同的框架来提高检测精度,如利用卷积神经网络[37]、注意力机制[38]、预训练模型[39]等单模型或者多模型融合[40]等方法。然而,社交网络评论领域往往存在数据类别失衡的问题,即评论数据集中某一类的数据占比过高,导致分类模型更趋向于将数据划分为大类别,使得模型泛化能力差,实用性降低。

在其他领域,有学者使用 GAN 来解决数据不平衡问题,提高模型的泛化能力。在虚假文本检测领域,文献[41]使用 GRU 网络构建生成器和判别器,通过生成器更好地学习谣言的潜在特征,提高谣言检测模型的精度。文献[42]通过循环生成对抗网络和 Wasserstein 距离改进生成损失,利用对抗网络提高谣言检测模型在样本不平衡和非配对情况下的稳定性和精确度。文献[43]基于 GAN 模型框架建立一个具有智能化自学习能力的谣言信息输入序列修正模型,解决了数据不平衡问题。

综上,在虚假评论文本识别中,主要从语言特征、文本特征、情感特征、产品特征等方面,采用统计学、自然语言处理、文本挖掘等方法进行研究。深入分析虚假评论检测领域所面临的挑战,可以将这些问题分为三大类:首先是数据层面的局限性,目前针对中文环境的虚假评论数据集稀缺,且现存数据中广泛存在标签缺失现象,这一状况直接制约了模型训练过程中的数据充分性,进而诱发欠拟合的风险;其次,鉴于评论文本固有的简短性特征,单纯依赖文本语义特征提取策略往往难以捕捉全面而深入的信息,导致特征表征不充分,进一步影响模型性能;最后,当前多数模型普遍面临泛化能力不足的问题,表现为在应对不同类别的虚假评论时,检测效果呈现出显著的差异性,难以保证一致的精准度。

2 DEBR-GAN 虚假评论检测方法

2.1 DEBR-GAN 模型构建

为充分挖掘评论文本深层特征,提高模型泛化能力,本文提出一种基于情感加权 BERT 和多任务对抗学习的虚假评论识别方法 DEBR-GAN。该方法整体框架如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同),模型由 2 个模块组成:第 1 个模块是融合情感语义的特征提取模块,该模块用于实现评论文本情感和语义关系的特征表示、提取与融合,获得样本特征集;第 2 个模块是基于领域对抗学习的虚假评论识别模块,该模块借用生成对抗网络的思想,使用领域对抗性训练强制生成器生成跨领域的共有特征,同时判别器学习区分不同领域的数据。图 1 左边为融合情感语义的特征提取模块的整体流程。首先,输入评论数据,进行数据预处理。其次,利用 BERT 提取评论文本语义特征,考虑到评论文本是消费者在使用消费产品后的个人感受,带有比较强的情感倾向,在提取语义特征的同时,情感特征抽取器先对评论文本进行情感分析,识别文本中的情感极性、情感强度、情感关联词以及一些附加特征等,然后进行综合分析。同时,引入情感加权策略,将情感特征与语义特征进行加权结合。通过这种方式,情感信息不仅能够与语义信息相互补充,还能通过调整情感和语义特征的权重,有效提升模型对细微情感变化的敏感度,使得虚假评论检测更加准确。随后,将融合消费者情感和语义特征的综合特征集通过一个 RNN 模块提取模型的序列依赖关系,提高模型对文本细节的敏感度。最后,将特征集输入基于 GAN 的识别模块进行训练。

计算机工程 www.ecice06.com

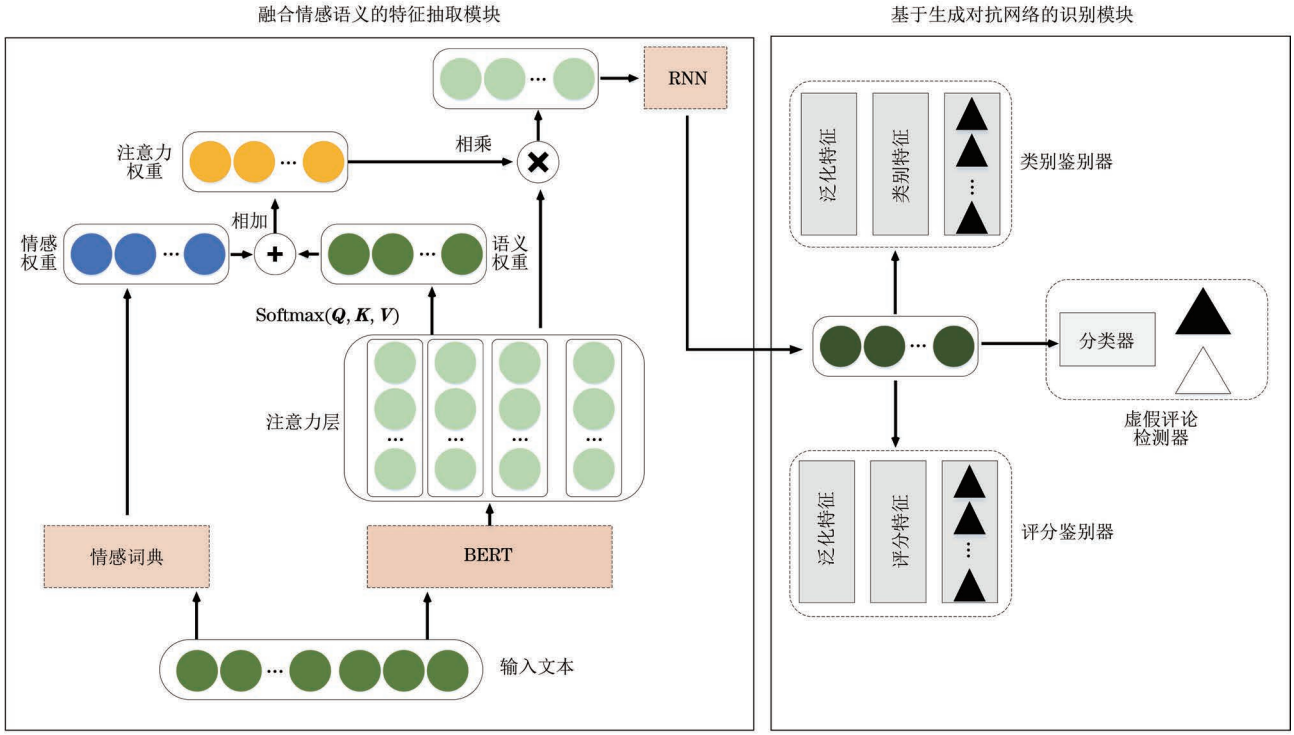


图 1 DEBR-GAN 框架

Fig. 1 Framework of DEBR-GAN

在 DEBR-GAN 模型中，为解决领域类别差异和用户评分偏好对特征提取的干扰问题，设计了一种基于多任务的对抗学习机制，如图 1 右侧所示，其核心思想是利用梯度反转技术，将特征生成器（虚假评论检测器）与两类鉴别器（类别鉴别器和评分鉴别器）构建为生成对抗网络，从而形成对抗博弈关系。生成器负责提取跨领域共享的虚假评论特征（如情感矛盾性和语义逻辑异常），类别鉴别器识别评论所属领域，评分鉴别器推断用户评分等级。在反向传播过程中，通过梯度反转层，将鉴别器的损失梯度乘以负系数（如−1）回传至生成器，迫使生成器的参数更新方向与鉴别器的分类目标相反。通过对抗训练，生成器能够“欺骗”鉴别器，从而隐式消除与领域类别和评分相关的干扰信息（如领域特有的高频词、与评分相关的情感强度），同时保留跨领域共享的判别性特征，显著提升模型的泛化能力。最终的损失函数通过加权求和得到，并通过调整权重参数找到性能最佳的平衡点。

2.2 基于情感词典的 BERT+RNN 深层词关系抽取器

2.2.1 语义编码器 BERT

预训练是指通过大规模的无监督学习的方式，将模型在大量未标记的数据上进行训练，学习到文本或者图像的共性表示，然后再将其共性迁移到另外特定领域的任务中，加入该领域少量的标注数据进行微调。由于在预训练阶段获得了丰富的知识表示，微调阶段就可以使模型很好地处理有限的目标任务，并获得不错的性能。最近几年，由于深度学习发展迅速，预训练模型在自然语言处理领域被广泛使用，尤其是基于 Transformer 架构[44]的 BERT、GPT 等预训练模型。社交网络评论攻击者往往会根据商家的不同类别，去设计一些针对性的虚假评论语言框架。例如，对于餐厅类的虚假评论，攻击者在评论时往往会夸张地表达对餐食的感受，这些问题增加了文本语义特征获取的难度。因此，本文选用目前主流的基于 Transformer 框架的预训练模型，以获取文本中的语义特征。

BERT-Base-Chinese 是 BERT 模型针对中文文本的版本，它能够学习到中文文本的语义和语法信息。BERT 模型的输入表示由 3 个部分组成，分别为词嵌入（Token Embeddings）、段嵌入（Segment Embeddings）和位置嵌入（Position Embeddings）。通过将输入 X 的字/词向量（Token Vector）、文本向量（Segment Vector）和位置向量（Position Vector）相加得到输入表示 E。

BERT 模型由 L 个 Transformer Encoder 层堆叠而成，每一层都使用自注意力机制。这种机制允许模型在处理序列时关注不同位置的 Token，并计算 Token 之间的注意力权重，从而捕捉输入序列中的依赖关系。对于第 $l(l\in(1,2,\cdots,L))$ 层，多头自

注意力机制的输出可以表示为：

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(h_{\text{head},1},\cdots,h_{\text{head},h})\boldsymbol{W}^{o} \tag{1}$$

$$h_{\text{head},i}=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{2}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V} \tag{3}$$

式中：$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$ 和 $\boldsymbol{W}^{o}$ 是可学习的参数矩阵；h 是头的数量；$d_{\boldsymbol{K}}$ 是 Key 向量的维度。

自注意力机制的输出会送入前馈神经网络进行进一步转换，以提取更高级别的特征。为了提高模型的训练稳定性和效果，每个编码器层均包含残差连接和层归一化，帮助缓解梯度消失和梯度爆炸问题，使模型更容易优化。残差连接将当前层的输入直接添加到输出中，而层归一化则对每一层的输出进行标准化处理，使得训练过程更加稳定。前馈神经网络的输出可以表示为：

$$\text{FFN}(x)=\sigma\left(\sum_{j=1}^{m}\omega_{jk}^{(2)}\sigma\left(\sum_{i=1}^{n}\omega_{ij}^{(1)}x_i+b_j^{(1)}\right)+b_k^{(2)}\right) \tag{4}$$

式中：$\omega_{ij}^{(1)}$ 是输入层到隐藏层第 j 个神经元的权重，$b_j^{(1)}$ 是该神经元的偏置；$\omega_{jk}^{(2)}$ 是隐藏层到输出层第 k 个神经元的权重，$b_k^{(2)}$ 是该神经元的偏置。

整个预训练阶段一般分为掩码语言模型 MLM 预测词语和对下一步双向句子预测模型两部分。在 MLM 任务阶段，一般会随机选择句子中的某些单词来掩码，然后根据其上下文的语义来预测这些被掩码的单词。该模型的损失函数如下：

$$l=-\sum_{i=1}^{N}\log_a P(X_i\mid X_{\text{context}}) \tag{5}$$

式中：X_i 指被掩码的字符；X_{context} 指上下文信息；$P(X_i\mid X_{\text{context}})$ 指预测概率。

2.2.2 情感编码器

BERT 预训练模型虽然具备良好的语义表征能力，但其架构设计主要针对英文语料库优化，在中文情感分析任务中存在显著局限性。真实用户的评论内容本质上是主观体验的语义投射，其文本生成过程必然伴随情感要素的嵌入。真实评论者与虚假评论者在情感表达策略上存在本质差异：前者遵循情感真实原则，通过多样化的语言形式（包括情感强度、情感极性及修辞手法）传递个体化体验；后者则表现出工具理性倾向，倾向于采用情感极化的表述策略（如夸张性形容词、重复性情感修饰等）来实现消费引导目的。情感词典包含表示感情、情绪、态度等的单词和短语。目前，典型的英文情感词典有 LIWC 情感词典[45]、SeniwordNet 词典[46] 等，中文有知网（Hownet）情感词典[47]、台湾大学（NTSUSD）简体中文情感极性词典、大连理工大学情感词汇本体[48] 等。本文从情感极性、情感词汇、情感强度和其他附加特征等方面对情感特征进行深度挖掘，最后构建情感特征抽取器。输入的评论文本序列表示为 $T=\{t_1,t_2,\cdots,t_{|T|}\}$。

1）情感极性特征。情感极性特征是一维正数或者负数，代表当前评论文本的积极或者消极倾向，本文使用 E_{pol} 表示情感极性特征得分。

2）情感词汇特征。评论文本通常会包含一些特定表达情感的词语。假设情感字典 $E=\{e_1,e_2,\cdots,e_{|E|}\}$，$e_i$ 代表情感类别，每一种情感类别 e 都包含若干表示该情感的词语 W_e，即 $e=\{W_{e,1},W_{e,2},\cdots,W_{e,|e|}\}$。另外，文本的情感表达不仅与当前情感词有关，还与其左侧若干个程度副词和否定词有关，所以通常需要对程度副词设置不同的权重，对于否定词，一般将其权重设置为 -1。最终，针对评论文本中的词汇 $t_i\in e$，得到它在情感类别 e 下的得分：

$$s(t_i,e)=\prod_{j=i-w}^{i-1}\text{neg}(t_j)*\prod_{j=i-w}^{i-1}\deg(t_j) \tag{6}$$

式中：$\text{neg}(t_j)$ 和 $\deg(t_j)$ 分别代表当前词的否定得分和程度得分。综合整条评论的所有情感词，获得这条评论文本在情感 e 上的得分：

$$s(T,e)=\sum_{i=1}^{|T|}s(t_i,e) \tag{7}$$

最后，将所有情感类别下的得分组合起来，构建为完整的情感词汇特征，将其表示为：

$$E_{\text{words}}=s(T,e_1)\oplus s(T,e_2)\oplus\cdots\oplus s(T,e_{|E|}) \tag{8}$$

3）情感强度特征。评论者针对同一种情感，由于感受程度不同，往往会采用不同的词汇来表达。如同样是表达生气的情感，“愤怒”的情感强度要高于“气愤”。本文提取情感强度特征的方法与提取情感词汇一致，将词汇 t_i 的词汇强度表示为 $\text{int}(t_i)$。计算出评论文本在情感类别 e 下的强度得分 s_{int}：

$$s_{\text{int}}(T,e)=\sum_{i=1}^{|E|}s_{\text{int}}(t_i,e) \tag{9}$$

最终，得到当前评论文本在所有情感类别下的得分：

$$E_{\text{int}}=s_{\text{int}}(T,e_1)\oplus s_{\text{int}}(T,e_2)\oplus\cdots\oplus s_{\text{int}}(T,e_{|E|}) \tag{10}$$

4）其他情感特征。将表情包、情感词、人称代词等诸多统计特征作为补充，表示为 E_{aux}。

将之前得到的所有情感特征进行组合，得到最

计算机工程 www.ecice.com

终的情感得分表示：

$$\xi=E_{\text{pol}}\oplus E_{\text{words}}\oplus E_{\text{int}}\oplus E_{\text{aux}}=\{\xi_1,\xi_2,\cdots,\xi_{|i|}\} \tag{11}$$

式中：ξ_i 为第 i 个特征。

2.2.3 联合微调

情感词典和预训练模型联合微调主要是实现一个带有情感词典的注意力层。首先，结合情感词典，计算数据集中每个词语的情感权重，将其存入权重张量中；然后，将权重张量变成 tensor 形式，再进行归一化处理使其变成情感向量。将每个词语所得的情感向量和 BERT 模型隐藏层提取出来的注意力权重向量进行加权微调，得到数据集情感加语义的权重向量；最后，将新的权重向量和隐藏层特征加权平均，输出文本的带情感加强的语义特征向量。情感微调过程如下：

$$\alpha*\text{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V}+\beta*\text{Softmax}(\xi_i) \tag{12}$$

式中：$\text{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V}$ 是文本语义特征的权重向量表示；$\text{Softmax}(\xi_i)$ 是情感特征的权重向量表示；α 和 β 分别是情感和语义微调的 2 个超参数。

2.2.4 RNN 序列特征连阶层

中文文本在一定程度上依赖特定顺序或结构，特别是在情感理解方面，不同的顺序有可能表达出不同的情感。RNN 通过记忆单元处理序列信息，能够捕捉序列依赖性，在文本分类任务中，RNN 能综合考虑文本的整体序列信息，帮助模型更好地理解句子的整体结构和语义流程。本文先使用 BERT 提取丰富的语义特征，后接 RNN 处理序列依赖关系，提高对文本细节的敏感度，可以更全面细致地理解文本内容。

RNN 的输入包括 2 个部分：一个是当前时刻的输入 x_t，用于实时更新状态；另一个是上一时刻的隐藏层状态 h_{t-1}，用于记忆状态。不同时刻的网络共用同一套参数 U、V、W。上一时刻的隐藏层和当前时刻的输入共同构成 RNN 的输入，在每一个当前时刻都能联系上一时刻，从而达到记忆序列特征的目的，公式表示为：

$$h_t=\tanh\left(W\begin{pmatrix}x_t\\h_{t-1}\end{pmatrix}\right) \tag{13}$$

式中：W 为可学习权重。

2.3 基于生成对抗网络的虚假评论识别

2.3.1 虚假评论检测器

虚假评论检测器的目标是鉴别评论文本的真假，将特征抽取器得到的特征输入到全连接层，其中激活函数采用 Softmax 函数。将全连接层得到的结果线性变换后再进行指数运算，然后对其归一化，最终得到虚假和真实的概率分布。将第 i 条评论文本表示为 T_i，则它是虚假评论的概率表示为 $P(T_i)$：

$$P(T_i)=\text{G}(F(T_i;\theta_{\text{f}});\theta_{\text{g}}) \tag{14}$$

式中：G 表示虚假评论检测器，θ_{g} 表示其参数。模型的交叉熵损失函数表示为：

$$L_{\text{g}}(\theta_{\text{f}},\theta_{\text{g}})=-\sum_{i=1}^{n}[y_i\log_a(P(T_i))+(1-y_i)\log_a(1-P(T_i))] \tag{15}$$

式中：L_{g} 表示虚假评论检测器的损失；n 表示评论的总数。参数求解公式表示为：

$$(\hat{\theta}_{\text{f}},\hat{\theta}_{\text{g}})=\underset{\theta_{\text{f}},\theta_{\text{g}}}{\text{argmin}}\,L_{\text{g}}(\theta_{\text{f}},\theta_{\text{g}}) \tag{16}$$

对于不同类别的评论文本以及不同的用户评分偏好，模型会学习到不同的文本特征，但模型学习到的特征只能用于特定评论类别和用户偏好，不具有泛化性能。因此，需要设计一个能够在所有类别和用户偏好上通用的泛模型。

2.3.2 评论类别鉴别器

本文采用评论类别鉴别器得到评论文本的类别，学习每种类别虚假评论的共同特征，使得模型可以扩展为多类别共用的虚假评论检测模型。该类别鉴别器由 3 层全连接神经网络构成，使用交叉熵损失函数计算类别鉴别器的损失 L_{c}：

$$L_{\text{c}}(\theta_{\text{f}},\theta_{\text{c}})=-\sum_{i=1}^{n}\sum_{j=1}^{J}1_{[j=y_{\text{c}}]}\log_a(\text{C}(F(T_i;\theta_{\text{f}}));\theta_{\text{c}}) \tag{17}$$

这里可以得到：

$$\hat{\theta}_{\text{c}}=\underset{\theta_{\text{c}}}{\text{argmin}}\,L_{\text{c}}(\theta_{\text{f}},\theta_{\text{c}}) \tag{18}$$

式中：C 表示评论类别鉴别器，θ_{c} 表示其参数；J 为评论文本类别数；y_{c} 为评论类别标签。

当损失较大时，认为预训练模型抽取得到的文本特征包含较少的类别信息，更具有泛化性能；反之，则表示当前学习到的评论文本特征类别性较强，不利于模型扩展，从而影响多类别虚假评论检测的性能。因此，需要找到合适的参数 θ_{c}，使得式(17)中的损失尽可能最大化，从而训练一个适用于多类别评论文本的虚假评论检测器。

2.3.3 用户评分偏好鉴别器

虚假评论用户一般出于某种商业利益，会恶意给一些商家极高或者极低的评分，不同评分情况下的评论文本差别较大，如某一条评论的分数是 5 分，但是评论文本会有明显对产品或者服务不满意的词

计算机工程 www.ecice.com

语。在此种情况下，若评论分数占主导地位，则有可能是好评，若评论文本占主导地位，则有可能是差评。因此，在判断是否为虚假评论时，还需要考虑评论分数和文本的匹配情况：若不对它们进行区分，则学习到的特征会注重用户评分偏好，而很难捕捉到针对虚假评论的独有特征；如果对它们分别构建模型，又会使模型结构复杂化。为了学习到不同评分偏好下的评论文本特征，本文使用 3 层全连接层神经网络构建评分偏好鉴别器，将评论文本的特征输入该鉴别器中，最终得到评论文本在各个评分上的概率分布。该鉴别器的损失函数如下：

$$L_s(\theta_f,\theta_s)=$$
$$-\sum_{i=1}^{n}\sum_{k=1}^{K}1_{[k=y_s]}\log_a(S(F(T_i;\theta_f));\theta_s) \tag{19}$$

进而可以得出：

$$\hat{\theta}_s=\underset{\theta_s}{\operatorname{argmin}}\,\theta_s L_s(\theta_f,\theta_s) \tag{20}$$

式中：S 代表用户评分偏好鉴别器，θ_s 为其参数；J 为评分类别数；y_s 为评分标签。若 L_s 值过小，则表示当前模型学习到的评论文本特征中包含较多用户评分偏好信息。通过调节参数 θ_s 使用户评分偏好损失最大化，进一步获得更加泛化的评论文本特征，再结合虚假评论检测器和评论类别鉴别器，训练出更具有普适性的虚假评论检测模型。

2.3.4 模块组合

本文借用生成对抗网络的思想，设法在两类鉴别器和虚假评论检测器之间获取平衡，从而将原本去除类别特征和用户评分偏好特征的虚假评论检测模型这一任务，转化为上文中类别鉴别器损失和虚假评论检测器损失之间的博弈。将两类鉴别器的损失进行梯度反转，然后在反向传播中与虚假评论检测器的损失作差，这样在使两类鉴别器的损失最大化的同时，又保证了虚假评论检测器的准确性。最终得到本文模型的损失函数为：

$$L_{au}(\theta_f,\theta_g,\theta_c,\theta_s)=$$
$$L_g(\theta_f,\theta_g)-\mu L_c(\theta_f,\theta_c)-\theta L_s(\theta_f,\theta_s) \tag{21}$$

参数 μ 和 θ 的作用是调节 3 个任务的重要程度，μ 和 θ 越小，则分别代表两类鉴别器的任务权重越小，越关注虚假评论检测器的结果；反之，若 μ 和 θ 越大，则代表评论类别和评分偏好的鉴别任务越重要。这里极大极小的博弈参数，需要通过求解目标函数的鞍点得到：

$$(\hat{\theta}_f,\hat{\theta}_g)=\underset{\theta_f,\theta_g}{\operatorname{argmin}}(\theta_f,\theta_g,\hat{\theta}_c,\hat{\theta}_s) \tag{22}$$

$$\hat{\theta}_c=\underset{\theta_c}{\operatorname{argmin}}\,L_{all}(\hat{\theta}_f,\hat{\theta}_g,\theta_c,\hat{\theta}_s) \tag{23}$$

$$\hat{\theta}_s=\underset{\theta_s}{\operatorname{argmin}}\,L_{all}(\hat{\theta}_f,\hat{\theta}_g,\hat{\theta}_c,\theta_s) \tag{24}$$

采用随机梯度下降的方法得到模型的最优解，固定学习率 r 后，参数的更新过程如下所示：

$$\theta_f:=\theta_f-r\left(\frac{\partial L_g}{\partial\theta_f}-\mu\frac{\partial L_c}{\partial\theta_f}-\theta\frac{\partial L_s}{\partial\theta_f}\right) \tag{25}$$

$$\theta_c:=\theta_c-r\frac{\partial L_c}{\partial\theta_f} \tag{26}$$

$$\theta_s:=\theta_s-r\frac{\partial L_s}{\partial\theta_s} \tag{27}$$

$$\theta_g:=\theta_g-r\frac{\partial L_g}{\partial\theta_g} \tag{28}$$

3 实验设计与结果分析

3.1 数据集

本文的评论数据集主要从大众点评网站的店铺爬取，为平衡数据需求并最小化其他外部环境对模型的影响，本文将数据采集周期设定为 2015 年 6 月 1 日—2024 年 1 月 1 日，最终收集到 11 221 位用户的 485 162 条评论。用于构建特征集的数据集包含用户个人信息和商家信息 2 个部分，包括用户 ID、用户名、评论数量、注册时间、评论时间、用户评分以及店铺 ID、店铺名称、店铺类型、店铺评分、评论数量、人均消费等。从原始文本中剔除无关评论，最后剩下 359 079 条评论信息。

对比数据集是带有标签的、公开的 Yelp 数据集（https://business.yelp.com/data/resources/open-dataset/），共有 6 990 280 条评论，涉及 150 346 个商家，11 个领域，包括评论日期、评论 ID、用户 ID、商家/产品 ID、标签（N 表示真实评论，Y 表示虚假评论）、星级评分等。

3.2 数据处理

1）数据预处理及标注。平台通过隐藏机制对部分评价实施过滤处理，此类评价不会展示在商户页面或用户主页。通过爬取平台显性评价与隐式过滤评价构建对比语料库。当评价存在于本地数据集但缺失于平台公开列表时，即可判定该评价已被平台隐藏。为建立高信度标注体系，构建跨学科专家协同标注框架，具体是联合 2 名老师、5 名硕博士及 256 名平台活跃用户，最后从 2 个维度制定虚假评论标注规则，具体如下：

（1）评价内容与商户特征的语义关联性：情绪极性偏离常态分布（如极端愤怒、亢奋表达）；专业术语密度异常（商品参数过度描述且缺乏消费情境细节）；诱导性言语行为标记（含“绝对值得”等强劝告句式及隐性利益诱导表述）。

（2）用户行为时间序列异常：相同 ID 短时间内

计算机工程 www.ecice06.com

出现频率较高，并发布多条极端好评或差评，发布时间与消费时间不符；直接复制或高度相似于其他用户的评价；群体评论文本具有逻辑互斥性（同商户评价情感极性相反）。

本文标注过程采用文献[49]中的改进多数表决机制，构建 7 人标注组进行同步验证。当真实票数大于等于 5 时判定为真实评论，否则归为非真实性评论。经计算，组内 Cohen's κ 系数达 0.75，满足社会计算研究标注一致性要求。

2）数据集展示。针对原始评论文本部分类别稀疏及特征提取需求，本文对语义相近的细粒度类别进行整合（如家装与车辆服务统一归并为生活服务类）。如表 1 所示，经调整后的 8 个主要类别呈现显著的不均衡分布，其中餐厅类评论文本量占比达 91.7%，构成典型的长尾分布。

统计用户评分比例如表 2 所示，评分出现分布不均匀的情况，即“4 分”和“5 分”的评论较多，说明用户评价行为存在显著的偏态特征，因此，在训练模型时需要考虑用户的喜好问题。

对 8 类评论文本构建图 2 所示的词云，不同类别的评论文本之间存在较大差异，餐厅类评论文本关注点较多集中在对食材和食物的感受，美容美发类则更注重服务体验和服务环境，宠物类除了包含领域特有名词外，还会有一些互动性的词语。

表 1 评论文本类别分布情况

Table 1 Distribution of comment text categories

单位：个

类别	数量
演出	3 824
餐厅	329 383
生活服务	993
美容美发	7 185
娱乐	8 302
购物	2 638
酒店	4 132
宠物	2 622

表 2 评分比例

Table 2 Rating ratio

分数	比例/%
5	40.86
4	36.23
3	16.15
2	3.82
1	2.94



图 2 不同评论类别的词云

Fig. 2 Word cloud for different comment categories

为进一步分析不同领域的情绪变化情况，根据文献[50]，本文采用情感词汇本体库[51]，先将真实文本和虚假文本拼接在一起进行归一化处理，然后分别对真实和虚假评论内容中的 7 种情绪进行分析，包括“乐”“好”“怒”“哀”“惧”“恶”“惊”。不同领域的评论情绪特征如图 3 和图 4 所示。真实评论的情感随着领域变化显得均匀和平滑，酒店类和娱乐类情感倾向会更加明显，而生活服务类情感更趋向于平和。虚假评论随不同领域的情绪波动比真实评论显著，这是因为虚假评论往往夸大负面情绪，或是表现出保守和缺乏深度的倾向。如虚假评论常忽略恐惧等情感，可能是因为它们倾向于塑造特定形象，而忽略用户面对风险或不确定性时的真实情感体验。

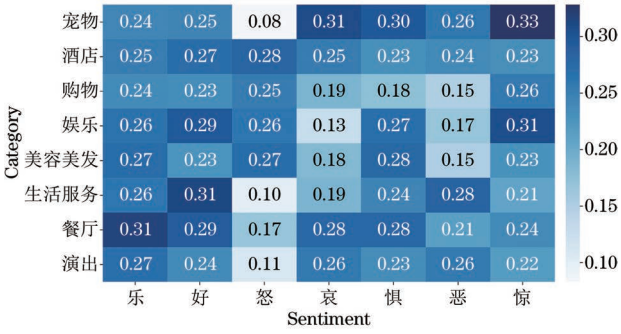


图 3 虚假评论情感热力图

Fig. 3 False comment sentiment heatmap

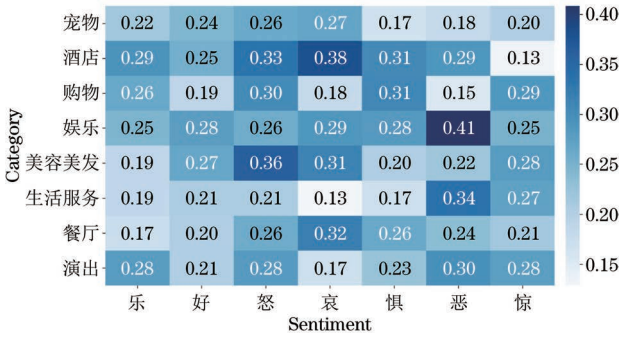


图 4 真实评论情感热力图

Fig. 4 Real comment sentiment heatmap

3.3 评价指标与实验设置

3.3.1 数据集划分

对未处理过的原始文本进行分析，发现不超过 30 个字符的评论文本占比最高，近乎一半的评论文本都在 60 个字符以下，而超过 120 个字符的文本仅仅只占总样本的 30%左右。选取训练样本时，要同时考虑不同类别评论样本的平衡问题和用户评分分布不均匀的影响。受到总样本数据集过大且很多评论文本质量不高的影响，本文在保证评分和类别样本数量均衡的情况下，从数据集中选取了质量和标注一致性都较高的 2 000 条真实评论和 2 000 条虚假评论作为样本数据集，其中 8 个类别中每种类别的真实评论和虚假评论都为 250 条，且 5 个评分中每个评分的真实评论和虚假评论均为 400 条。从真实评论和虚假评论中随机提取 70%的样本数据作为训练集，15%作为验证集，剩下 15%作为测试集。对比数据集的选取和划分标准与大众点评数据集一致。

3.3.2 评价指标

本文选取二分类任务中常用的基础指标，即准确率(Accuracy)、F1 值(F1)和概率指标 AUC。其中，准确率是预测正确样本的比例，F1 值是精确率和召回率的调和平均值，AUC 是 ROC 曲线下面积，反映模型在所有阈值下区分正负类的能力。每组实验进行 5 次并取平均值，避免数据集划分的随机性对模型性能产生影响。

3.3.3 实验设置

本文采用专门针对中文语言开发的预训练模型 BERT-Base-Chinese，包括 12 层 Transformer，每一层由前馈神经网络和多头自注意力机制 2 个子层组成，每个子层之间采用残差连接和层归一化技术，整个模型拥有约 1.1 亿的参数量。输入为文本序列，每一个序列的输出为一个 768 维的向量，用来表示该文本的语义特征。

DEBR-GAN 是在配备了 Intel Xeon W-2123 3.9 GHz CPU 和 64 GB RAM 的服务器上进行评估的。GPU 为 NVIDIA Tesla v100，CUDA 版本 10.2，基于 Python 3.8 和 PyTorch 1.10.1 来构建模型。模型参数设置如表 3 所示，在该参数下，一组实验大致需要 1 518 s。

表 3 模型参数设置

Table 3 Model parameters setting

参数	值
学习率	0.000 3
训练周期	100
Batch size	16
注意力层数	12
隐藏层维度	768
最大评论长度/字	170

3.4 实验结果分析

3.4.1 消融实验

为验证 DEBR-GAN 模型的有效性，进行 5 组消融实验，第 1 组～第 5 组实验分别去掉情感词典、RNN 模型、评分鉴别器、类别鉴别器，类别鉴别器和评分鉴别器，模型其他部分不变，验证去掉模块是否对虚假评论识别效果有影响。5 组实验结果如图 5 所示。

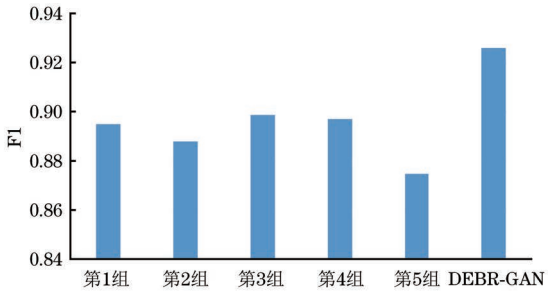


图 5 消融实验结果

Fig. 5 Results of ablation experiment

观察 5 组实验与本文 DEBR-GAN 模型结果可知，DEBR-GAN 模型的 F1 值均高于其他模型，达到 0.926，即使将两类鉴别器都去掉(第 5 组)，仅采用 BERT 结合情感词典对语义特征进行情感增强，F1 值也能达到 0.875。在加入评论类别鉴别器和用户偏好鉴别器后，两者都对检测模型的性能提升有所帮助，其中评论类别鉴别器效果更好。因为在特征抽取阶段，类别较多且类别之间差异明显，其中包含的评论类别信息多于评分信息。在前 2 组的特征抽取消融实验中，会发现去掉 RNN 模型对特征抽取影响较大，这说明中文文本在一定程度上依赖特定顺序或结构。

计算机工程 www.ccjeo6.com

3.4.2 对比实验

根据模型类型，本文将对比实验分为 2 组：第 1 组提取不同的文本特征，如用户行为、用户偏好等，用于识别虚假评论，如文献[52]方法采用传统特征与时空特征融合策略，文献[4]方法构建用户逻辑特征、用户偏好特征与传统特征的联合表征，文献[49，53-54]方法构建文本和图像融合特征；第 2 组采用 BiGRU[55]、TextCNN[56]、BERT 及其变体 RoBERTa[57]、MacBERT[58]等深度学习模型进行虚假评论识别。

从表 4 可以看出(最优结果加粗标注)，DEBR-GAN 模型在大众点评数据集上表现出了显著的优势，在 AUC、F1 值和准确率上，DEBR-GAN 均优于其他主流模型，尤其在 AUC 和 F1 值上分别达到了 0.976 8 和 0.926 0，分别领先于其他模型 3～5 百分点。在第 1 组实验中，DEBR-GAN 的 AUC 值显著高于其他两组，表明模型在整体性能上更有优越性，尤其是在虚假评论检测领域。通过融合情感、时序信息等多维特征表征，DEBR-GAN 显著提升了对虚假评论的检测准确性，这也验证了不同特征的联合建模在增强模型识别能力方面的重要性，特别是情感信息与时序依赖的结合，在捕捉虚假评论的微妙情绪变化方面发挥了关键作用。

表 4 大众点评数据集上的对比实验结果

Table 4 Comparative experimental results on the Dianping dataset

方法	AUC	Accuracy	F1
文献[52]方法	0.931 5	0.777 3	0.727 0
文献[4]方法	0.931 8	0.780 0	0.731 3
文献[53]方法	0.672 8	0.872 0	0.700 8
文献[54]方法	0.684 9	0.872 7	0.701 0
文献[49]方法	0.889 8	0.863 6	0.719 7
RNN	0.668 1	0.795 0	0.647 6
BERT	0.634 2	0.634 2	0.652 6
文献[55]方法	0.693 5	0.872 0	0.701 8
文献[56]方法	0.678 3	0.874 8	0.702 8
文献[57]方法	0.901 3	0.892 6	0.890 9
文献[58]方法	0.687 8	0.869 8	0.692 8
DEBR-GAN	**0.976 8**	**0.929 0**	**0.926 0**

在第 2 组实验中，尽管 RoBERTa 等深度学习模型普遍表现较好，但 DEBR-GAN 在所有评估指标上均展现了优异的性能，特别是在处理复杂评论内容时表现尤为突出。相比之下，尽管 RoBERTa 的整体性能较强，但它并未考虑情感信息和语义关系的深度融合，也未能有效克服模型在跨领域应用中的偏好偏差问题。DEBR-GAN 通过情感-语义耦合的特征提取阶段，充分发挥了情感特征的引导作用，同时借助 GAN 的对抗学习机制，成功消除了由用户偏好和领域特异性信息所引发的偏差，从而确保了在多个领域和情境下的高准确性和泛化能力。

从表 5 可以看出，DEBR-GAN 在 Yelp 数据集上的 F1 值为 0.818 5，表现突出。与其他模型相比，DEBR-GAN 在精确率和召回率的平衡上取得了显著优势，特别是相较于 BERT(0.779 4)和文献[56]方法(0.748 7)，展现了更强的虚假评论识别能力，能够有效减少漏检和误检，从而提升了检测的可靠性。此外，DEBR-GAN 在 Yelp 数据集上的各项指标普遍优于大多数现有方法，尤其在 F1 值和准确率上表现突出，表明其在处理不同情感倾向和复杂语义结构的评论时，仍能保持较高的识别准确性。综上所述，DEBR-GAN 在 Yelp 数据集上的优异表现验证了其在多个数据集中的强大适应性和鲁棒性，并展现了在不同场景下的广泛应用潜力。

表 5 Yelp 数据集上的对比实验结果

Table 5 Comparative experimental results on the Yelp dataset

方法	AUC	Accuracy	F1
文献[52]方法	0.875 3	0.808 3	0.802 7
文献[4]方法	0.840 4	0.761 3	0.760 4
文献[53]方法	0.842 0	0.761 1	0.757 2
文献[54]方法	0.855 0	0.776 3	0.773 8
文献[49]方法	**0.896 9**	0.817 8	0.817 5
RNN	0.864 3	0.756 7	0.757 2
BERT	0.860 3	0.785 0	0.779 4
文献[55]方法	0.833 9	0.757 0	0.750 4
文献[56]方法	0.824 5	0.750 1	0.748 7
文献[57]方法	0.841 1	0.819 7	0.770 3
文献[58]方法	0.848 8	0.769 4	0.766 1
DEBR-GAN	0.886 5	**0.822 1**	**0.818 5**

3.4.3 语义情感权重实验

DEBR-GAN 在特征提取阶段考虑了情感和语义双重特征，通过对情感和语义的权重进行微调，提高模型的识别率，如式(12)所示。本文对语义权重 α 和情感权重 β 设置不同的值进行验证，找到最合适的权重。实验分为两组：a 组固定情感权重 $\beta=1$，调节不同的语义权重 α，结果如图 6(a)所示；b 组固定语义权重 $\alpha=1$，调节不同的情感权重 β，结果如图 6(b)所示。从图 6 可以看出，两组实验整体表现出相同趋势，即固定其中一个权重，模型性能皆会随着另一个权重参数的增大而先提升后降低，并且固定一个参数为 1，当另一个参数达到 1.2 时，模型达到最佳性能。例如，a 组实验固定情感权重 $\beta=1$，当

www.ecice.com

情感权重 β 大于语义权重 α，语义表现不敏感，这是因为情感特征占主导地位，语义特征被弱化，不能体现语义特征的作用。当语义权重 $\alpha=1.2$ 时，模型达到了最佳性能，并且随着语义权重的逐渐增大，模型整体性能下降明显，主要因为语义权重越大，占主导地位越强，情感特征会表现得更不明显，模型性能会明显下降。b 组实验结果反之亦然。由实验结果可知，情感和语义权重仅能起到微调作用，在虚假评论文本里，语义特征能分析文本语境关系和句法结构等，但是评论是消费者使用后的感受，带有个人情感色彩，因而只有情感和语义特征相结合，才能使模型性能达到最优。

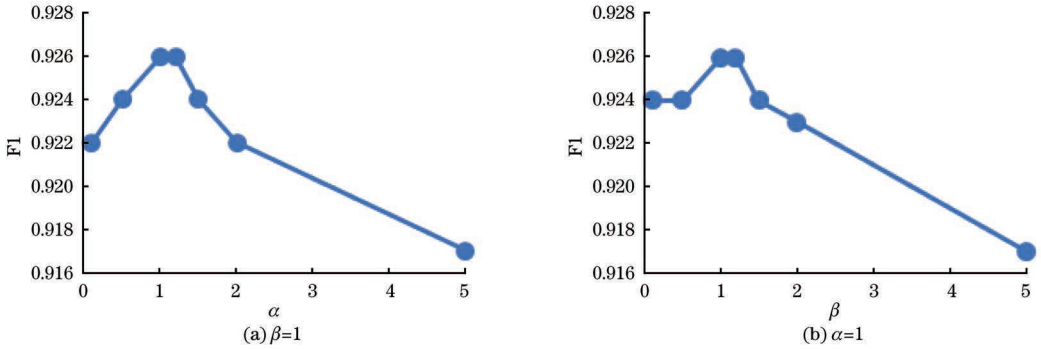


图 6 模型效率随语义和情感权重的变化情况

Fig. 6 The variation of model efficiency with semantic and emotional weights

3.4.4 对抗实验

在式(21)所示的损失函数中，参数 μ 和 θ 反映评论类别和用户评分偏好的重要性，μ 值越大，则代表模型中评论类别鉴别器的任务重要程度越高，θ 值越大，则表示用户评分鉴别器的任务更为重要。若两者都较小，则说明虚假评论检测器占据主导地位，μ 和 θ 的值将在很大程度上决定模型的优劣。图 7(a)和图 7(b)分别为固定 θ 和 μ 值为 1 时模型性能随 μ 和 θ 值的变化情况。当 $\mu=0$ 或者 $\theta=0$ 时，相当于去掉评论类别鉴别器或者评分类别鉴别器，会看到模型性能明显低于有评论类别鉴别器或评分类别鉴别器时的模型效果。

从图 7(a)实验可以看出，当固定用户评分偏好参数 $\theta=1$ 时，模型性能随着评论类别参数 μ 的增加先增强后减弱，当 $\mu=1$ 时，模型性能达到最佳。反之，固定评论类别参数 $\mu=1$ 时，模型性能也随着用户评分偏好参数 θ 的增加先增强后减弱。在 $\mu=1$ 和 $\theta=1$ 时，模型在数据集上的检测效果都达到最佳。因为伴随着另外一个参数继续增大，模型学习到的特征更加泛化，这也导致虚假评论检测任务的比重持续减小，造成了模型性能的损失，模型效果逐渐降低。因此，可以确定在 $\mu=1$ 和 $\theta=1$ 时，特征提取器、两类鉴别器和虚假评论检测器完成了相互之间的博弈，得到去除文本评论类别和用户评分偏好影响的泛化特征，达到了最佳的虚假评论检测结果。

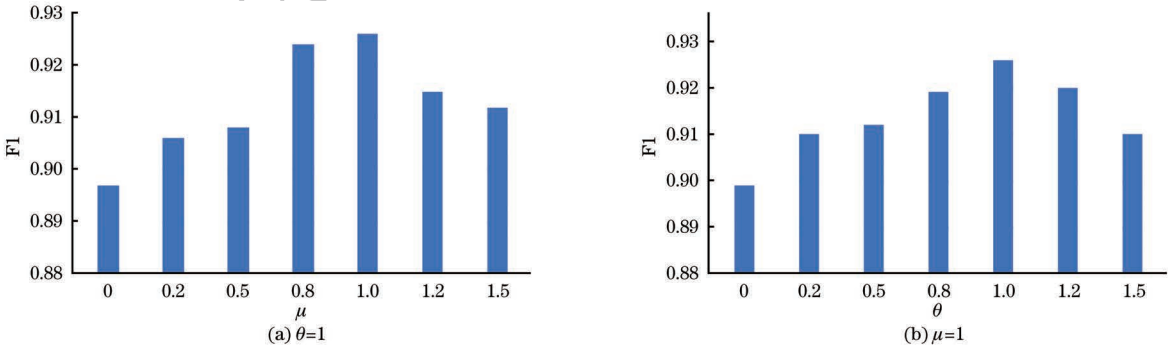


图 7 模型效率随 μ 和 θ 权重的变化情况

Fig. 7 The variation of model efficiency with the weights of μ and θ

4 结束语

本文针对社交网络虚假评论检测中存在的语义表征不充分、模型泛化能力不足的双重挑战，构建一种基于多维度信息融合的虚假评论检测框架 DEBR-GAN。在特征表征层面，创新性地提出情感-语义双通道增强预训练策略，通过领域自适应微调和注意力权重优化，实现了评论文本深层情感特征与上下文语义特征的协同提取。设计一种基于生成对抗网络的双偏好对抗机制，通过评分鉴别器与

类别判别器的动态博弈，有效缓解由跨平台数据分布漂移导致的模型性能衰减问题。实验结果表明，本框架在语义理解深度、特征融合效度及领域适应能力等维度均显著优于基准模型。

未来将从以下 3 个方面继续进行深入研究：第一，在数据层面构建跨模态动态增强数据集，通过迁移学习融合用户画像、社交图谱等辅助信息，结合主动学习策略缓解单一类型样本困境；第二，本文特征工程都是根据规则设计的，其优点是具有较强的可解释性，随着大语言模型的广泛应用，其在诸多任务上表现出了不错的性能，未来考虑结合大语言模型来提取特征；第三，本文应用场景局限于单一领域（社交媒体评论），接下来考虑将社交媒体评论虚假识别迁移到新闻类领域，在新闻类公开数据集和自制数据集上实现模型的跨场景识别。

参考文献

[1] 杨铭，祁巍，闫相斌，等. 在线商品评论的效用分析研究[J]. 管理科学学报，2012(5)：65-75.
YANG M，QI W，YAN X B，et al. Utility analysis for online product review[J]. Journal of Management Sciences in China，2012(5)：65-75.（in Chinese）

[2] 钟科，李佩锴，马士伟. 游客在线评论中嗅觉文字线索的价值[J]. 旅游科学，2019，33(6)：17-32.
ZHONG K，LI P M，MA S W. The value of olfactory text cues in visitors' online reviews[J]. Tourism Science，2019，33(6)：17-32.（in Chinese）

[3] 施运梅，袁博，张乐，等. IMTS：融合图像与文本语义的虚假评论检测方法[J]. 数据分析与知识发现，2022，6(8)：84-96.
SHI Y M，YUAN B，ZHANG L，et al. IMTS：detecting fake reviews with image and text semantics[J]. Data Analysis and Knowledge Discovery，2022，6(8)：84-96.（in Chinese）

[4] QU Z，LÜ C，CHI C H. MUSH：multi-stimuli hawkes process based sybil attacker detector for user-review social networks[J]. IEEE Transactions on Network and Service Management，2022，19(4)：4600-4614.

[5] 李璐旸，秦兵，刘挺. 虚假评论检测研究综述[J]. 计算机学报，2018，41(4)：946-968.
LI L Y，QIN B，LIU T. Survey on fake review detection research[J]. Chinese Journal of Computers，2018，41(4)：946-968.（in Chinese）

[6] LEE M，JEONG M，LEE J. Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website：a text mining approach[J]. International Journal of Contemporary Hospitality Management，2017，29(2)：762-783.

[7] YE J T，KUMAR S，AKOGLU L. Temporal opinion spam detection by multivariate indicative signals[J]. Proceedings of the International AAAI Conference on Web and Social Media，2016，10(1)：743-746.

[8] LIU A X，XIE Y，ZHANG J R. It's not just what you say，but how you say it：the effect of language style matching on perceived quality of consumer reviews[J]. Journal of Interactive Marketing，2019，46：70-86.

[9] KIM S M，PANTEL P，CHKLOVSKI T，et al. Automatically assessing review helpfulness[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. [S.l.]：ACL，2006：423.

[10] ANDERSON E，SIMESTER D. Deceptive reviews：the influential tail[EB/OL]. [2025-01-05]. https://www.semanticscholar.org/paper/Deceptive-Reviews-:-The-Influential-Tail-Simester/9061fb46185cffde44168aff4eb17f25d520f93a/figure/0.

[11] ZHU L，LIN Y，CHENG M M. Sentiment and guest satisfaction with peer-to-peer accommodation：when are online ratings more trustworthy?[J]. International Journal of Hospitality Management，2020，86：102369.

[12] 鄢慧丽，沈宸欣，熊浩. 反讽对评论有用性的机制研究：以酒店在线评论为情境[J/OL]. 旅游科学：1-23. [2025-01-05]. https://link.cnki.net/doi/10.16323/j.cnki.lykx.20240913.001.
YAN H L，SHEN C X，XIONG H. Research on mechanism of irony on review helpfulness：a scenario of hotel online reviews[J/OL]. Tourism Science：1-23. [2025-01-05]. https://link.cnki.net/doi/10.16323/j.cnki.lykx.20240913.001.（in Chinese）

[13] 曹蓓，赵奎. 基于双重情感和多特征融合的虚假新闻检测[J]. 计算机工程，2025，51(6)：193-203.
CAO B，ZHAO K. Dual emotion and multi-feature fusion based fake news detection[J]. Computer Engineering，2025，51(6)：193-203.（in Chinese）

[14] TRUEMAN T E，KUMAR A，NARAYANASAMY P，et al. Attention-based C-BiLSTM for fake news detection[J]. Applied Soft Computing，2021，110：107600.

[15] CHIN J Y，ZHAO K Q，JOTY S，et al. ANR：aspect-based neural recommender[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York，USA：ACM Press，2018：147-156.

[16] MCAULEY J，LESKOVEC J. Hidden factors and hidden topics：understanding rating dimensions with review text[C]//Proceedings of the 7th ACM Conference on Recommender Systems. New York，USA：ACM Press，2013：165-172.

[17] XU F，SHENG V S，WANG M W. Near real-time topic-driven rumor detection in source microblogs[J]. Knowledge-Based Systems，2020，207：106391.

[18] 王根生，朱奕，李胜. 一种融合知识图谱的图注意力神经网络谣言实时检测方法[J]. 数据分析与知识发现，2024，8(6)：95-106.
WANG G S，ZHU Y，LI S. A real-time rumor detection method based on the graph attention neural network integrated with the knowledge graph[J]. Data Analysis and Knowledge Discovery，2024，8(6)：95-106.（in Chinese）

[19] 黄学坚，王根生，罗远胜，等. 融合多元用户特征和内容特征的微博谣言实时检测模型[J]. 小型微型计算机系统，2022，43(12)：2518-2527.
HUANG X J，WANG G S，LUO Y S，et al. Weibo rumors real-time detection model based on fusion of multi user features and content features[J]. Journal of Chinese Computer Systems，2022，43(12)：2518-2527.（in Chinese）

[20] 郭贤伟，赖华，余正涛，等. 融合情绪知识的案件微博评论情绪分类[J]. 计算机学报，2021，44(3)：564-578.
GUO X W，LAI H，YU Z T，et al. Emotion classification of case-related microblog comments integrating emotional knowledge[J]. Chinese Journal of Computers，2021，44(3)：564-578.（in Chinese）

[21] 张换香，彭俊杰. 基于方面级情感分析的深度语义挖掘模型[J]. 电子学报，2024，52(7)：2307-2319.
ZHANG H X，PENG J J. A deep semantic mining model based on aspect-level sentiment analysis[J]. Acta Electronica Sinica，2024，52(7)：2307-2319.（in Chinese）

[22] 曾雪强，华鑫，刘平生，等. 基于情感轮和情感词典的文本情感分布标记增强方法[J]. 计算机学报，2021，44(6)：1080-1094.

计算机工程 www.ecice.com

ZENG X Q, HUA X, LIU P S, et al. Emotion wheel and lexicon based text emotion distribution label enhancement method[J]. Chinese Journal of Computers, 2021, 44(6): 1080-1094. (in Chinese)

[23] MENG Y, ZHANG Y Y, HUANG J X, et al. Text classification using label names only: a language model self-training approach[EB/OL]. [2025-01-05]. https://arxiv.org/abs/2010.07245.

[24] 程艳, 尧磊波, 张光河, 等. 基于注意力机制的多通道 CNN 和 BiGRU 的文本情感倾向性分析[J]. 计算机研究与发展, 2020, 57(12): 2583-2595.
CHENG Y, YAO L B, ZHANG G H, et al. Text sentiment orientation analysis of multi-channels CNN and BiGRU based on attention mechanism[J]. Journal of Computer Research and Development, 2020, 57(12): 2583-2595. (in Chinese)

[25] OTT M, CHOI Y, CARDIE C, et al. Finding deceptive opinion spam by any stretch of the imagination[EB/OL]. [2025-01-05]. https://arxiv.org/abs/1107.4557.

[26] DENG X L, CHEN R Y. Sentiment analysis based online restaurants fake reviews hype detection[EB/OL]. [2025-01-05]. https://link.springer.com/chapter/10.1007/978-3-319-11119-3_1.

[27] 赵军, 王红. 融合情感极性和逻辑回归的虚假评论检测方法[J]. 智能系统学报, 2016, 11(3): 336-342.
ZHAO J, WANG H. Detection of fake reviews based on emotional orientation and logistic regression[J]. CAAI Transactions on Intelligent Systems, 2016, 11(3): 336-342. (in Chinese)

[28] WU Z Y, PI D C, CHEN J F, et al. Rumor detection based on propagation graph neural network with attention mechanism[J]. Expert Systems with Applications, 2020, 158: 113595.

[29] 陈妍, 张小威, 金赞, 等. 基于加权 GraphSAGE 和生成对抗网络的医保欺诈识别方法[J]. 系统工程理论与实践, 2024, 44(2): 732-751.
CHEN Y, ZHANG X W, JIN Z, et al. Medical fraud detection method based on weighted GraphSAGE and generative adversarial network[J]. Systems Engineering-Theory & Practice, 2024, 44(2): 732-751. (in Chinese)

[30] 任亚峰, 尹兰, 姬东鸿. 基于语言结构和情感极性的虚假评论识别[J]. 计算机科学与探索, 2014, 8(3): 313-320.
REN Y F, YIN L, JI D H. Deceptive reviews detection based on language structure and sentiment polarity[J]. Journal of Frontiers of Computer Science & Technology, 2014, 8(3): 313-320. (in Chinese)

[31] FAN C, YAN H Y, DU J C, et al. A knowledge regularized hierarchical approach for emotion cause analysis[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.]: ACL, 2019: 5613-5623.

[32] SHIN B, LEE T, CHOI J D. Lexicon integrated CNN models with attention for sentiment analysis[EB/OL]. [2025-01-05]. https://arxiv.org/abs/1610.06272.

[33] JINDAL N, LIU B. Opinion spam and analysis[C]// Proceedings of the International Conference on Web Search and Web Data Mining. New York, USA: ACM Press, 2008: 219.

[34] LI X L, YU P S, LIU B, et al. Positive unlabeled learning for data stream classification[C]//Proceedings of the 2009 SIAM International Conference on Data Mining. New York, USA: ACM Press, 2009: 259-270.

[35] 杜姗, 杨敏, 仇蓉蓉. 基于 SMOTE-RF 与多维特征向量的在线商品虚假评论识别研究[J]. 情报杂志, 2023, 42(4): 156-164.
DU S, YANG M, QIU R R. Research on recognition of fake online product reviews based on SMOTE-RF and multidimensional feature vector[J]. Journal of Intelligence, 2023, 42(4): 156-164. (in Chinese)

[36] 张文, 王强, 步超骐, 等. 基于 Co-training 协同训练的在线虚假评论识别研究[J]. 系统工程理论与实践, 2020, 40(10): 2669-2683.
ZHANG W, WANG Q, BU C Q, et al. A study on deceptive review identification based on Co-training[J]. Systems Engineering-Theory & Practice, 2020, 40(10): 2669-2683. (in Chinese)

[37] 王一杰, 崔彩霞. 基于代价敏感图卷积网络的虚假评论检测研究[J]. 太原师范学院学报(自然科学版), 2023, 22(4): 37-42.
WANG Y J, CUI C X. Research on false review detection based on cost-sensitive graph convolutional network[J]. Journal of Taiyuan Normal University (Natural Science Edition), 2023, 22(4): 37-42. (in Chinese)

[38] XU Y Z, LI Q. Attention-based feature fusion network for fake reviews detection[C]// Proceedings of the 3rd International Conference on Artificial Intelligence and Advanced Manufacture. New York, USA: ACM Press, 2022: 666-671.

[39] MOHAWESH R, XU S X, TRAN S N, et al. Fake reviews detection: a survey[J]. IEEE Access, 2021, 9: 65771-65802.

[40] GOLDANI M H, MOMTAZI S, SAFABAKHSH R. Detecting fake news with capsule neural networks[J]. Applied Soft Computing, 2021, 101: 106991.

[41] MA J, GAO W, WONG K F. Detect rumors on Twitter by promoting information campaigns with generative adversarial learning[C]// Proceedings of the World Wide Web Conference. New York, USA: ACM Press, 2019: 3049-3055.

[42] 张洪志, 但志平, 董方敏, 等. 基于循环生成对抗网络和 Wasserstein 损失的谣言检测研究[J]. 数据分析与知识发现, 2024, 8(7): 32-43.
ZHANG H Z, DAN Z P, DONG F M, et al. Detecting rumors based on CycleGAN and Wasserstein loss[J]. Data Analysis and Knowledge Discovery, 2024, 8(7): 32-43. (in Chinese)

[43] CHENG M X, LI Y Z, NAZARIAN S, et al. From rumor to genetic mutation detection with explanations: a GAN approach[J]. Scientific Reports, 2021, 11: 5861.

[44] VASWANI A. Attention is all you need[EB/OL]. [2025-01-05]. https://arxiv.org/abs/1706.03762.

[45] PENNEBAKER J W, FRANCIS M E, BOOTH R J. Linguistic inquiry and word count: LIWC 2001[EB/OL]. [2025-01-05]. https://www.researchgate.net/publication/239667728_Linguistic_Inquiry_and_Word_Count_LIWC_LIWC2001.

[46] HU M Q, LIU B. Mining and summarizing customer reviews[C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2004: 168-177.

[47] 王科, 夏睿. 情感词典自动构建方法综述[J]. 自动化学报, 2016, 42(4): 495-511.
WANG K, XIA R. A survey on automatical construction methods of sentiment lexicons[J]. Acta Automatica Sinica, 2016, 42(4): 495-511. (in Chinese)

[48] 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1): 116-122.
XU L H, LIN H F, ZHAO J. Construction and analysis of emotional corpus[J]. Journal of Chinese Information Processing, 2008, 22(1): 116-122. (in Chinese)

[49] ZHENG H Z, XUE M H, LU H, et al. Smoke screener or straight shooter: detecting elite sybil attacks in user-review social networks[EB/OL]. [2025-01-05]. https://arxiv.org/abs/1709.06916.

[50] LI F T, HUANG M L, YANG Y, et al. Learning to

计算机工程 www.ecice16.com

identify review spam [C] // Proceedings of the 22nd International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2011: 2488-2493.
[51] SAMADI M, MOUSAVIAN M, MOMTAZI S. Deep contextualized text representation and learning for fake news detection[J]. Information Processing & Management, 2021, 58(6): 102723.
[52] LÜ C, HUANG D M, JIA Q Y, et al. Predictable model for detecting sybil attacks in mobile social networks [C] // Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC). Washington D.C., USA: IEEE Press, 2021: 1-6.
[53] WANG Y Q, MA F L, JIN Z W, et al. EANN: event adversarial neural networks for multi-modal fake news detection [C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM Press, 2018: 849-857.
[54] ZHANG X Y, CAO J, LI X R, et al. Mining dual emotion for fake news detection [C] // Proceedings of the Web Conference 2021. New York, USA: ACM Press, 2021: 3465-3476.
[55] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks [C] // Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2016: 3818-3824.
[56] RAKHLIN A. Convolutional neural networks for sentence classification[EB/OL]. [2025-01-05]. https://github.com/yoonkim/CNN_sentence.
[57] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. [2025-01-05]. https://arxiv.org/abs/1907.11692.
[58] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

文字编辑 吴云芳
栏目编辑 赖玉玲

计算机工程 www.ecice06.com