

基于图卷积网络的在线商品评论情感分析

尚 娜

燕山大学理学院, 河北 秦皇岛

收稿日期: 2024年5月8日; 录用日期: 2024年5月28日; 发布日期: 2024年6月5日

摘 要

现有的文本情感分析模型存在不能充分提取在线商品评论语义、全局特征信息导致分类准确率不高问题。预训练模型有较强的语义理解能力, 但是缺少全局特征信息, 而文本图卷积网络可以整合文本依赖信息和全局特征信息, 提高文本的表示能力。针对以上问题, 提出了一种新的情感分析模型ABGCN。首先使用ALBERT轻量级预训练模型进行词向量化并作为所构造文本图节点向量, 其次输入到文本图卷积网络中联合训练, 进行迭代更新得到在线商品评论特征信息, 解决了传统模型不能充分理解语义信息和全局结构信息的问题。最后发送给Softmax分类器进行情感分类。通过对比实验, 验证了提出的模型方法具有较高性能。

关键词

预训练模型, 图卷积网络, 情感分析

Sentiment Analysis of Online Product Reviews Based on Text Graph Convolutional Networks

Na Shang

School of Science, Yanshan University, Qinhuangdao Hebei

Received: May 8th, 2024; accepted: May 24th, 2024; published: Jun. 5th, 2024

Abstract

The existing text sentiment analysis model cannot fully extract the semantic and global feature information of online commodity reviews, resulting in low classification accuracy. Pre-training models possess strong semantic understanding capabilities, but lack global feature information. However, the text dependency information and global feature information can be integrated by

the text graph convolutional network, and the text representation ability is improved. To solve these problems, a new sentiment analysis model called ABGCN is proposed. Initially, the lightweight Pre-training model ALBERT is employed to perform word vectorization, generating node vectors for the constructed text graph. Subsequently, these embeddings are fed into the text graph convolutional network for joint training, enabling iterative updates to extract feature information from online product reviews, the problem of traditional models not fully understanding semantic information and global structural information has been resolved. Ultimately, the obtained features are forwarded to a softmax classifier for sentiment classification. Comparative experiments demonstrate the superior performance of the proposed model approach.

Keywords

Pre-Training Model, Text Graph Convolutional Network, Sentiment Analysis

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着科学技术的进步和互联网的普及,越来越多的网民不仅可以在互联网平台浏览文字新闻等信息,而且可以在互联网平台发表自己的观点和看法,比如电子商务网站中包含了众多商家对产品或服务的在线评论,文本信息呈爆炸性增长。从这些海量的文本信息中获取用户的首要选择或建议对于产品更新和服务改进至关重要,这就需要对文本进行情感分析。而情感分析是自然语言处理技术中文本分类的应用,是提取和识别文本中表达的观点或评价的情感极性的任务[1]。该技术在 Web 搜索、主题词提取、垃圾邮件或不良信息检测等领域有着广泛的应用。根据粒度的不同,可在文档级、句子级和方面级三个不同层次上考虑情感分类。

大量的文本数据如果只依靠人工进行筛选标注,会非常耗时耗力。目前,主要是用机器学习深度学习等方法,将收集数据集通过模型训练预测情感极性。张德阳[2]等提出了一种与主题相关的关键词提取算法,对微博情感倾向进行研究,得到微博句子情感极性。谷歌 Devlin [3]等提出 BERT (Bidirectional Encoder Representation from Transformers)模型,在自然语言处理任务中大放异彩,而且在情感分析任务中取得了良好的效果。主要是增加了掩码遮蔽策略并保留了传统的预测策略,该模型能获得比较完整的语义信息。但是,该模型随着语料库的增加和模型非常多而繁琐,训练时间和成本增加。因此谷歌 Lan [4]等在 BERT 模型基础上进行了改进,提出了 ALBERT (A Lite BERT)模型,该模型大大降低了参数量,提高了训练速度,而且效果不相上下。Kun [5]等为使学习任务能够主动适应预训练模型,可以通过添加模板改变模型输入,从而促进预训练模型的创新和适用性,促进预训练模型在情感分析等任务中的有效利用。对于传统的卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)和长短期记忆网络(Long Short-Term Memory, LSTM),在情感分析任务中有较好的性能和突破[6],但是受欧式结构化数据的限制。因此原本是非欧式结构化的文本等数据,只能单独对文本自身的上下文进行语义提取,而不能够对文本之间的相关信息进行表示,从而无法提取文本句法结构、空间等信息。Yao [7]等对于文本分析任务,首次提出文本图卷积网络(Text Graph Convolutional Network, TextGCN),该模型非常适合处理这类非欧式结构化数据。Lin [8]等先对数据集用 BERT 预训练,作为文本图节点特征,输入到图卷积网络中进行训练,弥补了 TextGCN 不能充分提取语义信息的不足,但是

BERT 参数大和文本图节点多，迭代更新会非常耗时。本文提出了 ABGCN 模型，首先用轻量级语言模型 ALBERT 预训练，作为所构造文本图的节点特征，可以提取文本语义特征信息。然后输入到文本图卷积网络中训练，能够提取节点更深层的特征信息并且去除无用信息的干扰，可增强文本特征。最终实现在线商品评论文本的情感分类，并有较好的性能。

2. 数据采集及预处理

本实验数据集使用网络爬虫技术爬取京东平台上各类商品的用户评价数据，采集时间为 2023 年 11 月 1 日，共包含 19,742 条数据，4 列，每行包含：用户名；评价星级(star 1~5)；用户评价内容；评价时间。数据示例见表 1 所示。

Table 1. Data sample
表 1. 数据示例

用户名	评价星级	用户评价内容	评价时间
2****	star5	耳机很好用，音质非常棒	2023-10-27 17:45
乔****	star3	外观很漂亮，但屏幕显示太模糊，显卡很差。	2023-07-02 11:48
j****	star1	容量小，基本上都是虚电，不建议购买	2023-10-31 18:44

首先需要对数据进行预处理，在线商品评论会存在大量复制粘贴的情况，为了降低训练时间和复杂度，对数据进行去重，去重后数据共有 16,507 条。将评价星级转为评价类型，分别为好评和差评，将 star 4~5 划分为好评，并设置标签为“1”；star 1~3 划分为差评，并设置标签为“0”，其中标签类别统计见图 1。

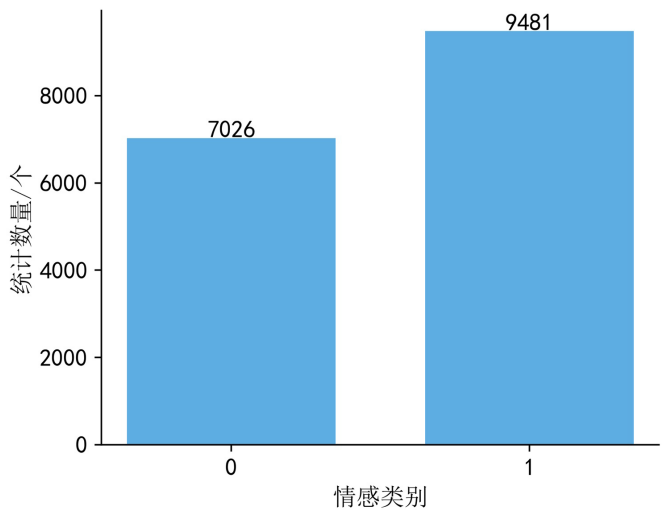


Figure 1. Category quantity statistics
图 1. 类别数量统计

删除数据中除字母、数字、汉字以外的所有符号，中文数据集需要用 jieba 进行分词处理，可以通过采用专门的算法和字典，能够将中文文本切分成有意义的词语，方便进行后续的处理和分析。去掉文本数据中被过滤或忽略的常见词语即停用词，如“的”、“是”、“我”等这些词语通常不携带太多的语义信息，因此去除它们可以减少数据中的噪声，使得后续的分析更加准确和可靠，数据集语料生成的词云见图 2。



Figure 2. Word cloud
图 2. 词云图

3. 基于文本图卷积网络的情感分析模型

本文基于图卷积网络的 ABGCN 模型使用 ALBERT 模型预训练词向量得到文本语义信息特征，构造文本图并将预训练的向量作为文本图节点特征向量，输入到两层文本图卷积网络模型训练，最后发送给 softmax 分类器进行情感分类。该模型可以充分提取文本的语义信息等，从而提高模型分类的准确率。

3.1. ALBERT 模型

3.1.1. 模型基本架构

ALBERT 模型是对采用了多层双向编码 Transform 编码的传统 BERT 模型轻量级改进[4]，模型图见图 3，预训练后的特征向量用 T 表示。除了保留了 BERT 模型优点外，还主要通过词嵌入参数因式分解和跨层参数共享两大机制进行了改进。

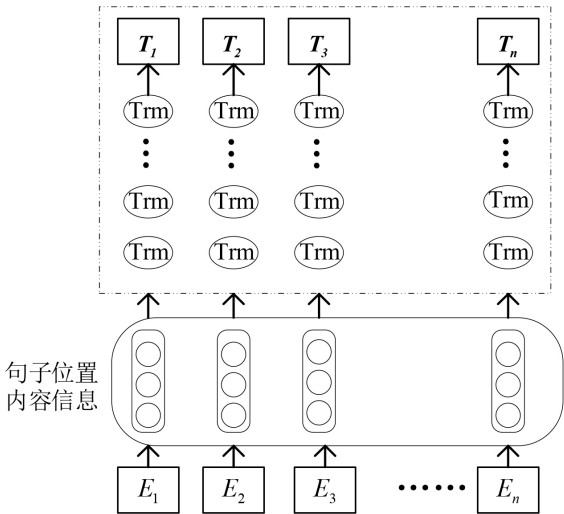


Figure 3. ALBERT model diagram
图 3. ALBERT 模型图

首先对词嵌入参数进行了因式分解,分解为两个小矩阵。将向量映射到一个低维词嵌入空间 E (BERT 模型 $H = E$, 参数多), 然后再映射到隐藏空间, 在词嵌入和隐藏层之间加入一个 project 层, 连接两个层。其次跨层参数共享的机制是单独用一个自注意机制循环 12 次, 每一层使用相同的参数。该模型还采取了句间语序预测 SOP (sentence-order prediction) 方法, 其组要是句子顺序预测, 正样本为正常顺序的 2 个相邻句子, 负样本为调换顺序的 2 个相邻句子, 所以此任务能够让模型学习到更多的信息[9]。

3.1.2. ABGCN 模型词嵌语义提取层

给定 $w_1, w_2, w_3, \dots, w_n$ 是评论文本一个句子中词语序列化的 n 个字符, 是组成句子的 n 个词语, 被映射到一个嵌入向量中进行输入。词嵌入是用 ALBERT 模型将句子中的每个词映射成词向量, 该模型减小了传统 BERT 模型参数量, 解决了传统模型太复杂参数过多会消耗大量时间的问题。该模型更加适合获取文本图节点特征对于文本图卷积网络节点多等较为复杂的模型。所有的文本形成一个词嵌入矩阵 $X_v \in \mathbf{R}^{v \times d}$, 其中 v 是所含词语的个数, d 是单词向量的维数。

3.2. 文本图卷积神经网络

3.2.1. 构造文本图

首先要将文本转换为图, 形成无向图 $G = (V, E)$, 图网络中有两种节点类型, 文本 document、词 word。词是文本中的不重复的词, 边一共有两种边: 文本和单词的边; 单词和单词的边。

将 ALBERT 生成后的特征向量作为文本图节点特征输入到构建的文本图之中, 可提取文本语义信息, 然后整合输入到改进的文本图卷积网络模型, 可提取语句全局结构信息, 最终将语义信息和全局结构信息进行联合特征融合。过程图见图 4, 其中示例 O1 文本: 灯造型大方美观, 性价比高, O2 文本: 电脑轻薄, 性价比高, O3 文本: 电脑外观美观大方。O 为文本节点, 其他是词节点, 粗边实线是文本和词节点边, 细边实线是词节点边, $R(x)$ 表示 x 通过预训练模型后特征向量嵌入。

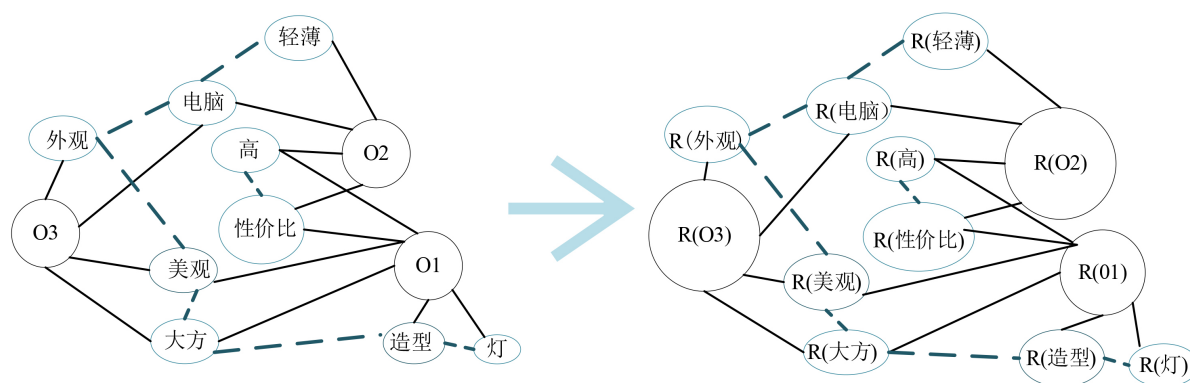


Figure 4. Text graph node feature extraction

图 4. 文本图节点特征提取

在无向图中我们利用点互信息 PMI (Point-wise Mutual Information) 来计算两个词节点之间的权重, PMI 的实验效果比用简单的单词共现频次作为边权重效果好。公式为式(1)~(3), 其中 W_i 是所有滑动窗口中包含单词 i 窗口个数, $W(i, j)$ 是指包含词 i 与词 j 的窗口个数, W 是总的滑动窗口个数。然后通过下面公式计算边权重。

$$\text{PMI}(i, j) = \log \frac{p(i, j)}{p(i)p(j)} \quad (1)$$

$$p(i, j) = \frac{\#W(i, j)}{\#W} \quad (2)$$

$$p(i) = \frac{\#W(i)}{\#W} \quad (3)$$

其中各边都带有不同的权重, 本文和词节点的边权重用 Tf-idf Transformer 定义, 并加入自循环, 边的权重 A_{ij} 总结通过式(4)来定义。

$$A_{ij} = \begin{cases} \text{PMI}(i, j) & i, j \text{ 词, } \text{PMI}(i, j) > 0 \\ \text{TF-IDF}_{i,j} & i \text{ 文本, } j \text{ 词} \\ 1 & i = j \\ 0 & \text{其他} \end{cases} \quad (4)$$

3.2.2. 图卷积层

图卷积网络是一个多层神经网络, 它直接在图上运行, 并根据节点的邻域属性诱导节点的嵌入向量。在形如 $G=(V, E)$ 图中, $V(|V|=n)$ 和 E 是节点和边的集合, $X \in \mathbf{R}^{n \times m}$ 是包含 n 的节点的特征矩阵, 其中 m 是特征向量的维度。在图结构中引入邻接矩阵 $A \in \mathbf{R}^{n \times n}$, 由于自循环对角元素设置为 1, D 为顶点的度矩阵 $D_{ii} = \sum_j A_{ij}$ 。对于单层图卷积网络, k 维度特征节点矩阵, 迭代公式为式(5)。

$$H^{(1)} = \rho(\tilde{A}XW_0) \quad (5)$$

其中 $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, $W_0 \in \mathbf{R}^{m \times k}$ 是归一化的邻接矩阵, 激活函数使用 ReLU, 本文 ABGCN 模型使用有两层 GCN, 使用图卷积运算对文本图进行卷积, 来合并高阶邻域信息, 其中 $H^{(0)} = X$, 则 j 层卷积迭代公式为式(6), W_{k-1} 为第 k 层的权重矩阵; h_k^i 为第 k 层 GCN 的输出, 经线性变换得到图卷积层整个更新迭代最后的输出为式(8)。

$$H^{(j+1)} = \rho(\tilde{A}H^{(j)}W_j) \quad (6)$$

$$H_i^{(k)} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H_i^{(k-1)}W_{k-1}\right) \quad (7)$$

3.2.3. 模型训练

用 Softmax 分别输出 ALBERT 层和 GCN 层的分类概率为式(8)~(9), 最后输出 ABGCN 模型和 ALBERT 模型输出分类概率的线性插值, 可以平衡两个模型分类效果从而提高 ABGCN 的模型情感分类性能。并加入 dropout 层来防止数据过拟合, 提升模型的泛化能力, 优化器选用 Adam。图模型有助于捕获句法全局信息, 以提高情感分类性能。

$$Z_{\text{GCN}} = \text{softmax}\left(\tilde{A} \text{ReLU}(\tilde{A}XW_0)W_1\right) \quad (8)$$

$$Z_{\text{ALBEET}} = \text{softmax}(WX) \quad (9)$$

$$Z = \lambda Z_{\text{GCN}} + (1 - \lambda) Z_{\text{ALBERT}} \quad (10)$$

损失函数采用交叉熵, y_D 是交叉熵的文本索引集, F 是输出的类别数, 本文数据集是二分类, Y 是标签指示器矩阵, 如式(11)。

$$\mathcal{L} = - \sum_{d \in y_D} \sum_{f=1}^F y_{df} \ln Z_{df} \quad (11)$$

4. 实验分析

4.1. 实验条件环境和参数

实验操作系统是在 Ubuntu20.04，显存：32 GB，处理器是 GPU：V100-32GB，CPU 为 10 vCPU Intel Xeon Processor (Sky-lake, IBRS)，采用 python 语言在 pycharm 工具下实现，开发环境为 python3.8、pytorch1.5.1。实验设置训练集测试集验证集比例为 8:1:1。实验模型中 ALBERT 采用 albert-Chinese-small，实验参数设置见表 2 所示。

Table 2. Parameter settings
表 2. 参数设置

参数	值
最大序列长度	128
Bach_size	128
Hidden_size	200
Dropout 率	0.1
bert_lr	2e-6
gcn_lr	1e-4
GCN layers	2
Epoch	40
λ	0.7

4.2. 对比实验设计

- 根据本文介绍为验证此模型的优越性，与现有基础模型作对比模型实验，共设计 6 种。
- 1) CNN：卷积网络 CNN 提取文本特征，最后用 softmax 情感分类。
 - 2) LSTM_att：长短时记忆网络 LSTM 进行文本特征提取，并加入注意力机制。
 - 3) BiLSTM_att：使用双向长短时记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)进行文本特征提取，并加入注意力机制。
 - 4) BERT：使用 bert-base-chinese 预训练模型[3]。
 - 5) ALBERT：使用 albert-chinese-small 预训练模型[4]。
 - 6) BERTGCN：使用 BERT 模型输出作为文本图的节点特征，输入到文本图卷积网络中[8]。

4.3. 评估指标

本模型主要用以下评价指标，分别为准确率(Accuracy, Acc)、精确率(Precision, P)、召回率(Recall, R)和 F1 值，准确率是分类正确文本占总文本的比例，精确率判断为正例子中被正确分类的文本比例，回召率是指真实为正例子中正确判断的比例，而 F1 分数值是综合性评价指标，可以表示模型整体文本情感的预测效果并且能衡量整个分类模型的整体性能，其计算公式为式(12)~(14)。

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{12}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{13}$$

$$F1_i = \frac{2P_iR_i}{P_i + R_i} \tag{14}$$

4.4. 实验结果分析

记录本文模型与其它 6 种模型的测试集平均分类准确率和 F1 值其结果见表 3 和图 5，本文模型训练准确率和损失值见图 6。

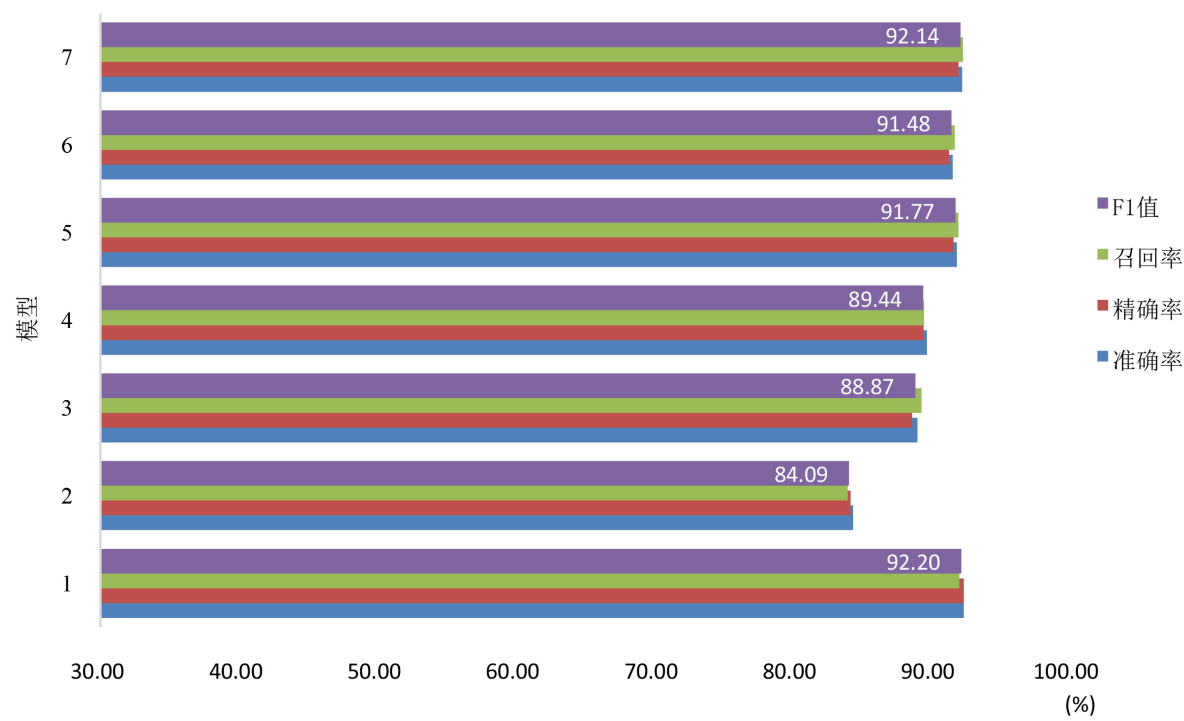


Figure 5. Compare each index value of model results
图 5. 对比模型结果各个指标值

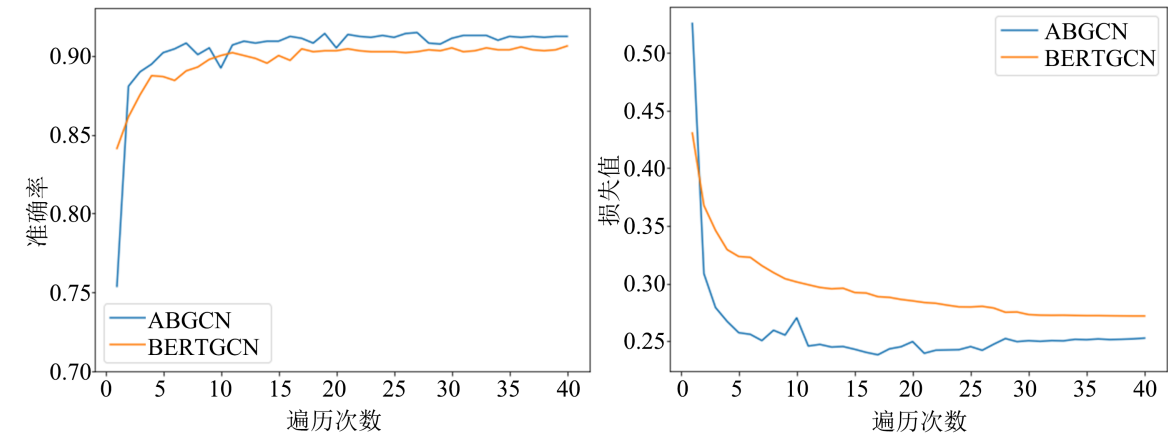


Figure 6. Models training accuracy graph
图 6. 模型训练准确率和损失值图

Table 3. Model comparison result
表 3. 模型对比结果

	模型	准确率	精确率	召回率	F1 值
1	ABGCN (本文模型)	92.37	92.39	92.05	92.20
2	CNN	84.37	84.20	84.00	84.09
3	LSTM_att	89.04	88.64	89.32	88.87
4	BiLSTM_att	89.70	89.47	89.49	89.44
5	BERT	91.88	91.63	91.98	91.77
6	ALBERT	91.58	91.31	91.73	91.48
7	BERTGCN	92.25	92.00	92.32	92.14

根据表 3 和图 5，分析本章模型和其它 6 组对比模型实验结果，通过实验 1、2、3、4 验证了本文模型比现有的卷积模型有更高的准确率和 F1 值。通过实验 5 [3]和实验 6 [4]，BERT 模型准确率和 F1 值比 ALBERT 模型略高，但是通过实验 7 [8]和实验 1，将 ALBERT 预训练模型替换 BERT 模型作为文本图节点特征向量，准确率和 F1 值分别升高了 0.12%、0.06%；根据图 6 所示，在模型训练达到稳定后 ABGCN 具有较高的准确率，并且其损失值更低，而且本文模型的模型参数量是实验 7 模型参数量的 0.05%，提高了运行效率。验证了 ALBERT 模型较适合作为图卷积网络节点特征向量，输入到文本图节点多而较复杂的图卷积网络。实验 6 模型和本章模型、实验 5 模型和实验 7 模型相比，去除了图卷积网络模块，发现准确率分别下降了 0.79%、0.37%，F1 值分别下降了 0.72%、0.37%，验证了图卷积网络的重要性。将节点进行特征表示能提取更深层语义特征信息，输入到图卷积网络中训练充分提取文本全局结构信息。综上所述，本文 ABGCN 模型准确率和 F1 值均高于对比模型，本章提出的模型可以更好提取文本语义和全局结构信息，对于在线商品评论文本情感分类有较好的性能。

5. 结论与展望

本文提出了 ABGCN 情感分析模型，本文实验数据集使用网络爬虫获取的 19,742 条京东商品评论文本，经过实验分析验证了以下结论。使用 ALBERT 预训练模型作为文本图特征节点输入，解决了传统模型不能充分提取文本语义信息和全局结构信息的问题，实验结果准确率和 F1 值分别达到了 92.37%、92.20%，相比传统模型提高了其性能，并且 ABGCN 较文献[8] BERTGCN 准确率和 F1 值分别提高了 0.12%、0.06%，ALBERT 模型参数少较适合联合训练图卷积网络文本图节点多等复杂的模型。本文模型可以更好提取文本语义和全局结构信息，通过实验验证了本文模型对在线商品评论文本情感分析有较好的性能。但是目前在构建文本图过程中还缺少使用其他外部语料知识等信息，后续研究可以进一步提升这类信息的特征提取能力，可以有更好的性能。

参考文献

[1] Phan, H.T., Nguyen, N.T., Tran, V.C., *et al.* (2021) An Approach for a Decision Making Support System Based on Measuring the User Satisfaction Level on Twitter. *Information Sciences*, **561**, 243-273.
<https://doi.org/10.1016/j.ins.2021.01.008>

[2] 张德阳, 韩益亮, 李晓龙. 基于主题的关键词提取对微博情感倾向的研究[J]. 燕山大学学报, 2018, 42(6): 552-560.

[3] Devlin, J., Chang, M., Lee, K., *et al.* (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computa-*

tional Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2-7 June 2019, 4171-4186.

- [4] Lan, Z.Z., Chen, M., Goodman, S., et al. (2020) ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *Proceedings of the International Conference on Learning Representations*, Addis Ababa, Ethiopia, 3-4 March 2020, 1-17.
- [5] Bu, K., Liu, Y.C. and Ju, X.L. (2023) Efficient Utilization of Pre-Trained Models: A Review of Sentiment Analysis via Prompt Learning. *Knowledge Based Systems*, **283**, Article 111148. <https://doi.org/10.1016/j.knosys.2023.111148>
- [6] Zhang, L., Wang, S. and Liu, B. (2018) Deep Learning for Sentiment Analysis: A Survey. *Wiley Inter Disciplinary Reviews Data Mining and Knowledge Discovery*, **8**, e1253. <https://doi.org/10.1002/widm.1253>
- [7] Yao, L., Mao, C.S. and Luo, Y. (2019) Graph Convolutional Networks for Text Classification. *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, **33**, 7370-7377. <https://doi.org/10.1609/aaai.v33i01.33017370>
- [8] Lin, Y.X., Meng, Y.X. and Sun, X.F. (2021) BertGCN: Transductive Text Classification by Combining GCN and BERT. In: Zong, C.Q., Xia, F., Li, W.J. and Navigli, R., Eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Bangkok, 1456-1462. <https://doi.org/10.18653/v1/2021.findings-acl.126>
- [9] 张楠, 苏南, 王贵阳, 等. 深度学习之自然语言实战[M]. 北京: 机械工业出版社, 2020: 178-180.