计算机科学
COMPUTER SCIENCE

结合上下文和依存句法信息的中文短文本情感分析

杜启明, 李男, 刘文甫, 杨舒丹, 岳峰

引用本文

杜启明, 李男, 刘文甫, 杨舒丹, 岳峰结合上下文和依存句法信息的中文短文本情感分析[J]. 计算机科学, 2023, 50(3): 307-314.

DU Qiming, LI Nan, LIU Wenfu, YANG Shudan, YUE Feng. Sentiment Analysis of Chinese Short Text Combining Context and Dependent Syntactic Information [J]. Computer Science, 2023, 50(3): 307-314.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200189

结合上下文和依存句法信息的中文短文本情感分析

杜启明[1,2]　李　男[1,2]　刘文甫[1,2,3]　杨舒丹[1,2]　岳　峰[1,2]
1 信息工程大学网络空间安全学院　郑州 450000
2 信息工程大学数学工程与先进计算国家重点实验室　郑州 450000
3 电子信息系统复杂电磁环境效应重点实验室　河南 洛阳 471003
(qimingducst@163.com)

摘　要　依存句法分析旨在从语言学的角度分析句子的句法结构。现有的研究表明,将这种类似于图结构的数据与图卷积神经网络(Graph Convolutional Network,GCN)进行结合,有助于模型更好地理解文本语义。然而,这些工作在将依存句法信息处理为邻接矩阵时,均忽略了句法依赖标签类型,同时也未考虑与依赖标签相关的单词语义,导致模型无法捕捉到文本中的深层情感特征。针对以上问题,提出了一种结合上下文和依存句法信息的中文短文本情感分析模型(Context and Dependency Syntactic Information,CDSI)。该模型不仅利用双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)提取文本的上下文语义,而且引入了一种基于依存关系感知的嵌入表示方法,以针对句法结构挖掘不同依赖路径对情感分类任务的贡献权重,然后利用GCN针对上下文和依存句法信息同时建模,以加强文本表示中的情感特征。基于SWB,NLPCC2014和SMP2020-EWEC数据集进行验证,实验表明CDSI模型能够有效融合语句中的语义以及句法结构信息,在中文短文本情感二分类以及多分类中均取得了较好的效果。

关键词:句法结构;上下文信息;GCN;中文短文本
中图法分类号　TP391.1

Sentiment Analysis of Chinese Short Text Combining Context and Dependent Syntactic Information

DU Qiming[1,2],LI Nan[1,2],LIU Wenfu[1,2,3],YANG Shudan[1,2] and YUE Feng[1,2]
1 School of Cyberspace Security Academy,Information Engineering University,Zhengzhou 450000,China
2 State Key Laboratory of Mathematical Engineering and Advanced Computing,Information Engineering University,Zhengzhou 450000,China
3 State Key Laboratory of Complex Electromagnetic Environment Effect on Electronic and Information System,Luoyang,Henan 471003,China

Abstract　Dependency parsing aims to analyze the syntactic structure of sentences from the perspective of linguistics. Existing studies suggest that combining such graph-like data with graph convolutional network(GCN) can help model better understand the text semantics. However,when dealing with dependency syntactic information as adjacency matrix,these methods ignore the types of syntactic dependency tags and the word semantics related to the tags,which makes the model unable to capture the deep emotional features. To solve the preceding problem,this paper proposes a Chinese short text sentiment analysis model CDSI(context and dependency syntactic information). This model can use BiLSTM(bidirectional long short-term memory) network to extract the context semantics of the text. Moreover,a dependency-aware embedding representation method is introduced to mine the contribution weights of different dependent paths to the sentiment classification task based on the syntactic structure. Then the GCN is used to model the context and dependent syntactic information at the same time,so as to strengthen the emotional features in the text representation. Based on SWB,NLPCC2014 and SMP2020-EWEC datasets,experimental results show that CDSI can effectively integrate the semantic and structural information in sentences,which achieves good results in both the Chinese short text sentiment binary classification and multi-classification tasks.

Keywords　Syntactic structure,Context information,GCN,Chinese short text

1　引言

随着信息时代的到来,以微博、今日头条为代表的数字化网络新闻平台在短短几年内迅速发展,成为了人们获取外界信息与发表个人观点的主要渠道。在此背景下,当前的网络新闻平台中充斥着大量的网民评论数据,其中往往蕴含了

到稿日期:2021-12-16　返修日期:2022-04-21
基金项目:国家自然科学基金(61802433)
This work was supported by the National Natural Science Foundation of China(61802433).
通信作者:李男(linan_happy@126.com)

大众对舆情事件的认知、对热点新闻的情感倾向等信息。利用文本情感分析技术对网络新闻下的网民评论进行情感分析，将有助于相关部门及时掌握网民情感倾向，这对舆情态势研判、舆情处置决策制定等具有重要的现实意义。

从语言学层面来讲，中文特有的语言结构较为复杂，而像网络新闻评论这种典型的短文本数据又存在特征稀疏、噪声大等问题[1-2]，这无疑给情感分析任务提出了更高的要求。近年来，基于深度学习的情感分析是较为主流的研究方向。其中，该类方法往往先借助词嵌入技术[3]将单词映射为低维稠密向量，随后将其输入深度学习网络模型中以针对文本中的复杂语义进行建模。常用的模型包括卷积神经网络[4]（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network，RNN）[5]、BiLSTM[6]等。相比传统的机器学习方法，深度学习网络提供了一种端到端的学习范式，很大程度上减少了手工特征提取的工作量。然而，基于深度学习的情感分析方法常常忽略自然语言中关于句法结构、情感的先验知识，导致情感分析不够精确，限制了其应用范围。

作为语言理解的重要一环，依存句法分析旨在从语言学的角度理解语句中各成分之间的结构关系，从而辅助模型更好地理解文本语义。基于该优势，一些工作尝试将依存句法分析技术引入情感分析任务中。根据信息利用方式的不同，结合依存句法分析技术的情感分析研究大致可以划分为3个发展阶段：1）根据依存句法分析结果构造情感计算规则，如基于情感特征词分布动态计算文本情感指数[7-8]；2）编码句法依赖关系以增强文本表示，如结合主题信息用于社会情绪分类[9]；3）将依存句法信息与GCN进行结合[10]，进一步提高了情感分析质量。现有工作表明，依存句法分析的引入有助于模型理解语言中的复杂语义关系，进一步为自然语言处理任务提供了新的解决思路，但仍存在一些不足之处，如常忽略句法依赖信息中的标签类型以及与标签相关的单词语义，实际上这些信息往往对最终情感分析结果的判定贡献较大。以短文本语句“我讨厌疫情!”为例，经过依存句法分析后可以得到如图1所示的解析效果。

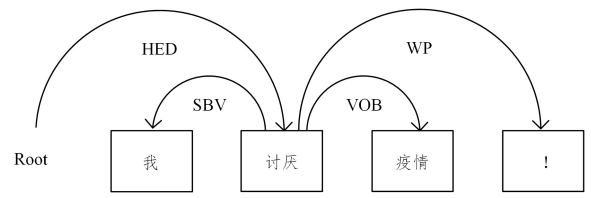

图1 依存句法分析结果（以“我讨厌疫情!”为例）

Fig. 1 Dependence parsing result (taking “I hate the epidemic!” as an example)

图1中，Root代表根节点，HED表示核心关系，SBV为主谓关系，VOB为动宾关系，WP指标点符号且与核心词之间存在句法依赖。依存句法分析结果认为该语句的核心成分为“讨厌”且句末标点符号为“!”，如果暂时忽略其他的主语、宾语等成分，结合“讨厌”和“!”的语义可以很容易地判定该语句的情感极性为消极。可以看出，依存句法结构图中的依赖标签类型以及与标签相关的单词语义对整个文本的情感色彩判定至关重要。如何能让模型有效利用文本中的语义以及句法结构信息，正是本文研究的出发点。

针对以上问题，本文提出了一种结合上下文和依存句法信息的中文短文本情感分析模型CDSI。该模型一方面利用BiLSTM提取文本的上下文语义，另一方面引入一种基于依存关系感知的嵌入表示方法，以针对句法结构挖掘不同依赖路径对情感分类任务的贡献权重。在此基础上利用GCN针对上下文和依存句法信息同时建模，加强了文本表示中的情感特征，达到了提高情感分析精度的目标。

本文的主要贡献概括如下：

（1）提出了一种结合上下文和依存句法信息的中文短文本情感分析模型，主要包括语义提取层、依存句法信息嵌入层、图卷积层和情感分类层4个部分。该模型针对上下文信息和依存句法信息同时建模，能够有效挖掘并利用文本中的情感特征。

（2）在依存句法信息嵌入层中设计实现了一种基于依存关系感知的嵌入表示方法。该方法不仅关注单词之间的依存相关性，而且针对依赖标签类型及与标签相关的单词语义同时进行编码，有利于模型挖掘出不同依赖路径对情感分类任务的贡献权重。

（3）基于3个公开的短文本评论数据集进行了相关实验，实验结果表明CDSI模型在中文短文本情感二分类以及多分类任务中均表现优异，体现了该模型的有效性。

2 相关工作

2.1 依存句法分析

依存句法分析旨在将文本解析为一棵依存句法树，继而可以得到单词之间的依赖关系和关联路径，也即表示该方法能从句子结构角度提取文本特征，有助于模型更好地理解自然语言[11]。近年来，依存句法分析受到越来越多研究者的关注，并在多个领域得到了广泛的应用。Guo等[12]以全依赖树作为图卷积网络的输入，并在训练过程中引入注意力机制，从而有选择性地关注依赖子结构，最终在多个关系抽取任务中表现优异。Zhang等[13]使用图卷积网络，针对加权图进行编码，同时引入多任务学习框架来减轻句法解析带来的错误传播，提升了模型在观点角色标注任务上的性能。Wang等[14]首先通过依存句法分析识别文本中的积极成分和消极成分，然后结合情感词典对新闻文章进行情感分析，改善了财经新闻的情感分类效果。Zhang等[15]将依存树中父节点和子节点的信息融入BiLSTM中进行编码，然后使用注意力机制动态选择两者特征，最终将其输入到条件随机场（Conditional Random Field，CRF）中进行命名实体识别。实验结果表明，该方法进一步提高了识别效果，并且在长实体识别上具有明显的优势。

2.2 图卷积神经网络

直观来看，文本的句法结构与图数据较为相似。对于这种非欧氏空间数据，传统的深度学习模型并不能有效利用甚至可能会破坏其内在信息。为此，Kipf等[16]尝试将卷积扩展到图结构数据，提出了著名的图卷积神经网络GCN。由于其可以针对现实中常见的图数据进行建模，继而挖掘其中的复杂关系，因此该模型在推荐系统、网络分析、自然语言处理等

领域得到了较为广泛的应用。

在情感分析领域，已有一些工作利用 GCN 对依存句法信息展开深入挖掘研究。Sun 等[17]提出一种基于依存句法树的卷积模型 CDT，该模型首先利用 BiLSTM 学习句子的特征表示，然后使用 GCN 针对句法依赖关系建模以进一步增强嵌入，最终实现了上下文和句法依赖信息从观点词到方面词的传播，在方面级情感分析任务中表现较好。Lai 等[18]将有句法依赖关系的两个单词之间的距离记为 1，无依赖关系则记为 0，继而得到语句的邻接矩阵表示，随后将单词向量表示和文本邻接矩阵表示共同输入 GCN 中，实验结果表明依存句法信息能够有效提高情感分类模型的性能。基于该工作，Fan 等[19]引入自注意力机制以充分学习网民评论中的情感信息，进一步优化了基于依存句法分析的文章级情感分析模型。

上述工作表明，依存句法分析和 GCN 的结合可以有效利用文本中的句法结构信息，在一定程度上弥补了传统的深度学习模型在相关句法约束和远距离单词依赖方面的缺陷，改善了情感分析质量。然而，现有的研究工作仍存在一些不足：1）在将依存句法信息处理为邻接矩阵时，它们往往忽略了句法依赖信息中的标签类型以及与标签相关的单词语义；2）将中文句法结构与 GCN 进行结合的研究较少，目前尚没有统一的中文句法依赖类型标准。为此，本文将结合 BiLSTM，GCN 以及依存句法分析技术，针对中文短文本情感分析问题进行探索研究。

3 模型设计与实现

假设待分类的短文本集合为 C，选取其中某条文本表示为 $D=(T_1,T_2,\cdots,T_n)$，T_j 表示文本 D 中第 j 个词语，n 表示 D 中所包含词语的个数，目标情感类别表示为 $S=(S_1,S_2,\cdots,S_m)$。则本文所描述的短文本情感分析任务的实质为判定 D 的情感极性，即使用 CDSI 模型来计算 D 在 S 中各类别上的似然概率分布，并求出最大概率分布值所对应的类别。其形式化定义描述如下：

$$CDSI:(D,S)\rightarrow \hat{S}_D=\mathrm{argmax}(s_1,s_2,\cdots,s_m) \tag{1}$$

CDSI 模型的具体结构如图 2 所示，主要包括 4 个部分：语义提取层、依存句法信息嵌入层、图卷积层和情感分类层。下文将对各层展开详细描述。

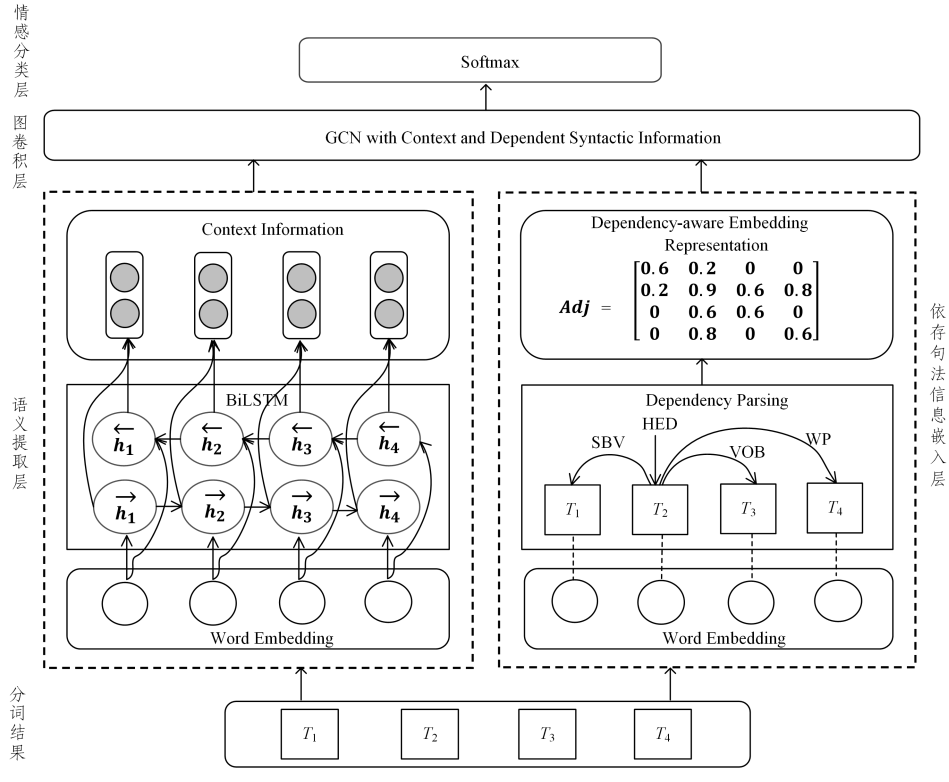

图 2 CDSI 模型结构

Fig. 2 Structure of CDSI model

3.1 语义提取层

语义提取层首先借助词嵌入模型将短文本 D 表示为低维向量，然后利用 BiLSTM 学习 D 的上下文信息，具体包括如下步骤。

（1）词向量表示。现有研究表明，基于语言学分布假设的词嵌入模型[20]可以挖掘文本中的语义信息，其输出的分布式稠密向量在多种自然语言处理任务中表现出显著的优势。假设当前单词与向量的映射关系为 $W\in R^{|v|\times d}$，其中 $|v|$ 为单词个数，d 为词向量维度。对于包含 n 个词语的短文本 D，考虑到其中的单词 T_j 不一定在 W 中存在映射关系，这种情况下本文将在区间 $[-0.25,0.25]$ 上使用均匀分布对其进行随机初始化，则关于单词 T_j 的词向量表示规则描述如下：

$$W(T_j)=\begin{cases} w_j^d, & T_j\in W \\ uniform(-0.25,0.25)^d, & T_j\notin W \end{cases} \tag{2}$$

基于映射关系 W，D 将被转化为一个实数向量矩阵 $\boldsymbol{M}_D\in R^{n\times d}$。

(2)上下文信息建模。作为循环神经网络 RNN 家族中的一员,BiLSTM 由一个前向的 LSTM 和后向的 LSTM 组合而成。相比经典的 RNN,该网络利用门控机制有效缓解了梯度爆炸或消失问题,而其中前后向的组合机制能更好地捕捉双向的语义依赖,非常适用于文本上下文信息建模。其计算规则描述如下:

$$\overrightarrow{h_t}=\overrightarrow{LSTM}(x_t,\overrightarrow{h_{t-1}}) \tag{3}$$

$$\overleftarrow{h_t}=\overleftarrow{LSTM}(x_t,\overleftarrow{h_{t+1}}) \tag{4}$$

$$h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]\in R^{2d_t} \tag{5}$$

其中,x_t为当前单元的输入,$\overrightarrow{h_{t-1}}$表示前一个单元的输出,$\overleftarrow{h_{t+1}}$表示后一个单元的输出,"→"表示前向,"←"表示后向,d_t表示当前单元在某一个方向上的输出向量维度,h_t表示将当前单元的前向输出和后向输出进行拼接。对于短文本 D,将基于映射关系得到的词嵌入表示矩阵$\boldsymbol{M}_D$输入 BiLSTM 中,经过计算后得到文本 D 所对应的上下文信息,记为$\boldsymbol{H}_D=(h_1,h_2,\cdots,h_n)\in R^{n\times 2d_t}$。

3.2 依存句法信息嵌入层

依存句法信息嵌入层首先利用开源的 LTP 工具获取文本中的依存结构信息,随后使用一种基于依存关系感知的嵌入表示方法针对该信息进行嵌入编码,具体包括如下步骤。

(1)中文依存句法分析。依存句法分析旨在获得文本中各成分之间的依赖关系(与"依存关系"含义一致)和关联路径,其分析结果往往作为一种先验知识来辅助模型更好地理解文本语义。然而在当前的中文依存句法分析领域中,尚不存在统一的依赖类型标准,为此本文以哈尔滨工业大学提供的 LTP 依存句法分析标注集[21]作为参考。与此同时,文献[22]指出在构造邻接矩阵时考虑自环和对边情况能提升 GCN 的性能,文献[23]考虑了标点符号对语句情感色彩判定的影响。结合上述工作,本文额外添加了 SL 和 WP 两种特殊依赖关系,构建了一个更加完善的中文句法依赖类型集合,将其定义为 Ω。其中,SL 代表自相邻关系,表明文本中各单词自身与自身之间存在依赖关系;而 WP 代表标点符号修饰关系,表明该成分是文本中的标点符号,且与核心词之间存在依赖关系。集合 Ω 中所包含的句法依赖关系信息如表 1 所列。

表 1 依赖关系类型描述

Table 1 Description of dependency types

符号	依赖关系类型	符号	依赖关系类型
SL	自相邻关系	CMP	动补结构
SBV	主谓关系	COO	并列关系
VOB	动宾关系	POB	介宾关系
IOB	间宾关系	LAD	左附加关系
FOB	前置宾语	RAD	右附加关系
DBL	兼语	IS	独立结构
ATT	定中关系	HED	核心关系
ADV	状中结构	WP	标点符号修饰关系

对于短文本 D,基于该步骤得到的依存句法分析结果与图结构数据类似,其形式化定义描述如下:

$$G_D=(V,E) \tag{6}$$

其中,$V=\{T_i\mid T_i\in D\}$,$E=\{e_{ij}\mid e_{ij}\in\Omega\}$。下一步骤中将结合该分析结果以及单词语义信息进行嵌入编码。

(2)基于依存关系感知的嵌入表示。实质上,该步骤旨在根据语句的依存结构生成相应的邻接矩阵 $\boldsymbol{A}_D\in R^{n\times n}$。传统的依存关系嵌入方法常使用 1,0 来编码单词之间的句法依赖关系,也即表示邻接矩阵 $\boldsymbol{A}_D$ 中的元素取值仅为 1 或 0。然而,这种方式忽略了不同的依赖关系对目标任务的影响,同时也引入了一定的冗余特征。在关系抽取领域中,Guo 等[12]根据依赖路径的距离或单词的相关性对冗余的词依赖弧进行修剪,进而改善了模型的效果。受此启发,为更加充分地利用句法依赖关系来加强最终文本表示的情感特征,本文在该层引入了一种基于依存关系感知的嵌入表示方法。该方法不仅关注单词之间的依存相关性,还考虑了依赖标签类型以及与标签相关的单词语义,能够挖掘不同依赖路径对情感分类任务的贡献权重。

同样以短文本语句"我讨厌疫情!"为例,图 3 给出了基于该方法的嵌入效果。可以看到,在生成的邻接矩阵中,每个元素的值在 0 到 1 之间,不同的值对应了不同依赖关系对情感分类任务的贡献权重,其值越大则意味着该依赖路径对目标任务价值越高。

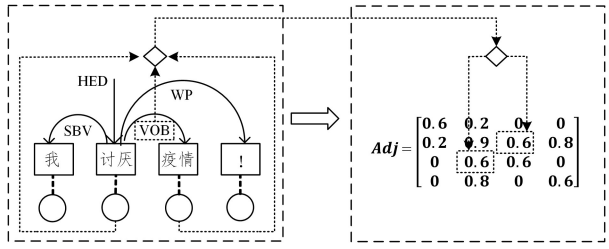

图 3 依存句法信息嵌入示例(以"我讨厌疫情!"为例)

Fig. 3 Demo of dependency syntax information embedding (taking "I hate the epidemic!" as an example)

下面将介绍关于基于依存关系感知的嵌入表示方法的实现细节。当前已知针对短文本 D 的依存句法分析结果G_D,如果其中的单词T_i和T_j之间存在依赖关系,则定义其依赖类型为 φ,那么相应的依赖类型嵌入向量可表示为$\text{ऽ}_\varphi\in R^{d_\varphi\times 1}$。结合相应的单词向量表示,则单词$T_i$和$T_j$之间的依存关系可嵌入表示为:

$$a_{ij}=\text{sigmoid}(\text{avg}[W(T_i),W(T_j)]\times\omega_\varphi\times\text{ऽ}_\varphi+b_\varphi) \tag{7}$$

其中,$\omega_\varphi\in R^{d_{W(T)}\times d_\varphi}$,$b_\varphi$均为可训练的参数,$d_\varphi$表示依赖类型嵌入维度,$d_{W(T)}$表示词向量的维度,$W(T_i)$和 $W(T_j)$为单词的词嵌入表示,avg 表示求其平均值,sigmoid 表示激活函数,ऽ_φ在模型训练前进行初始化并在训练过程中更新。相应地,邻接矩阵 $\boldsymbol{A}$ 的计算规则为:若单词T_i和T_j之间存在句法依赖关系,则有$A_{ij}=a_{ij}$,反之$A_{ij}=0$。此外,为将本文所提出的嵌入表示方法适用于各种依赖关系,规定式(7)中当 φ 为 HED 或 SL 时,$T_i=T_j$。

3.3 图卷积层

图卷积层主要负责基于前两层所提取的特征进行卷积操作。实质上,3.1 节中得到的向量$\boldsymbol{H}_D$即为节点特征,而 3.2 节中得到的矩阵$\boldsymbol{A}_D$即为边特征,将两者共同输入到 l 层的 GCN 中,随后按照以下公式进行卷积计算:

$$H^{(l)}=\text{Relu}(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{A}_D\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l-1)}+b_l) \tag{8}$$

其中，Relu 为激活函数，$\tilde{\boldsymbol{D}}$为$\boldsymbol{A}_D$的度矩阵，$\boldsymbol{H}^{(l-1)}$表示第 $l-1$ 层的节点特征，而 l 为 1 时，$\boldsymbol{H}^{(l-1)}=\boldsymbol{H}_D$，$\boldsymbol{W}^{(l-1)}$表示 GCN 中第 $l-1$ 层的权重矩阵。

随后对 GCN 的输出进行平均池化。为进一步提升模型泛化能力，后续还添加了 Dropout 和 Relu 层，最终将结果输入到一个全连接层中：

$$\boldsymbol{H}^{*}=\omega * \mathrm{Relu}(Dropout(Pooling(\boldsymbol{H}^{(l)})))+b \tag{9}$$

其中，$\boldsymbol{H}^{*}$即为经过图卷积层计算后所得到的文本特征表示，其有效融合了文本中的语义以及句法结构信息。

3.4 情感分类层

情感分类层主要根据图卷积层的输出对文本语句进行情感色彩的判定。本文在该层使用经典的 softmax 激活函数，结合式(1)，预测公式描述如下：

$$\hat{S}_D=\mathrm{argmax}(\mathrm{softmax}(\boldsymbol{H}^{*})) \tag{10}$$

其中，$\hat{S}_D$表示情感标签预测结果。针对本文要解决的短文本情感分类问题，CDSI 模型采用交叉熵(cross-entropy)作为损失函数，同时在训练过程中加入L_2正则化系数以避免过拟合。关于该模型的损失函数描述如下：

$$loss=-\sum_{i=1}^{|S|}S_D\log\hat{S}_D+\lambda\|\theta\|^2 \tag{11}$$

其中，S_D表示真实情感标签，λ为L_2正则化系数，θ为正则化参数。

4 实验与分析

实验的软件环境为 Python 3.6.13，Pytorch 1.6.0，Pyltp 0.2.1，Pyhanlp 0.1.79，HarvestText 0.8.1.4。

4.1 实验准备

(1)数据集介绍。下文实验使用 3 个公开的短文本情感分析数据集来验证 CDSI 模型的有效性。其中，SWB 数据集[1)]来源于新浪微博，源数据中已标注好正负向情感标签；NLPCC2014 情感评测数据集[2)]来源于第三届自然语言处理与中文计算会议所发布的开放评测任务，类别包括 none，like，disgust，surprise，happiness，fear，sadness 和 anger；SMP2020-EWECT 情绪分类数据集[3)]来源于第九届全国社会媒体处理大会所发布的公开评测任务，类别包括积极、愤怒、悲伤、恐惧、惊奇和无情绪。

(2)实验数据与工具准备。考虑到数据集中存在类别不平衡现象，针对上述 3 个数据集进行了类别合并、预处理、数据重组等操作。具体来讲，对于 SWB 数据集，本文各抽取 5000 条数据用于测试 CDSI 模型在情感二分类任务中的效果。对于 NLPCC2014 数据集，本文将 happiness，like 视为积极标签，将 disgust，surprise，fear，sadness 和 anger 视为消极标签，将 none 视为客观标签，最终从 3 类数据中各抽取 5000 条用于测试情感三分类任务。对于 SMP2020-EWECT 数据集，本文将其中的通用微博数据集和疫情微博数据集进行合并，最终从 6 类标签中各抽取 2000 条用于测试情感六分类任务。实验数据集信息详见表 2。

表 2 实验数据集概述

Table 2 Overview of experimental datasets

数据集	数量	类别数	平均长度
SWB	10000	2	31.17
NLPCC2014	15000	3	18.19
SMP2020-EWECT	12000	6	28.65

源数据集中存在与情感分析无关的冗余信息，如网站链接、“@”、冗余的标点符号等。为此，本文使用工具 HarvestText 和 Pyhanlp 对其过滤删除。在此过程中，一些与情感色彩判定相关的标点符号将会被保留，如“!”“?”。对于预处理后的文本数据，本文使用 Pyltp 对其进行分词。考虑到学术界中大多认为短文本长度不超过 140[18]，本文在对文本数据进行预处理时将针对过短文本采取“0”填充操作，针对过长文本采取截断操作，最终保证分词后$|D|$的值为 140。另外，下文实验将按照 8∶1∶1 的比例进行数据集划分。

(3)模型参数设置。CDSI 模型的参数设置详见表 3。为方便比较，关于对比模型与 CDSI 模型中所使用的词嵌入模块，本文将统一使用基于微博语料预训练好的 Word2vec[24]模型[4)]。

表 3 CDSI 模型的参数设置

Table 3 Parameters setting of CDSI model

模块	参数/超参数	值/描述
Word embedding	size(per word)	300
Bi-LSTM	units	150
	layers	2
	dropout	0.5
Graph embedding	size(per sentence)	140×140
GCN	channels	200
	layer	1
Training	loss	cross-entropy
	L2 regularization	1×10^{-8}
	optimizer	Adam
	batch size	32
	learning rate	1×10^{-3}

4.2 评价指标

本文所研究的问题本质上为单标签多分类任务。考虑到 4.1 节中的实验数据准备阶段一定程度上避免了数据不平衡问题，后续实验将使用准确率(Accuracy，Acc)和宏平均 F1 值(Macro-F1，MF1)来评估模型性能。对于情感类别为 i 的样本集合，其相应的$F1_i$值计算公式描述如下：

$$Precision=\frac{TP}{TP+FP} \tag{12}$$

$$Recall=\frac{TP}{TP+FN} \tag{13}$$

$$F1_i=\frac{2\times Precision_i\times Recall_i}{Precision_i+Recall_i} \tag{14}$$

对于所有样本，$MF1$ 和 Acc 的计算公式为：

1) https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k

2) http://tcci.ccf.org.cn/conference/2014/pages/page04_dg.html

3) https://smp2020ewect.github.io/

4) https://github.com/Embedding/Chinese-Word-Vectors

$$MF1=\frac{\sum_{i=1}^{n}F1_i}{n} \tag{15}$$

$$Acc=\frac{\sum_{i=1}^{n}TP_i+TN_i}{TP_i+FN_i+FP_i+TN_i} \tag{16}$$

其中，$TP+TN$ 表示被正确预测的样本数量，FP 指其他类别被预测为当前类别的样本数量，FN 指当前类别被预测为其他类别的样本数量。

4.3 与相关工作比较

为验证本文所提出模型的有效性，选取以下 5 个模型作为基线。

(1)CNN[4]：基于经典的卷积神经网络提取文本局部特征信息，随后输入 softmax 层进行分类。

(2)BiLSTM[6]：利用 BiLSTM 提取短文本局部特征信息，然后连接 softmax 层进行情感分类。

(3)CNN-BiLSTM[25]：利用 CNN 和 BiLSTM 分别提取文本局部特征以及上下文全局特征，然后将两种特征融合后输入 softmax 层进行情感分类。

(4)BiLSTM-GCN[18]：首先使用 BiLSTM 抽取文本特征，然后在构造邻接矩阵时使用 1 和 0 来编码语句各成分之间的依赖关系，最后将单词的向量表示和邻接矩阵共同输入到 GCN 进行情感分类。

(5)BiLSTM-ATT-GCN[19]：该模型使用 BiLSTM 抽取文本特征，随后在此基础上引入自注意力机制丰富语句的情感信息，其余模块与 BiLSTM-GCN 模型保持一致。

本文基于 SWB，NLPCC2014 和 SMP2020-EWECT 数据集对上述模型以及 CDSI 模型进行了测试，具体的实验结果如表 4 所列。

表 4 短文本情感分类实验结果对比

Table 4 Experimental results comparison of short text sentiment classification

(单位：%)

模型	SWB		NLPCC2014		SMP2020-EWECT	
	Acc	*MF*1	*Acc*	*MF*1	*Acc*	*MF*1
CNN	92.34	92.11	64.72	64.66	61.57	61.83
BiLSTM	93.25	93.01	66.02	65.73	62.43	62.31
CNN-BiLSTM	93.72	93.98	67.13	66.52	63.69	63.55
BiLSTM-GCN	94.19	94.23	67.73	67.89	64.61	64.25
BiLSTM-ATT-GCN	94.47	94.45	68.63	68.71	64.89	64.86
CDSI	**95.22**	**95.18**	**71.13**	**70.88**	**66.91**	**66.66**

从表 4 可以看出，本文所提出的 CDSI 模型在 3 个短文本数据集上的表现均优于基线。相较于经典的 CNN 网络，BiLSTM 能够更好地学习到文本中的上下文信息。特别是对于短文本数据，其自身特征较为稀疏，因此上下文信息中所蕴含的长距离依赖、双向语义依赖关系等对文本情感色彩的判定更为关键。前 3 个模型的预测效果表明了文本的上下文信息对短文本情感分类的重要性。依存句法信息是一种类似于图结构的数据，其可以作为一种先验知识辅助模型深入理解自然语言。表 4 中关于后 3 个模型的实验结果，均能说明该信息对短文本情感分析具有重要的影响。相较于 BiLSTM-GCN，BiLSTM-ATT-GCN 通过引入自注意力机制增强了文本表示中的情感信息，在 3 个数据集上均表现出了一定的优势。与 BiLSTM-GCN 和 BiLSTM-ATT-GCN 相比，本文提出的 CDSI 模型针对上下文和依存句法信息同时建模，同时还考虑了单词之间的依赖类型以及相应的单词语义，最终在短文本情感二分类以及多分类中均取得了较好的效果。

4.4 依赖类型嵌入向量维度对模型性能的影响

本文设计实现了一种基于依存关系感知的嵌入表示方法，其本质在于同时考虑了单词之间的依赖类型以及相应的单词语义信息，这将有利于 CDSI 模型更多地关注对情感分类任务有益的依赖路径。对于中文短文本，句法依赖类型 φ 的取值范围在表 1 中已描述。而在针对依存句法信息进行嵌入编码时，相应的 $\wp_\varphi$ 表示依存句法依赖类型嵌入向量，其维度值是一个可自定义的参数。本小节探究了 $\wp_\varphi$ 的维度值对模型性能的影响，如表 5 所列。可以看出，在不同的依赖类型嵌入维度下，CDSI 模型表现较为稳定。而对于本文所使用的 3 个短文本数据集，当 $\wp_\varphi$ 维度值为 25 时，模型性能最佳。

表 5 依赖类型嵌入向量维度对模型的影响

Table 5 Impact of different dimensions of dependent type embedding on model

(单位：%)

维度	SWB		NLPCC2014		SMP2020-EWECT	
	Acc	*MF*1	*Acc*	*MF*1	*Acc*	*MF*1
5	94.84	94.72	69.64	69.39	64.99	64.86
15	95.01	94.86	70.72	70.54	65.78	65.65
25	**95.22**	**95.18**	**71.13**	**70.88**	**66.91**	**66.66**
35	95.14	95.11	70.89	70.61	65.86	65.72

4.5 针对依存句法信息嵌入的可视化分析

短文本具有特征稀疏、表达信息能力不足等特点，因此文本中各成分之间的句法依赖关系以及单词语义信息对情感色彩的判定影响较大。为更直观地观察依存句法信息嵌入所发挥的作用，本节将根据 CDSI 模型在测试集上的实际预测效果来统计每个依赖类型所对应的分值，随后求平均值并进行可视化。如图 4 所示，模型对不同依赖关系的重视程度各异。其中，HED 和 WP 依赖类型的分值最高，表明语句中核心词以及标点符号对情感色彩的判定较为重要。POB，RAD 和 LAD 依赖类型通常与语句中的名词相关，而名词在短文本语句中通常作为情感的主体或者客体，在一定程度上表明 CDSI 模型将情感特征与句子成分进行了关联融合，有利于针对整个文本进行情感分析。同理，POB 和 IOB 通常与语句中的宾语也即动作的受体相关，将情感特征与宾语成分进行关联同样有益于结果分析。此外，一些不重要的依赖类型，如 CMP 和 COO 等，与语句情感色彩的相关度较低，因此最终的嵌入分值较小。

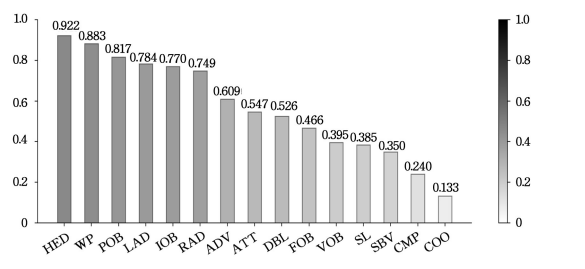

图 4 不同依赖类型所对应的嵌入分值

Fig. 4 Embedding scores corresponding to different dependency types

以正向评论“非常感谢老师的悉心指导！”为例，本小节进行了案例分析，其依存句法信息嵌入结果如图 5 所示。可以看到，“感谢”为核心词汇，HED 依赖类型所对应的嵌入分值为 0.96；“！”为标点符号并与核心词汇之间存在依赖关系，其 WP 依赖类型嵌入分值为 0.85；“老师”作为“指导”的定语，所对应的依赖类型 ATT 嵌入分值为 0.81；而对于“指导”与“感谢”之间的依赖类型 VOB，其对应的嵌入分值仅为 0.38。从语言学的角度看，“感谢”“老师”“指导”以及“！”这 4 个句子成分足以表达整个文本的语义，同时能够决定整个语句的情感极性为正向。结合依存句法信息嵌入结果来看，CDSI 模型对“Root”→“感谢”、“感谢”→“！”以及“老师”→“指导”的句法依赖路径关注较高，这表明本文提出的嵌入表示方法有利于模型挖掘出不同依赖路径对情感分类任务的贡献权重。

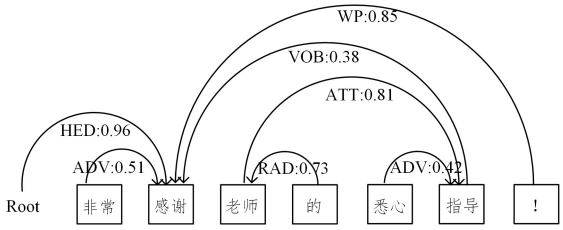

图 5 案例分析

Fig. 5 Case study

4.6 文本长度对预测效果的影响

在 CDSI 模型中，语义提取层、依存句法信息嵌入层均能够克服文本的长距离依赖，并分别从语义和语言学的角度挖掘文本的深层情感特征，有益于最终的情感分析任务。为证实这一特性，本小节将探究文本长度对 CDSI 模型性能的影响。以 NLPCC2014 数据集为例，本次实验从中随机抽取 1 000 条长度各异的文本进行模型测试，结果如图 6 和图 7 所示。

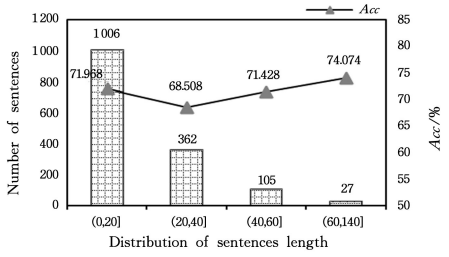

图 6 句子长度对 *Acc* 值的影响

Fig. 6 Impact of sentence length on *Acc*

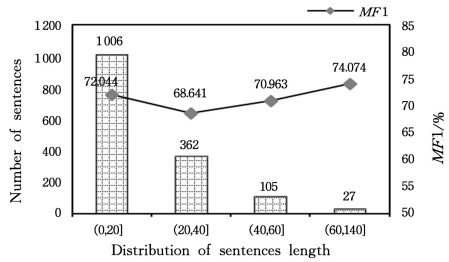

图 7 句子长度对 *MF*1 值的影响

Fig. 7 Impact of sentence length on *MF*1

从图 6 和图 7 可以看到，模型在(0，20]，(40，60]，(60，140]范围语句上的预测效果均高于整体的平均线（见表 4），特别是从(20，40]范围开始，其预测效果随着语句长度的增加呈上升趋势。结果表明，本文提出的 CDSI 模型有效融合了依存句法信息，即使句子长度增加，模型仍能较好地理解语句中各成分之间的关系；此外，模型还考虑了文本的上下文信息，因此在应对长句子时仍表现出较好的情感分析效果。

结束语 本文提出了一种结合上下文和依存句法信息的中文短文本情感分析模型 CDSI，其主要由语义提取层、依存句法信息嵌入层、图卷积层和情感分类层 4 部分组成。首先，利用词嵌入技术将短文本表示为低维稠密向量，并将其输入到 BiLSTM 中学习上下文信息；其次，对文本进行依存句法分析，综合考虑语句各成分之间的依赖类型以及单词语义以实现依存句法信息嵌入；然后，将前两层的输出分别作为节点特征和边特征输入图卷积层计算，得到文本表征信息；最后，将该信息输入情感分类层，实现短文本情感分类。本文在 3 个开源数据集 SWB，NLPCC2014 和 SMP2020-EWEC 上进行了一系列的实验，结果表明本文提出的 CDSI 模型能够充分挖掘句法结构中不同依赖路径对情感分析的贡献权重，同时针对语句上下文以及句法结构共同建模实现了信息的有效融合，在中文短文本情感二分类以及多分类任务中的表现均优于基线。

相较于现有的相关工作，本文构建了一个更加完善的中文句法依赖类型集合，并首次探索了利用图卷积网络针对上下文语义、句法依赖关系类型以及与该依赖关系相关的单词语义信息同时建模的可能性，其后续的实验结果表明了 CDSI 模型在中文短文本情感分类任务中的优越性。然而，该模型也存在一定的局限性：1）模型中的依存句法信息嵌入层依赖于 LTP 工具针对文本数据的分析结果，其可能出现的分析误差会传播至后续的图卷积层以及情感分类层，影响最终的分类效果；2）模型使用 Word2vec 进行词嵌入表示，无法解决一词多义问题，并且其本身包含的语义信息较为简单，限制了其应用范围。

未来有以下问题值得进一步研究：1）改进文本的语义表示，尝试融入更多情感特征，提高情感分类效果；2）改进依存句法分析工具，设法减小分析误差；3）结合中、英文依存句法分析特点，尝试提出一种通用的结合依存句法信息的短文本情感分类方法。

参 考 文 献

[1] ZHANG Y，XU H，XU K. Chinese Short Text Classification based on Dependency Syntax Information[C]// ICCDA 2021：The 5th International Conference on Compute and Data Analysis. Sanya：ACM，2021：133-138.

[2] LI C B，DUAN Q J，JI C H，et al. Method of Short Text Classification Based on CHI and TF-IWF Feature Selection [J]. Journal of Chongqing University of Technology(Natural Science)，2021，35(5)：135-140，222.

[3] QIU X，SUN T，XU Y，et al. Pre-trained Models for Natural Language Processing：Asurvey[J]. Science China Technological Sciences，2020，63(10)：1-26.

[4] KIM Y. Convolutional Neural Networks for Sentence Classifica-

tion[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Proceeding. Doha: ACL, 2014: 1746-1751.
[5] LENG X L, MIAO X A, LIU T. Using Recurrent Neural Network Structure with Enhanced Multi-Head Self-Attention for Sentiment Analysis[J]. Multimedia Tools and Applications, 2021, 80(8): 12581-12600.
[6] XU G, MENG Y, QIU X, et al. Sentiment Analysis of Comment Texts based on BiLSTM[J]. IEEE Access, 2019, 7: 51522-51532.
[7] XIAO H, XU S H. Analysis on Web Public Opinion Orientation based on Syntactic Parsing and Emotional Dictionary[J]. Small Microcomputer System, 2014, 35(4): 811-813.
[8] LI X H, GUO H, YAN H T. Micro-blog Sentiment Analysis based on Improved DependencyParsing[J]. Computer and Digital Engineering, 2017, 45(3): 506-511.
[9] WANG C, WANG B, XIANG W, et al. Encoding Syntactic Dependency and Topical Information for Social Emotion Classification[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019: 881-884.
[10] TANG H, JI D, LI C, et al. Dependency Graph Enhanced Dual-Transformer Structure for Aspect-based Sentiment Classification[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 6578-6588.
[11] ZHANG M, LI Z, FU G, et al. Dependency-based Syntax-Aware Word Representations[J]. Artificial Intelligence, 2021, 292(4): 103427.
[12] GUO Z, ZHANG Y, LU W. Attention Guided Graph Convolutional Networks for Relation Extraction[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 241-251.
[13] ZHANG B, ZHANG Y, WANG R, et al. Syntax-Aware Opinion Role Labeling with Dependency GraphConvolutional Networks [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 3249-3258.
[14] WANG J H, WANG H H, WANG L. Dependency Parsing of Financial News to Improve Sentiment Analysis for Predicting Market Prices[C]// International Conference on Technologies and Applications of Artificial Intelligence. Taipei: IEEE, 2020: 1-7.
[15] ZHANG X S, GUO R Q, HUANG D G. Named Entity Recognition Based on Dependency[J]. Journal of Chinese Information Processing, 2021, 35(6): 63-73.
[16] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[J]. arXiv: 1609.02907, 2016.
[17] SUN K, ZHANG R C, MENSAH S, et al. Aspect-Level Sentiment Analysis Via Convolution over Dependency Tree[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 5679-5688.
[18] LAI Y, ZHANG L, HAN D, et al. Fine-Grained Emotion Classification of ChineseMicroblogs based on Graph Convolution Networks[J]. World Wide Web, 2020, 23(5): 2771-2787.
[19] FAN T, WANG H, WU P. Sentiment Analysis of Online Users' Negative Emotions based on GraphConvolutional Neural Network and Dependency Parsing [J]. Data Analysis and Knowledge Discovery, 2021, 5(9): 97-106.
[20] PARK J, PARK C, KIM J, et al. ADC: Advanced Document Clustering Using Contextualized Representations[J]. Expert Systems with Applications, 2019, 137: 157-166.
[21] CHE W, LI Z, LIU T. LTP: A Chinese Language Technology Platform[C]// COLING 2010, 23rd International Conference on Computational Linguistics. Beijing: Demonstrations Volume, 2010: 13-16.
[22] MARCHEGGIANI D, TITOV I. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 1506-1515.
[23] KARAMI M, MOSALLANEZHAD A, MANCENIDO M V, et al. "Let's Eat Grandma": When Punctuation Matters in Sentence Representation for SentimentAnalysis[J]. arXiv: 2101.03029, 2020.
[24] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and Their Compositionality[J]. Advances in Neural Information Proces-sing Systems, 2013, 26(5): 3111-3119.
[25] LI Y, DONG H B. Text Sentiment Analysis based on Feature Fusion of Convolution Neural Network and Bidirectional Long Short-Term Memory Network[J]. Computer Applications, 2018, 38(11): 3075-3080.

DU Qiming, born in 1998, postgraduate. His main research interests include big data analysis, natural language processing and so on.

LI Nan, born in 1977, Ph. D, associate professor. His main research interests include high-performance computing, big data analysis, big data security and so on.

(责任编辑:柯颖)