文章编号:1008-1534(2024)02-0092-07

基于结构-语义融合的评论文本情感分类研究

马艳珍[1],勾智楠[2],池云仙[1],高 凯[1]

(1. 河北科技大学信息科学与工程学院,河北石家庄 050018;2. 河北经贸大学信息技术学院,河北石家庄 050061)

摘 要:为了解决当前部分情感分析模型片面依赖文本语义特征、忽视句法结构特征的问题,提出了一种基于结构-语义融合的情感分类模型 SF-TLSTM(sentiment-fusion-tree LSTM),即将句法结构信息与语义信息进行融合,全面提取文本特征。首先,将 BERT(bidirectional encoder representation from transformers)模型引入 TreeLSTM(tree-structured bidirectional LSTM)网络结构中;其次,利用 SimCSE(simple contrastive learning of sentence embeddings)模型的自监督训练对 BERT 获得的向量表示进行数据增强;最后,通过节点编码的方式在构建的 TreeLSTM 网络上实现结构语义特征融合,并与基线模型进行多组对比分析。结果表明:在斯坦福大学发布的 SST(stanford sentiment tree-bank)数据集上,SF-TLSTM 模型相较于经典树形结构情感分类模型获得更高的准确率,其中在二分类任务中的准确率达到了 96.79%。所提方法能够全面有效地提取评论文本的特征,增强文本的向量表示,对自然语言处理领域中的文本处理具有重要意义。

关键词:自然语言处理;情感分类;语义融合;预训练模型;句法结构

中图分类号:TP391.1 **文献标识码**:A **DOI**:10.7535/hbgykj.2024yx02002

A study on sentiment classification of commentary text based on structural-semantic fusion

MA Yanzhen[1], GOU Zhinan[2], CHI Yunxian[1], GAO Kai[1]

(1. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China; 2. School of Information Technology, Hebei University of Economics and Business, Shijiazhuang, Hebei 050061, China)

Abstract: To solve the problem that some current sentiment analysis models unilaterally rely on textual semantic features and neglect syntactic structural features, a sentiment classification method based on structural-semantic fusion was proposed. The method fused syntactic structural information with semantic information to comprehensively extract text features. The BERT (Bidirectional Encoder Representation from Transformers) model was introduced into the TreeLSTM (Tree-structured bidirectional LSTM) network structure. Then the data enhancement of the vector representation obtained by BERT was performed by utilizing the SimCSE (Simple Contrastive Learning of Sentence Embeddings) model's self-supervised training.

收稿日期:2023-10-10;修回日期:2024-02-28;责任编辑:王淑霞

基金项目:2024 年度人文社会科学研究重大课题攻关项目(ZD202402)

第一作者简介:马艳珍(1997—),女,河北张家口人,硕士研究生,主要从事自然语言处理方面的研究。

通信作者:高凯教授。E-mail: gaokai@hebust.edu.cn

马艳珍,勾智楠,池云仙,等. 基于结构-语义融合的评论文本情感分类研究[J]. 河北工业科技,2024,41(2):92-98.

MA Yanzhen, GOU Zhinan, CHI Yunxian, et al. A study on sentiment classification of commentary text based on structural-semantic fusion[J]. Hebei Journal of Industrial Science and Technology, 2024, 41(2): 92-98.

Finally the structural semantic feature fusion was realized on the constructed TreeLSTM network by means of node encoding, and was analyzed in a multi-group comparison with the baseline model. The experimental results show that on the SST (Stanford Sentiment Tree-bank) dataset released by Stanford University, the structural-semantic fusion-based sentiment classification method obtains higher accuracy compared to the classical tree-structured sentiment classification model, with an accuracy rate of 96.79% in the binary classification task. The proposed method can comprehensively and effectively extract the features of the comment text and enhance the vector representation of the text, which is important for text processing in the field of natural language processing.

Keywords: natural language processing; sentiment classification; semantic fusion; pre-trained model; syntactic structure

情感分析是对文本信息中蕴含的情感信息进行分析，可划分为积极、消极或中立等不同的情感类别，输出情感标签。早期的情感分析方法通常依赖于手动设计的特征和分类器。基于机器学习的情感分析方法多数是基于 SVM 算法和最大熵算法进行研究的，需要大量的人工参与[1-2]。随着深度学习技术的发展，卷积神经网络(CNN)和递归神经网络(RNN)可以进一步提取文本特征用于文本分类[3-4]。部分研究人员将普通序列结构转换为树状结构，将句法结构特征纳入文本的表示学习中[5-6]。TreeGRU 模型根据句法结构建模，通过门控机制处理单元内的信息流，用 GRU 单元取代了 LSTM 单元[7]。MOU 等[8]提出了一种以并行化方式提取成分树上的特征模型 TBCNN。ZHU 等[9]提出的 S-LSTM 可以代替递归神经网络中的语义组合层，从而更好地学习不同文本片段组合之后的文本语义表征。TAI 等[10]提出的 Constituency Tree-LSTMs 设计句子依存树，从多个孩子节点中获取语义信息。LE 等[11]提出的 LSTM-RNN 可以进一步获取文本长距离依赖的信息。现阶段的研究倾向于对先进的预训练模型进行微调或改进。目前较为通用的预训练模型有：BERT、ELMo、XL-NET、ALBERT 等[12-14]。预训练模型在向下游任务迁移时可以产生高质量的文本表示。相比于预训练模型获取的词嵌入，句子嵌入考虑了更多的语义和上下文信息。CONNEAU 等[15]提出 InferSent 模型，基于 7 个不同的网络架构研究使用监督数据进行句子嵌入训练。Sentence-BERT(SBERT)[16]采用孪生网络和三重网络结构训练句子嵌入，用于 Reddit 对话生成任务。SimCSE[17]通过比较类似话语的对比学习，通过自监督训练和无监督训练，对文本嵌入表示进行改善。

尽管在情感分类领域已有大量研究成果，然而基于序列结构的方法在表示文本信息时片面提取语义特征，忽略了句法结构对句子含义的影响。而基于树形结构建模的方法则在长距离编码时会逐渐丢失底层单词的语义信息。因此，本文提出一种基于结构-语义融合的情感分类模型 SF-TLSTM(sentiment-fusion-tree LSTM)。通过无监督数据增强方法，simCSE 模型对文本嵌入表示进行改善，以节点编码的方式将句法结构与语义特征有机融合，实现对文本特征地全面提取。此外，将句子拆分为短语组合，逐级获取节点信息、增强节点记忆、提高文本情感分析能力。

1　混合神经网络模型 SF-TLSTM

1.1　任务定义

情感分析任务可以定义为一个分类问题，即输入一个文本序列 X，机器通过对文本进行分析、处理、归纳和推理后自动输出不同的情感标签，本文定义文本输入序列 $X=\{x_1,x_2,\cdots,x_l\}$，其中 l 为序列长度。经模型计算生成情感预测标签 $\hat{y}_l$。情感预测包含五分类和二分类 2 种任务，五分类任务情感预测标签分为非常消极、消极、中立、积极和非常积极 5 种结果，二分类任务简化为积极和消极 2 种结果。

1.2　模型结构

本文提出一种基于结构-语义融合的情感分类混合模型 SF-TLSTM，如图 1 所示。模型由文本嵌入、自监督训练、结构语义特征融合、情感分类预测 4 部分组成。首先，采用 BERT 对文本进行向量化表示；其次，利用无监督数据增强方法 SimCSE 对文本嵌入表示进行改善；再次，根据句法结构将文本序列解析为成分树，以节点编码的方式与语义特征实现有机融合；最后，使用 TreeLSTM 算法进行计算，输出离散的情感标签。

1.2.1　文本嵌入

BERT 模型是基于 Tranformer 的深度双向语言表征模型，即利用 Transformer 结构构造了一个多层双向的 Encoder 网络[12]。BERT 模型所有网络层均联合上下文语境进行预训练，能够获得上下文的双向特征表示，使训练得到的词嵌入能更好地用于下游任务。因此，采用 BERT 模型获取文本序

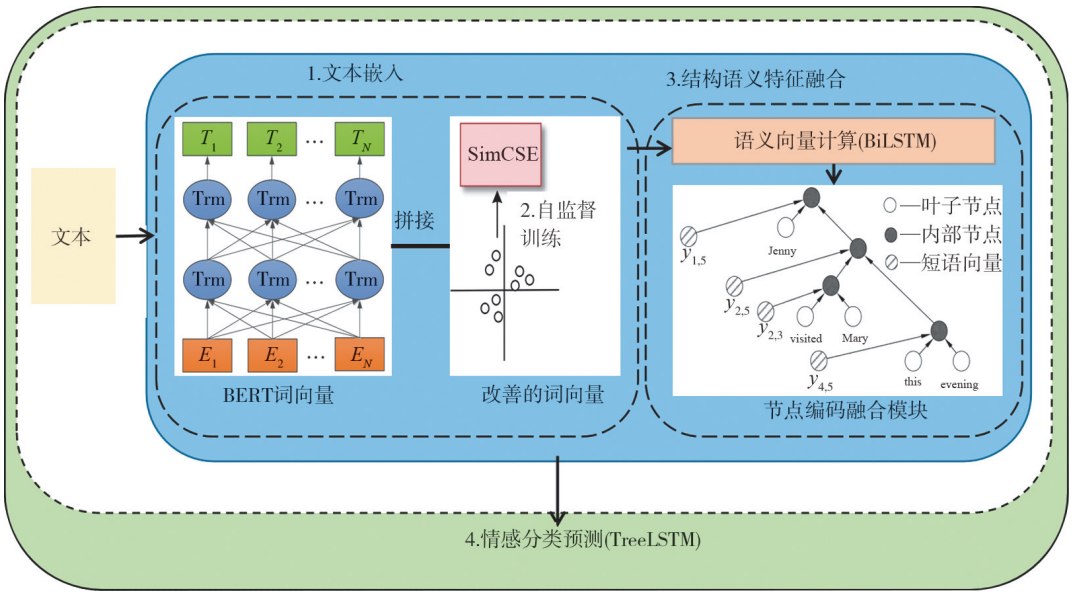

图 1 SF-TLSTM 模型框架图

Fig. 1 SF-TLSTM model framework

列初步词向量 w_j。然而，BERT 模型得到的词嵌入表示在空间呈锥形不均匀分布[18-19]，高频词分布在狭小的区域，靠近原点，低频词训练不充分，分布相对稀疏，远离原点。尤其在针对分类任务时 BERT 模型获得的[CLS] token 所包含的句子信息往往稀疏不均，效果难以让人满意。

针对上述问题，本文引入对比学习思想使用 SimCSE 模型进行数据增强，并在训练过程中以自监督的方式训练模型。实验在 huggingface 网站直接调用封装好后的 SimCSE 对经由 BERT 初步处理后的词向量进行处理，从而得到更好的词向量表示 e_j。

1.2.2 数据增强

本文采用将 SimCSE[17] 进行数据增强。SimCSE 模型能够利用预训练语言模型和 dropout 机制的随机性，在无监督的情况下学习文本嵌入。它对同一文本多次前向传播得到不同的嵌入表示，并使用对比损失函数拉近这些表示的距离，以此来增强模型对相似文本的识别能力，实现数据增强。其训练损失函数公式如式 (1)、式 (2)所示。

$$L_i = -\log \frac{e^x}{\sum_{j=1}^{N} e^x}, \tag{1}$$

$$x = \mathrm{sim}(h_i, h_j^+)/\tau, \tag{2}$$

式中：τ 为温度系数，用来调节模型对负样本的关注度，实验设置与文献[17]相同；$\mathrm{sim}(h_i, h_j^+)$ 是样本间的余弦相似度。

1.2.3 结构语义特征融合

如何将句法结构信息与文本语义信息有效的融合是本模型的核心部分，本模型将结构语义特征融合分为 2 个模块：语义向量计算模块和节点编码融合模块。语义向量模块负责对上一节中得到的向量表示进行编码计算；节点编码融合模块负责借助树形结构网络特点以节点编码方式逐步将句法结构信息融合。

1) 语义向量计算模块　对于输入序列 $X=\{x_1, x_2, \cdots, x_l\}$，首先，通过 BERT 模型和 simCSE 模型分别训练，分别获取文本的词向量表示；然后，使用双向长短时记忆网络 BiLSTM 对其进一步编码，计算公式如式(3)－式(5)所示。

$$\overrightarrow{h_j} = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{h_{j-1}}, [w_j, e_j]), \tag{3}$$

$$\overleftarrow{h_j} = \overleftarrow{\mathrm{LSTM}}(\overleftarrow{h_{j-1}}, [w_j, e_j]), \tag{4}$$

$$h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}], \tag{5}$$

式中：w_j 是经由 BERT 预先训练好的词向量；e_j 是由 SimCSE 改善得到的词向量，文献[20]指出二者拼接作为相应文本单词的语义表示时效果最好；$\overrightarrow{h_j}$ 为前向 LSTM 的隐藏状态；$\overleftarrow{h_j}$ 为后向 LSTM 的隐藏状态，通过对前一步隐藏状态和文本单词的网络编码计算得到；h_j 为正反双向隐藏状态二者的拼接，表示文本中第 j 个单词的语义向量。

为了增强节点的记忆，本文采用文献[20]的方法，将句子看作是短语的组合，对短语的语义向量进行计算。句子中的任意短语 $s_{m,n}$ 的语义向量计算如式(6)所示。

$$s_{m,n} = [\overrightarrow{h_m}; \overleftarrow{h_n}]。 \tag{6}$$

由于短语中包含 2 个或 2 个以上单词，所以采用拼接单词隐藏状态的方式表示短语的语义向量。其中：$\overrightarrow{h_m}$ 表示前向 BiLSTM 网络中第 1 个单词的

隐藏状态；$\overleftarrow{h}_n$ 为反向 BiLSTM 网络最后 1 个单词的隐藏状态。

2）节点编码融合模块　TreeLSTM 是一种树状的递归神经网络模型，它能够将文本序列根据句法结构信息处理成树形数据，并根据不同的句法结构捕捉不同层次的结构信息。在 TreeLSTM 模型中，树形的结构形式分为父节点和孩子节点，每个节点将其左右孩子节点的表示作为输入，通过将左右孩子节点的表示加权求和的结果向上传递给父节点。以“John hit the ball”为例，这句话被解析为名词 John 加动词短语 hit the ball，而动词短语里嵌套动词 hit 加名词短语 the ball，名词短语中包含冠词 the 和名词 ball，如图 2 所示。

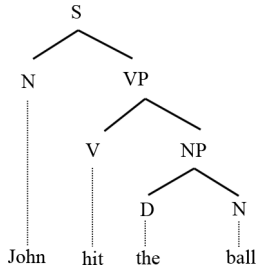

图 2　句法结构成分树

Fig. 2　Syntactic structure component tree

首先，根据文本序列名词 John 加动词短语 hit the ball 的句法结构将其解析为成分树；其次，关注各个内部节点，包括根节点，动词短语节点及名词短语节点，对这 3 个短语节点经过语义向量计算模块进行计算得到短语语义向量 $s_{m,n}$，注入相应内部节点位置。第 2 层内部节点 NP(名词短语)表示短语 the ball，在计算此节点记忆单元传递信息前需要注入 the ball 的语义向量 $s_{3,4}$，第 3 层节点 VP(动词短语)表示的短语是 hit the ball，在计算此节点记忆单元时注入 hit the ball 的语义向量 $s_{2,4}$。以此类推，通过节点编码的方式，以句法结构选择相应节点语义向量，将句子的语义向量逐步输入编码计算，实现在句子建模时对文本语义和句法结构的兼顾，从而获得完整的文本特征。节点编码计算如图 3 所示。

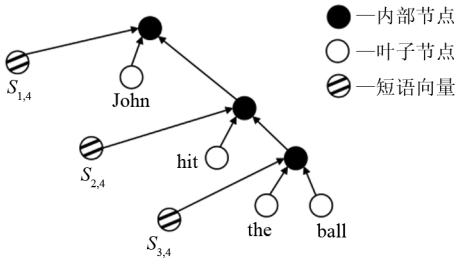

图 3　节点编码计算结构

Fig. 3　Node code fusion module

模型针对句子的句法结构成分树构建 Tree-LSTM 网络，相比于 LSTM，以树节点的形式表示 LSTM 的单元。LSTM 单元中任意节点的隐藏状态 h_t，在树状的 LSTM 网络中被分解为左孩子节点的隐藏状态 h_{t-1}^l 和右孩子节点的隐藏状态 h_{t-1}^r，由左右孩子节点的隐藏状态加权求和向上传递信息。同样，节点的记忆单元 c_t 被分解为左孩子节点的记忆单元 c_{t-1}^l 和右孩子节点的记忆单元 c_{t-1}^r。而叶子节点用来表示单个单词，它们不存在遗忘门。句法结构成分树的输入自底向上，从叶子节点处直接输入各单词的词嵌入向量。

具体计算时，将语义向量模块计算得到的 $s_{m,n}$ 与句法结构成分树中原有门机制中的公式相结合，内部节点的记忆单元 c_t 计算公式如式(7)所示，t 为节点在句法结构成分树中的索引。

$$c_t = f_t^l \otimes c_{t-1}^l + f_t^r \otimes c_{t-1}^r + i_t \otimes \widetilde{c_t}\ , \tag{7}$$

其中，

$$\widetilde{c_t} = \tan h\left(W_{s_{\tilde{c}}}^D s_{m,n} + \sum_{D\in\{l,r\}} W_{h_{\tilde{c}}}^D h_{t-1}^D + b_{\tilde{c}}\right)\ , \tag{8}$$

$$i_t = \sigma\left(W_{s_i}^D s_{m,n} + \sum_{D\in\{l,r\}} \left(W_{h_i}^D h_{t-1}^D + W_{c_i}^D c_{t-1}^D\right) + b_i\right), \tag{9}$$

$$f_t^l = \sigma\left(W_{s_l}^D s_{m,n} + \sum_{D\in\{l,r\}} \left(W_{f_l}^D h_{t-1}^D + W_{c_l}^D c_{t-1}^D\right) + b_{f_l}\right), \tag{10}$$

$$f_t^l = \sigma\left(W_{s_r}^D s_{m,n} + \sum_{D\in\{l,r\}} \left(W_{f_r}^D h_{t-1}^D + W_{c_r}^D c_{t-1}^D\right) + b_{f_r}\right)。 \tag{11}$$

输出门的结果 o_t、当前节点的隐藏状态 h_t，计算公式如式(12)、式(13)所示。

$$o_t = \sigma\left(W_{s_o}^D s_{m,n} + \sum_{D\in\{l,r\}} \left(W_{h_o}^D h_{t-1}^D + W_{c_o}^D c_t\right) + b_o\right), \tag{12}$$

$$h_t = o_t \otimes \tan h\left(c_t\right)\ , \tag{13}$$

式(7)—式(12)中：tan 为正切激活函数；c_t 为当前记忆单元；$W_{s_{\tilde{c}}}^D$、$W_{h_{\tilde{c}}}^D$、$b_{\tilde{c}}$、$W_{s_i}^D$、$W_{h_i}^D$、$W_{c_i}^D$、b_i、$W_{s_l}^D$、$W_{f_l}^D$、$W_{c_l}^D$、b_{f_l}、$W_{s_r}^D$、$W_{f_r}^D$、$W_{c_r}^D$ b_{f_r}、$W_{s_o}^D$、$W_{h_o}^D$、$W_{c_o}^D$ 及 b_o 均为模型参数。

1.2.4　情感分类预测

SF-TLSTM 模型需要从离散的标签集合 y_t 中预测最终的情感标签 $\hat{y}_t$，在每个节点 t 处，使用 softmax 分类器完成一次分类计算，预测输入序列 X 时的情感标签 $\hat{y}_t$，分类器将该节点的隐藏状态 h_t 作为输入。如式(14)、式(15)所示。

$$\hat{p}_\theta(y \mid \{x\}_t) = \mathrm{softmax}(w_t h_t + b_t)\ , \tag{14}$$

$$\hat{y}_t = \underset{y}{\mathrm{argmax}}\,\hat{p}_\theta(y \mid \{x\}_t)\ , \tag{15}$$

式中：w_t、b_t 均为模型可学习参数。

1.3 模型训练

训练过程中，本模型对训练参数 θ 的目标函数如下。

$$L(\theta)=-\frac{1}{S}\sum_{S=1}^{S}y_t\log\hat{y}_t+\frac{\lambda}{2}\|\theta\|^2, \quad (16)$$

式中：S 是数据集中训练样本的数目；λ 是 L2 正则化超参数。

2 实验设计

2.1 数据集和评价指标选取

为了测试 SF-TLSTM 模型的性能，实验选取斯坦福大学发布的权威数据集 SST(stanford sentimenttree-bank)[21]，有 2 个版本可用。一个具有细粒度标签(五分类)，对应 5 种情绪类别：非常消极、消极、中立、积极和非常积极。另一个为二进制标签(二分类)，去除标记为中立的样本，对应积极和消极 2 种情绪类别。该数据集由 11 855 条电影评论组成，对于细粒度标签下的五分类任务划分为 8 544 个训练样本、1 101 个开发样本和 2 210 个测试样本，分别作为训练集、验证集和测试集，对二分类任务，训练集、验证集和测试集大小分别为 6 920、872 和 1 821。实验采用准确度指标评估模型的有效性。

2.2 对比基线模型和参数设置

为证明 SF-TLSTM 模型在情感分类上表现出的良好性能，选取以下方法作为对比方法。

1) BiLSTM 方法　基于简单加权求和模型进行情感分类[22]，该方法引入了对上下文敏感的词库，使用 BiLSTM 模型来学习词库情感的归纳强度、强化和否定，从而进行情感分类。

2) S-LSTM 方法　将 LSTM 扩展为树形结构，可以视为递归神经网络和循环神经网络二者的结合，利用 S-LSTM 来代替递归神经网络中的语义组合层，从而更好地学习不同文本片段组合之后的文本语义表征[9]。与先前的循环神经网络相比，S-LSTM 方法有避免梯度消失的潜力，通过树结构对长距离交互进行建模。

3) Constituency Tree-LSTMs 方法　解决了 LSTM 无法对句子的结构信息进行提取等问题，提出了树形的 LSTM 模型[10]。实验使用其中的 Constituency Tree 部分模型进行对比。

4) LCT-LSTM 方法　在 Constituency Tree-LSTMs 方法基础上，需要处理 2 个方向的节点编码过程[24]。

实验使用 PyTorch 深度学习开发框架，用 Python 作为开发语言，GPU 为 NVIDIA GeForce RTX 3090，使用 Stanford PCFG Parser 解析输入的序列。在构建模型时，将隐藏层维度设置为 150、输入维度设置为 300、学习率设置为 0.01、批量大小设置为 25，为防止过拟合，L2 正则化系数设置为0.000 1。

2.3 基于预训练模型对比实验设计

SST 数据集中的文本数据形式不同于一般序列结构数据构建形式，其基于树结构构建，包括根结点、孩子节点等不同位置信息，较为特殊，在其上验证实验的模型较少，所以选择同样为电影评论的 IMDB 数据集，采用基于预训练模型的通用情感分析方法进行对比实验，验证本模型的效果。IMDB 数据集是为电影评论的二元情感分类任务而开发的。IMDB 数据集由相同数量的正面和负面评论组成，平均分配为训练集和测试集，训练集和测试集分别有 25 000 条评论。对比基线模型如下。

1) BERT 模型[13]　使用基本模型 BERT_base，12 层 Transformer 编码器块，768 个隐藏单元和 12 个自注意头。

2) BERT+CNN 模型　以 BERT 模型训练词向量，通过 CNN 模型进行分类预测。

3) BERT+RNN 模型　以 BERT 模型训练词向量，通过 RNN 模型进行分类预测。

4) BERT+TRS 模型　以 BERT 模型训练词向量，通过 Transformer 模型进行分类预测。Transformer 是一种标准的编码器-解码器结构。编码器和解码器由多个相同的层组成，主要包括多头注意力层和前馈神经网络层。

2.4 消融实验设计

为了更好地验证本文提出的模型框架对改善向量表示的有效性，针对只经过 BERT 模型表示向量及直接由 SimCSE 模型改善句子嵌入 2 种情况进行消融实验。

3 实验结果与分析

3.1 对比基线模型结果分析

表 1 展示了各种 LSTM 及其变体模型在情感分类任务上训练后的结果。模型对句子情感标签分类的准确度在五分类和二分类 2 种情况下进行了评估。其中，BiLSTM 模型是基于顺序结构的神经网络模型[22]，S-LSTM、Constituency Tree-LSTMs 和 LCT-LSTM 模型均为针对句法结构成分树的 LSTM 变体[9-10,23]。因此，将各模型与 SF-TreeLSTM 模型进行了比较。

表 1　在 SST 上的测试准确率

Tab. 1　Accuracy of the test on SST

%

模　型	准确率	
	五分类	二分类
BiLSTM	82.31	90.25
S-LSTM	83.91	91.01
Constituency-TreeLSTMs	83.62	90.83
LCT-LSTM	86.92	93.59
SF-TLSTM	**87.61**	**96.79**

由表 1 可知，与序列结构 BiLSTM 相比，关注句法结构以树形结构建模的 S-LSTM、Constituency Tree-LSTMs、LCT-LSTM 和本模型的准确率均有所提升，说明句法结构特征的融入在情感分析中发挥关键作用，尤其是 SF-TLSTM 模型在五分类和二分类测试集的准确度得到了普遍提高。对于五分类任务，本模型相较于其他模型，其准确度提高了 1～5 个百分点，对于二分类任务，提升更为明显，提高了 5～6 个百分点，这说明进一步完整提取特征对提高二分类任务的准确度作用显著，这是由于二分类任务中句中的标注只有消极和积极 2 种。S-LSTM 和 Constituency Tree-LSTMs 模型在长距离编码中丢失了部分信息，导致效果不佳，而本模型可以以逐步编码的方式增加节点对文本特征的记忆，因此取得更优效果。对于 LCT-LSTM 模型，虽然准确度得到提升，但由于其在编码时无法并行化导致耗时较长。而本模型能够全面提取文本特征，增强了节点对文本特征的记忆，提高了情感分类任务的准确度。

3.2　基于预训练模型对比实验结果分析

由表 2 可知，对比实验都是基于 BERT 模型，依赖于其强大的向量表征能力，分别与 CNN、RNN、Transformer 相结合，准确率逐步提高，是因为在词嵌入表示一致时，RNN 和 Transformer 模型处理长距离依赖的能力更强。但 3 种模型准确率提升效果不大，这是由于 BERT 模型得到的词嵌入表示在空间上分布不均匀，虽然数据集内容类型相同、结构不同，但本模型的在二分类的预测效果上明显提升，准确率提升了 12.68 个百分点，说明本模型在情感分类预测上凭借着 simCSE 的自监督训练有效改善了词嵌入表示，从而使分类预测更加准确。

表 2　其他模型对比结果

Tab. 2　Comparative results of other models

%

模　型	准确率
BERT	84.11
BERT＋CNN	84.16
BERT＋RNN	84.35
BERT＋Transformer	84.44
SF-TLSTM	**96.79**

3.3　消融实验结果分析

如表 3 所示，从消融实验中，结合表 1 实验结果，可以得到如下信息。BERT 预训练模型能够获得上下文的双向特征表示，但是由于其训练得到的词嵌入表示空间分布不均匀，在单独与 Tree-LSTM 模型结合后准确率提升度较小，甚至在五分类任务上有所下降；将 SimCSE 训练向量与 TreeLSTM 相结合，由于其更侧重于对句子整体的训练，不能对语义向量计算，因此效果反而下降；而 SF-TLSTM 模型通过逐步编码增强了节点记忆，提升了对远距离文本信息的捕捉能力，从而提高了情感分类标签预测的准确度。

表 3　消融实验结果

Tab. 3　Results of ablation experiments

%

模　型	准确率	
	五分类	二分类
BERT＋Tree LSTM	83.16	91.25
SimCSE＋Tree LSTM	83.33	89.82
SF-TLSTM	**87.61**	**96.79**

4　结　语

本文提出一种基于结构-语义融合的评论文本情感分类方法，解决了传统情感分析模型方法无法有效充分利用文本句法结构信息的问题。通过实验数据分析得出以下结论。

1）句法结构特征的融入有效提升了情感分类的准确性。对于五分类任务，本文提出的 SF-TLSTM 方法相较于其他 BERT 模型的准确率提高了 1～5 个百分点，对于二分类任务，提升更为明显，达到了 5～6 个百分点。

2）文本向量表示的数据增强策略提高了文本特征抽取的能力。simCSE 的自监督训练有效改善了词嵌入表示，相比于原始 BERT 向量表示学习训练得到的模型，本文提出的 SF-TLSTM 模型在二分类任务上提升了 12.68 个百分点。

所提 SF-TLSTM 模型虽然在当前数据集的情感分类预测效果上有一定提升，但是由于受数据集结构形式局限，对其他自然语言处理任务的预测效果还需要进一步验证。今后的工作将研究 SF-TLSTM 模型在更加细粒度下的情感分类泛化能力。

参考文献/References：

[1] KUMAR S, GAHALAWAT M, ROY PP, et al. Exploring impact of age and gender on sentiment analysis using machine

learning[J]. Electronics,2020,9(2):374.

[2] YASIN S,ULLAH K,NAWAZ S,et al. Dual language sentiment analysis model for YouTube videos ranking based on machine learning techniques[J]. Pakistan Journal of Engineering and Technology,2020,3(2):213-218.

[3] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics,2014:1746-1751.

[4] 邵良杉,周玉. 基于语义规则与RNN模型的在线评论情感分类研究[J]. 中文信息学报,2019,33(6):124-131.
SHAO Liangshan,ZHOU Yu. Semantic rules and RNN based sentiment classification for online reviews[J]. Journal of Chinese Information Processing,2019,33(6):124-131.

[5] MIWA M,BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics,2016:1105-1116.

[6] CHEN H,HUANG S,CHIANG D,et al. Improved neural machine translation with a syntax-aware encoder and decoder [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics,2017:1936-1945.

[7] MOU Lili,PENG Hao,LI Ge,et al. Discriminative neural sentence modeling by tree-based convolution[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics,2015:2315-2325.

[8] MOU Lili,YAN Rui,LI Ge,et al. Backward and forward language modeling for constrained sentence generation[EB/OL]. (2015-12-21). https://arxiv.org/abs/1512.06612.

[9] ZHU X,SOBHANI P,GUO H. Long Short-Term Memory over tree structures[C]//Proceedings of the 32nd International Conference on Machine Learning. France:[s.n.],2016:1604-1612.

[10] TAI K,SOCHER R,MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics,2015:1556-1566.

[11] LE P,ZUIDEMA W. The forest convolutional network: Compositional distributional semantics with a neural chart and without binarization[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics,2015:1155-1164.

[12] DEVLIN J,CHANG Mingwei,LEE K,et al. BERT:Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics,2019:4171-4186.

[13] YANG Zhilin,DAI Zihang,YANG Yiming. XLNet: Generalized autoregressive pretraining for language understanding [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Canada:[s.n.],2019:5754-5764.

[14] 支世尧,吴贞如,陈涛,等. 基于ALBERT-BiLSTM-Att的微博评论情感分析研究[J]. 计算机时代,2022(2):19-22.
ZHI Shiyao,WU Zhenru,CHEN Tao,et al. ALBERT-BiLSTM-Att models for sentiment analysis of Weibo comments[J]. Computer Era,2022(2):19-22.

[15] CONNEAU A,KJELA D,SCHWENK H,et al. Supervised learning of universal sentence representations from natural language inference data[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics,2017:670-680.

[16] REIMERS N,GUREVVCH I. Sentence-BERT: Sentence embeddings using siamese BERT-networks[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,2019:3982-3992.

[17] GAO Tianyu,YAO Xingcheng,CHEN Danqi. SimCSE: Simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana: Association for Computational Linguistics,2021:6894-6910.

[18] GAO Jun,HE Di,XU Tan,et al. Representation degeneration problem in training natural language generation models[C]// In 7th International Conference on Learning Representations. [S.l.]:[s.n.],2019:1-14.

[19] WANG L,HUANG J,HUANG K,et al. Improving neural language generation with spectrum control[C]//In 8th International Conference on Learning Representations. Addis Ababa:[s.n.],2020:1-15.

[20] 霍欢,薛瑶环,黄君扬,等. 一种针对成分树的混合神经网络模型[J]. 中文信息学报,2019,33(3):8-16.
HUO Huan,XUE Yaohuan,HUANG Junyang,et al. A hybrid neural network model on constituent tree structure[J]. Journal of Chinese Information Processing,2019,33(3):8-16.

[21] SOCHER R,PERELVGIN A,WU J,et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics,2013:1631-1642.

[22] GRAVES A,SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks,2005,18(5/6):602-610.

[23] GAN Ling,GONG Houyu. Text sentiment analysis based on fusion of structural information and serialization information [C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing. Taipe: Asian Federation of Natural Language Processing,2017:336-341.